Development and Psychometric Testing of a Total Quality Management Scale

Dr. Sanjay Kumar
Dean-Academics, (Associate Professor), New Era College of Science and Technology
Ghaziabad (U.P.) India

ABSTRACT

The purpose of this paper is to describe the procedures followed in an instrument development and validation to measure the TQM practices in higher education institutions. The paper explores the process by which a reliable and valid questionnaire can be developed. The paper adopted the mixed methodology, qualitative and quantitative. The qualitative approach was used at the initial stage of item generation and development for the instrument while the quantitative method was predominantly used to empirically establish the psychometric properties of the instrument. The paper finds that the TQM Scale is psychometrically sound. The reliability index for all the dimensions are well above 0.75 and the various empirical analyses, providing evidence of convergent, concurrent, and discriminant validity as well as dimensionality. The paper provides detailed information on developing and validating a new measuring instrument. The method and procedure can be a good reference for researchers interested in developing measuring instruments on educational practices. The soft system methodology, intervention enhances holistic discussion in relation to the focal at the item generation phase. This allows a greater amount of creative debate resulting in a comprehensive list of indicators to be used the instrument. The final questionnaire consists of 30 items and the psychometry indicated that it is valid and reliable.

Keywords: Construct Identification, Item Generation, Psychometric Testing. Validity, Reliability and Construct Validity.

I INTRODUCTION

Fast economic growth, increasing global competition, privatization, globalization and the liberalization of world education has caused an immediate need of TQM implementation in Indian higher education. One of the researchers (Thiagarajan, 1996) suggested total quality management as most prestigious quality indicator for measuring the overall quality of an educational institution. As quality is a fundamental issue in education, the Indian government has also emphasized in improving the overall education quality. Moreover, the Indian government also encourages both private and public universities to acquire quality certification and TQM adoption with the purpose of meeting international standards. However, the current practices of TQM among the various higher learning institutions in India are not much satisfactory. Therefore, it is an urgent need to evaluate and upgrade the TQM practices in India. The purpose of this paper is to describe the procedures followed in an instrument development and validation to measure the TQM practices in higher education institutions. The paper explains the procedure by which a reliable and valid measuring instrument can be developed. The paper adopted the mixed methodology, qualitative and quantitative. The qualitative approach was used at the initial stage of item generation and development for the instrument while the quantitative method was predominantly used to empirically establish the psychometric properties of the measuring instrument.

II CONSTRUCT IDENTIFICATION

The writings of Crosby (1979), Deming (1986), Feigenbaum (1991) and others have developed certain propositions in the area of quality management. Their insights into quality management provide a good understanding of quality management principles. There are several Quality Awards, such as the Deming Prize in Japan, the European Quality Award and the Malcolm Baldrige National Quality Award, which are based on a perceived model of TQM. The constructs of these TQM models must be operational for empirical work. Based on the comprehensive review of the TQM literature, only 7 constructs are believed to be related to the education system. In this paper, The Malcolm Baldrige Award was chosen as the theoretical basis for the identification of TQM constructs because it incorporates a number of different perspectives on quality, rather than focusing on one specific view. The key dimensions of the total quality management identified are categorized as Management Support, Strategic Planning, Student Teachers Satisfaction, Teaching Staff Participation, Training & Development, Continuous Improvement and Information & Analysis. These seven principles can be best applied to higher education Institutions. In this study, the initial draft of TQM scale consisted of 7 constructs of the above TQM principles with modification to suit the higher education institutions setting. These seven constructs of TQM were collectively labeled as Total Quality Management Practices.

III MEASUREMENT APPROACH

Researchers use measuring instruments to measure factual knowledge, judgments, opinions, perceptions, attitudes, feeling, emotions, human behavior and interests. This approach captures the self-reported observations of the individual and is commonly used in educational research. Factual information under a research study is collected using a questionnaire. The rating scale is used for obtaining judgments or opinions or perception of the degree to which an individual possesses certain behavioral traits and attributes not readily detectable by objective tests. Attitude scales are used for finding the attitudes of persons on issues like co-education and religious education, etc. A list of statements is provided to respondents and asks them to respond to each statement in accordance with their true feelings. Tests are designed to measure and describe an aspect of human behavior. Inventory is the most appropriate tool to assess one's interest in a particular phenomenon. In the present paper, researcher is developing a measuring instrument to measure total quality management practices of higher education institutions on the basis of opinion or perception of the teachers and students. For the purpose of the present study, rating scales are best suited. Therefore, measurement of total quality management practices in higher education institutions can be done using rating scales only. There are various types of rating scales used in educational research. Now next question arises, which response style is best suited for the developing TQM scale.

IV RESPONSE STYLES

A number of response styles may be used when a researcher develops a new measuring instrument. These produce different types or levels of data and this will influence the analysis part of the study. Therefore, when developing a new measuring instrument, it is important to be clear which scale and response style to use. Frequency scales may be used when it is important to establish how often a target behavior or event has occurred (Rattray et al. 2004). Thurstone scales are less commonly used in educational research. Thurstone scaling approximates an interval level of measurement (Miller, 1991). Developing a true Thurstone scale is considerably more difficult than describing one (Nunnally, 1994).Guttman scaling is a hierarchical scaling technique that ranks items such that individuals who agree with an item will also agree with items of a lower rank (Katz et al. 1963). Knowledge questionnaires may be helpful when evaluating the outcome of an educational programme (Furze, 2001). They generally offer multiple choice or dichotomous yes/no response options. Likert's scale is widely and very commonly used response style in questionnaires because of its easy construction, high reliability, and successful adaptation to measure many types of affective characteristics. (Nunnally, 1994).

