Journal of Data Analysis and Information Processing, 2020, 8, 358-377
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Scale for Evaluation of Humanities Courses: Development, Reliability and Validity

Raghavendra Dwivedi[1]*, Nagendra Nath Pandey[2]

[1]Shri Jagdish Prasad Jhabarmal Tibrewala University, Vidyanagari, Jhunjhunu, India
[2]P.D.Lions College of Commerce & Economics [MS41], Malad (West), Mumbai, India
Email: *raghavendra.jjtu@gmail.com

How to cite this paper: R. Dwivedi and Pandey, N.N. (2020) Scale for Evaluation of Humanities Courses: Development, Reliability and Validity. *Journal of Data Analysis and Information Processing*, **8**, 358-377.
https://doi.org/10.4236/jdaip.2020.84021

Received: September 19, 2020
Accepted: November 27, 2020
Published: November 30, 2020

Abstract

The present paper deals with development of a course evaluation survey (CES) questionnaire regarding evaluation of humanities courses (e.g., Commerce/Economics/Law) and assessment of its validity and reliability. Taking help from available questionnaires and also incorporating views from the teachers of the related subjects, the relevant Likert type items were selected. The drafted version of the questionnaire was finalized taking into account feedback from the subject experts. The focus was on simple language to ensure clarity to the students to facilitate accuracy of the collected data. The developed questionnaire consists of five domains involving fifteen Likert type items including a global item. For assessment of its reliability and validity, data were collected among two groups of undergraduate students studying two courses namely Accounts (n = 190) and Law (n = 90). For assessing questionnaire, Karl Pearson Correlation Coefficient, Cronbach's Alpha coefficient and four factor solutions under factor analysis were used. The analytical results were quite consistent across the data sets (that are, accounts, law and pooled together) that exhibit good psychometric properties. It may thus be concluded that the questionnaire is valid, reliable and may measure what it is intended to measure.

Keywords

Course Evaluation Survey, Reliability, Validity, Factor Analysis

1. Introduction

There are various requirements to ensure quality in higher education among which an important aspect is to adopt evaluation system on a regular basis. The reputed academic institutions globally have often integrated program of evaluating various components of educational system, mainly by the students, for example, evaluation of each and every course under an academic program, evaluation of lecturing by faculty of various courses, evaluation of facilities like library, computational facilities, and sports facilities. In some countries, facilities related to religious practices within an academic institution also need to be evaluated. The ultimate goal of integrating such evaluations within institutional routine activities is