In the present study, TQM scale is being developed to measure the total quality management practices of higher education institutions. Likert scale suited the purpose of the researcher's study; therefore it was kept as a response style for the questions on TQM scale. These scales use fixed choice response formats and are designed to measure attitudes or opinions (Bowling 1997, Burns & Grove 1997). On the Likert rating scale, a respondent indicates agreement or disagreement with a variety of statements on an intensity scale. The five-point "strongly agree" to "strongly disagree" format is used. Responses are then summed across the items to generate a score on the effective instrument. The simplicity and ease of use of the Likert scale are its real strength. The Likert scale can provide an ordinal-level measure of a person's attitude (Babbie, 2001). Gathering and processing the Likert responses are efficient. When several items are combined, more comprehensive multiple-indicator measurement is possible. The rating scales have the advantage of providing data that use values rather than merely categories (Edwards, Thomas, Rosenfeld, and Booth-Kewley, 1997). This feature can provide greater flexibility for data analysis.

V ITEM GENERATION

Questions can be generated from a number of sources, including consultation with experts in the field, proposed respondents and review of associated literature (Priest et al. 1995, Bowling 1997). A key strategy in item generation is to review the research items repeatedly to ensure that items reflect the relevant construct. The draft of TQM scale was derived from the relevant literature and four existing 'quality management' tools: TQM implementation constructs (Ahire, Golhar & Waller,1996), The School Quality Management Scale (Lee 2009), Measure of Quality Management Practices in Education (Jose & Angel, 2014) and Quality Process Management instrument for higher education (David, Nurahimah & Arsaythamby,2016). The initial draft of the TQM scale contained 42 items in seven dimensions: Management Support, Strategic Planning, Student Teachers Satisfaction, Teaching Staff Participation, Training & Development, Continuous Improvement and Information & Analysis.

The most clear and understandable the item; the better the results (Covert, 1984).Poorly framed item produces meaningless responses. Writing items are more of an art than a science (Neuman, 1997).Many authors (Dillman, 2000; Dornyei, 2003) suggested in their studies to keep the grammatical complexities to a minimum. All items in the measuring instrument are grammatically checked. Professional terms, technical terms and abbreviation may have different meanings to different background respondents ((Edwards, Thomas, Rosenfeld, and Booth-Kewley, 1997).The technical terms and professional terms were used to keep in mind the respondents. A short question avoids the ambiguity, confusion, and the vagueness (Neuman, 1997) in the mind of a respondent. The maximum length of a item in the TQM scale is 15 words.

Certain questions should be avoided, e.g. those that leads or includes double negatives or double-barreled questions (Bowling 1997). Fischer and Lewis (1983) find in their survey research that leading questions can influence the answer of respondents to another question on a survey. The total quality management practices scale has no leading statements. The general advice is against the inclusion of negatively worded questions (Foddy, 1993) as they takes longer time to process (Weems, 2002) and have greater chance of mistakes by respondent (Dudycha & Carpenter, 1973). Negatively worded questions are avoided in the TQMP scale. The more general the question, the wider will be the range of interpretations (Converse and Presser, 1986).Specific statement were asked Answers to a hypothetical circumstance are not very reliable, but being explicit will reduce respondents' frustration (Kent, 1993) and able to provide useful information (Babbie, 2001). People tend to under report unhealthy life style practices and over report healthy ones (Brace, 2004) because of social prestige. No sensitive questions were kept in the TQMP scale. Question order effects arise when answering behavior changes depending on the position of a question during the interview (Baker 2003). Consideration is given to the order in which items are presented. All possible considerations were taken into consideration in creating effective items.

VI PILOT WORK

It is important to ensure that sufficient pilot work is carried out during the development of a new measuring instrument. The piloting of a newly developed measuring instrument is done to more clarify its items. A measuring instrument is piloted on a small size sample of potential respondents. Item analysis is one way to pilot a questionnaire. It provides a range of simple heuristics on item retention or deletion. High endorsement of an option within a particular item suggests poor discriminatory power or the redundancy of an item that requires the deletion (Priest et al. 1995). Alternatively, a Cronbach's a < 0.70 may suggest that items in a questionnaire or subscale are poorly grouped. To identify specific items that do not add to the explanatory power of the questionnaire or subscale an item-total correlation cut-off of<0.3 can be used (Ferketich 1991, Kline 1993). However, retain items that are thought to reflect the concerned theoretical domains of the questionnaire despite poor psychometric analysis. Problem items should be identified if higher levels of non-response exist.

A comprehensive item analysis evaluates the items to the scales in the development of a measuring instrument. The item-score to scale score correlation are used to determine if an item belongs to the scale as assigned, is part of another scale, or should be discarded (Nunnally, 1978).The scale-score of each of the seven constructs in the TQM Scale was obtained by computing the mean of the scores of the items that comprise that scale. If an item did not correlate highly with any of the scales, it had to be discarded. Saraph (1989) used this method to evaluate the assignment of items to scales for developing their instrument. Generally, a correlation of item values greater than 0.5 indicates the items has been assigned appropriately to the relative scale. Whereas, item values lower than 0.5 do not share enough variance with the rest of the items in that scale. It is therefore assumed that the items are not measuring the same construct, and that it should be deleted from the scale (Creswell, 2002). The TQM scale was piloted keeping in mind the above discussed points.

(a) **Psychometric Testing-**In the present study, the Total Quality Management Scale was developed to measure quality practice in higher education institutions. This scale consisted of 7 sub-scales of TQM practices which had to be empirically tested and validated. Many methods are available for empirically assessing the reliability and validity of a measurement scale. This section addresses how the reliability and validity of this scale is evaluated to determine the efficacy of TQM Scale in measuring the total quality management practices in higher education institutions.

(b) **Validity Testing-**Validity is the most important consideration in developing and psychometric testing of a measuring instrument. Validity refers to whether a questionnaire is measuring what it purports to (Bryman & Cramer 1997). There are several different types of validity (Polgar & Thomas 1995, Bowling 1997).

(c) **Content validity** was undertaken to ascertain whether the content of the scale was appropriate and relevant to the study purpose. Content validity indicates the content reflects a complete range of the attributes under study and is usually undertaken by seven or more experts (Pilot & Hunger 1999; DeVon et al. 2007). To estimate the content validity of the TQM Scale, the researchers clearly defined the conceptual framework of total quality management by undertaking a thorough literature review and seeking expert opinion. Once the conceptual framework was established, eight purposely chosen experts in the areas of quality, questionnaire design, and academic administration were asked to review the draft of 42-items to ensure that it was consistent with the conceptual framework. Each reviewer independently rated the relevance of each item on the TQM Scale to the conceptual framework using a 4-point Likert scale (1=not relevant, 2=somewhat relevant, 3=relevant, 4=very relevant). The Content Validity Index (CVI) was used to estimate the validity of the items (Lynn 1996). According to the CVI index, a rating of three or four indicates the content is valid and consistent with the conceptual framework (Lynn 1996). For example, if five of eight content experts rate an item as relevant (3 or 4) the CVI would be 5/8=0.62, which does not meet the 0.87 (7/8) level required, and indicates the item should be dropped (Devon et al. 2007). Therefore, three items on the draft TQM Scale were deemed to be invalid because they yielded CVIs of 5/8=0.62 to 6/8=0.75 and were removed from the scale.

(d) **Face validity** indicates the scale appears to be appropriate to the study purpose and content area. It is the easiest validation process to undertake, but it is the weakest form of validity. It evaluates the appearance of the questionnaire in terms of feasibility, readability, consistency of style and formatting, and the clarity of the language used (Haladyna 1999; Trochim 2001; DeVon et al. 2007). Thus, face validity is a form of usability rather than reliability. To determine the face validity of the TQM scale, an evaluation form was developed to help respondents assess each question in terms of the clarity of the wording; the likelihood the target audience would be able to answer the questions, the layout and style. Fifty respondents were randomly selected from twenty educational institutions and they completed the face validity form on a Likert scale of 1-4, strongly disagree= 1, disagree= 2, agree= 3, and strongly agree= 4. All respondents rated each parameter at three or four on a Likert scale of 1-4. Ninety five percent indicated that they understood the questions and found them easy to answer, and 90% indicated the appearance and layout would be acceptable to the intended target respondents.

(e) **Construct validity** refers to the degree to which the items on an instrument relate to the relevant theoretical construct (Kane 2001; DeVon 2007). Construct validity is a quantitative value rather than a qualitative distinction between 'valid' and 'invalid'. It refers to the degree to which the intended independent variable (construct) relates to the proxy independent variable (indicator) (Hunter & Schmidt 1990). When an indicator consists of multiple items, factor analysis is one statistical technique that can be used to determine the constructs validity within the developing measure. Factor Analysis is a statistical method commonly used during measuring instrument development to cluster items into common factors, interpret each factor according to the items having a high loading on it, and summarize the items into a small number of factors (Bryman & Cramer 1999). Loading refers to the measure of association between an item and a factor (Bryman & Cramer 2005). A factor is a list of items that belong together. Related items define the part of the construct that can be grouped together. Unrelated items, those that do not belong together, do not define the construct and should be deleted (Munro 2005). Following the initial pilot work and item deletion, the TQM Scale should be administered to a sample of sufficient size to allow exploratory factor analytic techniques to be performed. Ferguson and Cox (1993) suggest that 100 respondents are the absolute minimum number to be able to undertake this analysis. However, others would suggest that this is insufficient and a rule of thumb would be five respondents per item (Bryman & Cramer 1999). The sampling population for factor analysis was 350 respondents from the higher education institutions in the NCR. The sample was selected using a stratified sampling technique.

(f) **Fairly Well Correlated Items:** In the process of factor analysis, first we get correlation matrix for preliminary analysis. The top half of this matrix contains the Pearson correlation coefficient between all pairs of questions where as the bottom half contains the one-tailed significance of these coefficients. This correlation matrixes check whether all the question items in a developing measuring instrument correlate fairly well. First, the significance values were scanned and found that majority of values was less than 0.05 except one significance value. Then correlation coefficient were scanned and found that majority of correlation coefficient were less than 0.9 except one correlation coefficient value. One significance value greater than 0.05 and one correlation coefficient value greater than 0.9 could arise a problem because of singularity in data. After checking the determinant of the correlation matrix, if necessary, one of the two question items causing the problem can be eliminated. The value of the determinant of the correlation matrix is 0.0005271 which is greater than the necessary value of 0.00001.Therefore; multicollinearity is not a problem in case of TQM scale data. It concludes that all questions in the TQM scale correlate fairly well and there is no need to eliminate any question item at this stage.

(g) Sampling Adequacy: It is essential to have a sufficiently large sample to enable factoring analysis to be undertaken reliably (Bryman & Cramer 2005). Ferguson and Cox (1993) suggest that 100 respondents are the absolute minimum number to be able to undertake this analysis. Although, the number of participants required undertaking factor analysis remains under debate, a minimum of five participants per variable is generally recommended (Bryman & Cramer 1999, Munro 2005). However, to ensure an appropriate sample size was obtained in the current study to enable factoring analysis, Kaiser-Meyer-Olkin (KMO) sampling adequacy was used. The KMO statistic varies between 0 and 1. A value of 0 indicates that the sum of partial correlations is large in comparison to the sum of correlations, which indicates diffusion in the pattern of correlation, and that factor analysis is inappropriate. A value close to one indicates that factor analysis will yield distinct and reliable factors (Field 2005). Kaiser (1974) recommended accepting values ≥ 0.5 and values below this should lead you to either collect more data or rethink which question items to include. Furthermore, described values between 0.5 and 0.7 as mediocre; 0.7 and 0.8 as good, 0.8 and 0.9 as great, and > 0.9 as superb. Therefore, using Kaiser's scale, the sampling adequacy value of 0.91(table-1) for the total quality management scale was superb. So it is clear that the sample size of 200 is appropriate to enable factor analysis to be undertaken. According to George & Mallery (2003), the Bartlett's test also examines whether the correlation matrix is an identity matrix. A significance value < 0.05 indicates that the data do not produce an identity matrix and are thus approximately multivariate normal and acceptable for factor analysis. The significance value of < 0.01 in this study was ideal for factor analysis as depicted in table 1.

Table-1
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measures of Sampling Adequecy		0.91
Bartlett's Test of Sphericity	Approx. Chi-Square	19046.86
	Degree of Freedom (df)	349
	Significance	0.000

(h) Factor Extraction: Several types of extraction methods are used to undertake factor analysis. The most common method: Principal Component Analysis (PCA) was performed on the TQM Scale.PCA explores the inter-relationship of variables. It provides a basis for the removal of unnecessary items in a developing measure (Anthony, 1999) and can identify the associated underlying concepts, domains or subscales of a questionnaire (Oppenheim 1992, Ferguson & Cox, 1993). Therefore, PCA is assumed to be perfectly reliable and without error (Bryman & Cramer 2005). In this study, the principal component analysis was selected and conducted on the measurement instrument. The following five criteria were used in this study to determine the number of factors to be retained.

(i) Factors with eigenvalues greater than 1.0
(ii) Examination of the Scree plot
(iii) The magnitude of factor loading score greater than 0.40
(iv) The presence of the correlation with the other resulting factors
(v) The conceptual meaningfulness of the factors (Tabachink & Fidell,1996)

On the first run PCA, the eigenvalues associated with each linear component (factors) before extraction, after extraction and after rotation are explained. The eigenvalues associated with each factor represent the variance explained by that particular linear component. SPSS also displays the eigenvalue in terms of the percentage of variance explained. First Factor of TQM Scale explains 41.33% of total variance which means the first factor explains relatively large amounts of variance, whereas subsequent factors explain only a small amount of variance. SPSS then extracts all factors with eigenvalues greater than 1. An eigenvalue is an estimate of variance explained by a factor in a data set (Ferguson & Cox 1993), and a value >1 indicates greater than an average variance. In the present study, SPSS extracts 6 factors for TQM Scale. Rotation technique equalizes the relative importance of these extracted factors. Before rotation first factor accounted for considerably more variance than the remaining five (41.33% compared to 6.65%, 5.07%, 3.88%, 3.12% and 2.57%), however, after rotation it accounts for only 21.37% as compared to 13.47%, 10.92%, 6.76%, 5.48% and 4.62% respectively.

SPSS shows the table of the communities before and after extraction. PCA works on the initial assumption that all variance is common; therefore before the extraction value of all the commonalities is 1. After extracting some of the factors are discarded and so some information is lost. The amount of variation in each variable that can be explained by the retained factors is represented by the communities after extraction. Further SPSS also shows the component matrix before rotation. This matrix contains the loadings of each variable onto each factor. Factor loadings greater than 0.30 are considered significant; loadings of 0.40 are considered most important; if the loadings are 0.50 or greater, they are considered very significant (Hair et al., 1992). In this study, a factor loading of 0.40 was used as the cutoff point. It is suggested that each factor should have at least three items with significant structure coefficients (Tabachinck & Fidell,1996).

At this stage SPSS has extracted 6 factors. It becomes necessary to know whether the decision on the number factors to extract is correct. According to Steven (2002) and Field (2005), Eigen values and the Scree plot are accurate to determine how many factors should be retained. By the Kaiser's criterion we should extract six factors and this is what SPSS has done. However, this criterion is accurate if there are less than 30 variables and communalities after extraction are greater than 0.7 or if the sample size exceeds 250 and the average communality is greater than 0.6 then retain all factors with Eigenvalues above 1 **(Kaiser's criterion)**.The commonalities obtained in the present study are not greater than 0.7 and the average of the communalities (0.53) at not greater than 0.6.So, on both grounds Kaisher's rule may not be accurate. If Kaisher's rule does not apply, a Scree plot can be used when the sample size is large (around 300 or more cases). A scree test is the graphic representation of eigenvalues. An inspection of the scree plot indicated a sudden drop in the scree beginning with the seventh factor as depicted in figure 1, hence only 6 factors were kept for further analysis.

In interpretation of the factors, factor structure coefficients (faster loading) were used to describe correlations between each variable in the original variable set and each of the factors that was retained. Only factor structure coefficients of 0.40 or greater were considered to be significant and used to interpret the factors. It is suggested that each factor should have at least three items with significant structure coefficients (Tabachinck & Fidell, 1996). The relationship between the relevant subscales was examined by conducting Pearson correlations among these subscales and significant correlations were observed among these subscales as depicted in Table 4.2. Item content was examined for each factor to see if an underlying theme was identifiable. It was decided that the first 6 factors were conceptually meaningful and had greatest conceptual clarity in describing quality improvement strategies in school management. Therefore, they were retained in the final instrument. These 6 factors accounted for 62.62% of the total variance.

Table-2
Coefficient of Correlation among sub-scales

	MS	SP	TSP	TAD	CI	IAN
Management Support (MS)	1					
Strategic Planning (SP)	0.77**	1				
Teaching Staff Participation (TSP)	0.57**	0.59**	1			
Training and Development (TAD)	0.71**	0.68**	0.66**	1		
Continuous Improvement (CI)	0.59**	0.52**	0.58**	0.72**	1	
Information and Analysis (IAN)	0.75**	0.73**	0.63**	0.63**	0.73**	1

****Significant at the 0.01 level**

(i) Factor Rotation: The two methods of factor rotation are used under factor analysis. The orthogonal rotation is used when it is assumed that factors should theoretically independent and oblique rotation is used when it is assumed that the factors should be related to each other. The orthogonal rotation was used in the present study. It provides the rotated component matrix or rotated factor matrix which is a matrix of the factor loadings for each variable onto each factor. This matrix contains the same information as component matrix except that it is calculated after rotation. Before rotation, most variables loaded highly onto the first factor and the remaining factors did not really get a look into. But this problem is solved after the rotation of the factors.

The Results of the Final Six Factors solution of the Total Quality Management Practices according to Principal Component Analysis with Varimax Rotation and the internal consistency of each factor.

Table-3 (a)
Factor Loadings of TQM Practices and Cronbach's Coefficient Alpha for Factor-1*

Items (Alpha value= 0.87, Eigenvalue=18.07, Variance=41.33% and CV=41.33%)	Loadings
Management is committed to promote quality within the Institution.	0.93
Management clearly communicates vision, mission, objectives and policy statements.	0.83
Quality awareness programs are conducted by the Institution regularly.	0.85
Management provides internet facility to explore new areas of knowledge.	0.82
The institution has strong linkages with Industry for training and placement.	0.74
The institution has well- established system to handle inquiries and complaints.	0.73
Management ensures availability of qualified and competent teaching staff.	0.68
Management provides good quality of Infrastructural facilities.	0.63
Extraction Method: Principal Component Analysis **Rotation Method:**Varimax with Kaiser Normalization	

***Factor 1: Management Support**

Table-3 (b)

Factor Loadings of TQM Practices and Cronbach's Coefficient Alphas for Factor-2*

Items (Alpha value= 0.83, Eigenvalue=2.23,Variance=6.65% and CV=47.98%)	Loadings
The Planning of the Institution involves both teachers and Students.	0.91
Regular meetings are conducted to improve the quality of education.	0.85
The Institute develops an annual plan which is implemented and updated regularly.	0.82
Information collected from students is used for Institutional Planning.	0.81
The institution has quality cell for planning & maintaining quality of education.	0.67
Relevant information is communicated to the students well in Advance.	0.61
Extraction Method: Principal Component Analysis **Rotation Method:**Varimax with Kaiser Normalization	

***Factor 2: Strategic Planning**

Table-3 (c)

Factor Loadings of TQM Practices and Cronbach's Coefficient Alphas for Factor-3*

Items (Alpha value= 0.84, Eigenvalue=1.92,Variance=5.07% and CV=53.05%)	Loadings
Teachers are ready to help students inside and outside the class.	0.91
Students are encouraged to participate in extra-curricular Activities.	0.82
Teachers encourage student teachers to raise doubts and ask questions.	0.74
The creativity of the students is always encouraged and supported by teachers.	0.48
Teachers take an interest in the character building of Student teachers.	0.59
Extraction Method: Principal Component Analysis **Rotation Method:**Varimax with Kaiser Normalization	

***Factor 3: Teaching Staff Participation**

Table-3 (d)

Factor Loadings of TQM Practices and Cronbach's Coefficient Alphas for Factor-4*

Items (Alpha value= 0.87, Eigenvalue=1.53, Variance=3.88% and CV=56.93%)	Loadings
Special classes are conducted to improve the communication skills of the Students.	0.87

Special arrangements are made for the professional training of the students.	0.72
Teachers get a regular review of the training and development of students.	0.61
Teachers conduct seminars / workshops / conferences regularly.	0.48
Extraction Method: Principal Component Analysis **Rotation Method:** Varimax with Kaiser Normalization	

***Factor 4: Training and Development**

Table-3 (e)
Factor Loadings of TQM Practices and Cronbach's Coefficient Alphas for Factor-5*

Items (Alpha value= 0.80, Eigenvalue=1.37, Variance=3.12% and CV=60.05%)	Loadings
The regular assessment system is adopted to improve the performance of the students.	0.79
The complaints from students and stakeholders are immediately looked into.	0.67
Repair and maintenance of the instructional infrastructure are done regularly.	0.53
The college has developed quality standards that every teacher and students must meet.	0.47
Extraction Method: Principal Component Analysis **Rotation Method:** Varimax with Kaiser Normalization	

***Factor 5: Continuous Improvement**

Table-3 (f)
Factor Loadings of TQM Practices and Cronbach's Coefficient Alphas for Factor-6*

Items (Alpha value= 0.78, Eigenvalue=1.07, Variance=2.57% and CV=62.62%)	Loadings
Institution collects feedback to determine the quality of its services.	0.77
Teachers use feedback data in their decision making.	0.72
Institution regularly updates and maintains its website.	0.62
Extraction Method: Principal Component Analysis **Rotation Method:** Varimax with Kaiser Normalization	

***Factor 6: Information and Analysis**

Table-3 (a) to 3 (f) shows the factor loading of all the 30 items from the six components in the TQM scale obtained from factor analysis. Eigenvalues, cumulative variance explained by each individual factor and Cronbach's Coefficient Alphas were also presented. Eight items (1,2,3,4,5,6,34,35) loaded solely on the first factor (management support), Six items (8,10,11,12,13,25) on the second factor (strategic planning), five items (16,17,18,20,21) on the third factor (teaching staff participation), four items (22,23,26,28) loaded solely on the fourth factor (training and development), four items (29,30,32,33) on the fifth factor (continuous improvement) and tree items (37, 40,42) loaded on sixth factor (information & analysis).

The factor loading criterion of 0.40 or higher was used in this study for factors extraction. These resulted only 30 items were retained in the original 42- item TQM scale after performing a thorough examination on the item-factor loading and the reliability coefficient. This remaining 30 viable items categorized into six components or constructs of the Total Quality Management Practices. Table-4 presents the item list each of the six sub-scales of the total quality management practices in this study based on the results of principal component analysis (after rotation).

Table-4
Six Sub-scales of Total Quality Management Practices

Sub-scales	Item List	Number of Items
Management Support	1,2,3,4,5,6,34,35	8
Strategic Planning	8,10,11,12,13,25	6
Teaching Staff Participation	16,17,18,20,21	5
Training and Development	22,23,26,28	4
Continuous Improvement	29,30,32,33	4
Information and Analysis	37,40,42	3

(j) Reliability Testing-Once the validity procedures were completed, the final version of the TQM scale was examined to assess its reliability. Reliability refers to the ability of a questionnaire to consistently measure an attribute and how well the items fit together, conceptually (Haladyna 1999; Devon et al. 2007). Although reliability is necessary, is not sufficient to validate an instrument, because an instrument may be reliable but not valid (Beanland et al. 1999; Pilot & Hunger 1999, DeVon et al. 2007). Cronbach & Shavelson (2004) suggested researchers should consider the following issues when determining reliability:

(i) The standard error of the instrument, which is the most important reliability information to report.
(ii) Independence of sampling.
(iii) Heterogeneity of content.
(iv) How the instrument is used.

Two estimates of reliability are commonly used: internal consistency reliability and test-retest reliability: both were used to examine the reliability of the TQM Scale.

(k) Internal consistency reliability examines the inter-item correlations within an instrument and indicates how well the items fit together conceptually (Nunnally & Bernstein 1994; DeVon et al. 2007). In addition, a total score of all the items is computed to estimate the consistency of the whole questionnaire. Internal consistency is measured in two ways: Split-Half reliability and Cronbach's alpha correlation coefficient (Trochim 2001). In **Split-Half reliability**, all items that measure the same construct are divided into two sets and the correlation between the two sets is computed i.e 0.87.

(l) Cronbach's alpha is equivalent to the average of the all possible split-half estimates and is the most frequently used reliability statistic to establish internal consistency reliability (Trochim 2001; DeVon et al. 2007). Cronbach's alpha was computed to examine the internal consistency of the TQM Scale. If an instrument contains two or more subscales, Cronbach's alpha should be computed for each subscale as well as the entire scale (Nunnally & Bernstein 1994; DeVon et al. 2007). Therefore, Cronbach's alpha was computed for each subscale. Cronbach's alpha was computed for the revised TQM Scale after construct validation was computed and was 0.91, which indicates a high correlation between the items and the questionnaire is consistently reliable. Opinions differ about the ideal alpha value. Some experts recommend the alpha should be at least 0.90 for instruments used in clinical settings (Nunnally & Bernstein 1994). Others suggest an alpha of 0.70 is acceptable for a new instrument (DeVellis 1991; DeVon et al. 2007). The alpha computed for each of the four subscales also exceeded the minimum value for a new tool: all subscales were ≥ 0.70, see Table 4.1 to 4.7.

(m) Test–retest reliability can assess stability of a measure over time and this should be included in the process of any questionnaire development. This is of particular importance if the intended use of the measure is to assess change over time or responsiveness. Test-retest reliability is estimated by administering the same tool to the same sample on two different occasions on the assumption there will be no substantial change in the construct under study between the two sampling time points (Trochim 2001; DeVon et al. 2007). A high correlation between the scores at the two time points indicates the instrument is stable over time (Haladyna 1999; DeVon et al. 2007). The duration of time between the two tests is critical. The shorter the interval, the higher the correlation between the two tests, the longer the interval, the lower the correlation (Trochim 2001). However, very long test intervals can affect the results because of changes in participants or their environment (Linn & Gronlund 2000; DeVon et al. 2007). Currently, there is no definite evidence about the best time interval to allow between the test and the retest. Researchers need to consider factors such as the effects of time on health status such as deterioration or improvement in health and what the results will be used for, to make an appropriate decision about the time interval between tests (Concidine, Botti & Thomas 2005). Test-Retest reliability of the TQM Scale was undertaken by administrating the questionnaire to 25 teachers randomly selected from four higher education institutions in an inner city area. They completed the TQM Scale on two different occasions; at baseline and eight weeks later. The coefficient of correlation of test-retest method was 0.77. Because ordinal data were obtained from the questionnaire using a five point Likert scale rated from strongly disagree to strongly agree; and the scale was not continuous, non-parametric statistical tests were deemed to be more appropriate than Pearson Correlation Coefficient. Therefore, the analysis of responses between the test and the retest was conducted using Wilcoxon Non-parametric Statistical Test to determine whether there were any significant differences between the responses at each time point.

VII CONCLUSION

This paper emphasizes the need to adopt a logical, systematic and structured approach to scale development. The researcher has presented a framework that supports this type of approach and has illustrated the rating scale development process using item analysis, factor analysis and related methods and has demonstrated strategies to demonstrate the reliability and validity of the new and developing measures. If a measure is poorly designed and has had insufficient psychometric evaluation, it may be difficult to judge between such competing explanations. In addition, it may not be possible to use the findings from an established measure, if that measure cannot be shown to be reliable in a particular sample. If educational practice is to be enhanced or changed using findings derived from scale-based methods, it is vital that the scale has been sufficiently developed. This paper presents a critical evaluation of the scale design and development process and demonstrates best practice at each stage of this process. This paper will enable the informed researcher to plan the design and development of their own research tools, to evaluate the quality of existing educational measures, and to inspire confidence in applying findings into practice.

REFERENCES

[1] Ahire, S., Golhar, D., & Waller, M. (1996). Development and validation of TQM implementation constructs. Decision Sciences, 27 (1), 23-56.

[2] Anthony, D. (1999). Understanding Advanced Statistics: A Guide for Nurses and Health Care Researchers. Churchill Livingstone,Edinburgh.

[3] Babbie, E. R. (2001). The practice of social research. Belmont, CA: Wadsworth.

[4] Baker M.J. (2003). Data collection – questionnaire design. Marketing Review, 3, 3, pp. 343–370.

[5] Beanland C., Schneider Z., LoBiondo-Wood G. and Haber, J. (1999). Nursing Research: methods, critical appraisal and utilization. Sydney: Mosby.

[6] Bowling, A. (1997). Research Methods in Health. Open University Press, Buckingham.

[7] Brace, I. (2004). Questionnaire Design. How to Plan, Structure and Write Survey Material for Effective Market Research. London: Kogan Page.

[8] Bryman, A., Cramer, D. (1999). Quantitative Data Analysis with SPSS Release 8 for Windows. A Guide for Social Scientists. London: Routledge.

[9] Bryman, A., Cramer, D. (2005). Quantitative Data Analysis with SPSS12 and 13. A Guide for Social Scientists. East Sussex Routledge.

[10] Burns, N. & Grove, SK. (1997). The Practice of Nursing Research Conduct, Critique, & Utilization. W.B. Saunders and Co., Philadelphia.

[11] Concidine, J., Botti, M., Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. Collegian, 12 (1), 19-24.

[12] Converse, J. M., and Presser, S. (1986). Survey questions: Handcrafting the standardized questionnaire. Newbury Park, CA: Sage.

[13] Covert, R. W. (1984). A checklist for developing questionnaires. Evaluation News, 5(3), 74–78.

[14] Creswell, J. W. (2002). Educational research: planning, conducting and

evaluating qualitative & quantitative research. New Jersey: Merill Prentice Hall.

[15] Cronbach, L.J., Shavelson, R.J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures Educational and Psychological Measurement, 64 (3), 391-418.

[16] Crosby, P. B. (1979). Quality is free. New York: McGraw-Hill.

[17] David, J. K., Nurahimah, M.Y. & Arsaythamby,V. (2016).Validating Quality Process Management instrument for higher education using structural equation modeling,International Journal for Quality Research, 10(2), 341-354.

[18] Deming, W. E. (1986). Out of the crisis. Cambridge, MA: MIT Center for Advanced Engineering Study.

[19] DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J. et al. (2007). A psychometric Toolbox for testing Validity and Reliability. Journal of Nursing scholarship, 39 (2), 155-164.

[20] Dillman, D. (2000). Mail and Internet Surveys. The Tailored Design Method. New York: JohnWiley & Sons, Inc.

[21] Dörnyei, Z. (2003). Questionnaires in Second Language Research. Mahwah, NJ: Lawrence Erlbaum.

[22] Dudycha, A.L. & Carpenter, J.B. (1973). Effects of item format on item discrimination and difficulty. Journal of Applied Psychology, 58, pp. 116–121.

[23] Edwards, J. E., Thomas, M. D., Rosenfeld, P., and Booth-Kewley, S. (1997). How to conduct organizational surveys: A step-by-step guide. Thousand Oaks, CA: Sage.

[24] Feigenbaum, A. V. (1991). Total quality control. New York: McGraw-Hill.

[25] Ferguson E & Cox T (1993). Exploratory factor analysis: a user’s guide. International Journal of Selection and Assessment 1,84–94.

[26] Ferketich, S. (1991). Focus on psychometric: aspects of item analysis. Research in Nursing and Health 14, 165–168.

[27] Field, A. P. (2005). Discovering Statistics using SPSS. London: Sage.

[28] Field, J. C. (1994). Total quality for schools: A guide for implementation. Milwaukee, WI: ASQC.

[29] Fischer, A.J., and Lewis, R.D (2008). “Deliberately Biasing Survey Questions.” Australian and New Zealand Journal of Statistics. Volume 27, issue 1.

[30] Foddy, W. (1993). Constructing Questions for Interviews and Questionnaires. Theory and Practice in Social Research. Cambridge, UK: Cambridge University Press.

[31] Furze G, Lewin RJP, Roebuck A, Thompson DR & Bull P (2001). Attributions and misconceptions in angina: an exploratory study. Journal of Health Psychology, 6, 501–510.

[32] George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference 11.0 update. Boston: Allyn & Bacon

[33] Hair, J.F.; Anderson, R.E.; Tatham, R. L.; Black, W.C. (1998). Multivariate Data Analysis, Fifth Edition, Prentice-Hall: Upper Saddle River.

[34] Haladyna, T. (1999). Developing and Validating multiple-choice test items. New Jersey: Lawrence Erlbaum.

[35] Hunter, J.E., Schmidt, F. L. (1990). Methods of meta-analysis: Correcting errors and bias in research findings. Newsbury Park: Sage Publications.

[36] Jose, S.G. & Angel,R.M. (2014).Development and Validation of a measure of the quality management practices in education, Total Quality Management and Business Excellence, 25 (1-2)

[37] Kaiser, H. (1974). An index of factorial simplicity Psychometrika, 39 (1), 31-36.

[38] Kane, M. (2001). Current concerns in validity theory. Journal of Educational Measurement, 38, 319-342.

[39] Katz S, Ford A & Moskowitz R (1963) Studies of illness in the aged: the index of ADL. A standardised measure of biological

and psychosocial function. Journal of American Medical Association 185, 914–919.

[40] Kent, R. A. (1993). Marketing research in action. London, England: Routledge.

[41] Kline, P. (1993). The Handbook of Psychological Testing. Routledge, London.

[42] Lee, S.S. (2009). The construction and validation of a school quality management scale for quality improvement in school management, Thesis, University of Malaya, Malaysia.

[43] Lynn, M.R. (1996). Determination and quantification of content validity. Nursing Research, 35, 382-385.

[44] Miller, D. C. (1991). Handbook for research design and social measurement. Newbury Park, CA: Sage.

[45] Munro, B.H. (2005). Statistical methods for health care research. Philadelphia: Lippincott, Williams & Wilkins.

[46] Neuman, W. L. (1997). Social research methods: Qualitative and quantitative approaches. Boston: Allyn & Bacon.

[47] Nunnally, J.C., Bernstein, I.H. (1994). Psychometric theory. New York: McGraw-Hill.

[48] Oppenheim AN (1992). Questionnaire Design, Interviewing and Attitude Measurement. Pinter, London.

[49] Pilot, D., & Hunger, B. (1999). Nursing research: principals and methods. Philadelphia: Lippincott Williams & Wilkins.

[50] Priest J, McColl BA, Thomas L & Bond S (1995). Developing and refining a new measurement tool. Nurse Researcher 2, 69–81.

[51] Rattray JE, Johnston M & Wildsmith JAW (2004). The intensive care experience: development of the intensive care experience (ICE) questionnaire. Journal of Advanced Nursing 47, 64–73.

[52] Saraph,.J., Benson, G., & Schroeder, R. (1989). An Instrument for measuring the critical factors of quality management. Decision Sciences, 20 (4),810-829.

[53] Stevens, J. (2002). Applied Multivariate Statistics for the Social Sciences (4th Edition). Mahwah, NJ: Lawrence Erlbaum Associates.

[54] Tabachnick, B. G., & Fidell, L. S. (1996). Using Multivariate Statistics (3rd ed.). New York: Harper Collins.

[55] Thiagarajan, T. (1996). An empirical study of total quality management (TQM) in Malaysia: a proposed framework of generic application , unpublished PhD thesis, University of Bradford, Bradford.

[56] Trochim, W.M.K. (2001). The Research Methods Knowledge Base. Cincinnati: Atomic Dog.

[57] Weems, G.H., Onwugbuzie, A.J. & Lustig, D. (2002). Profiles of respondents who respond inconsistently to positively and negatively worded items on rating scales. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Chattanooga, TN.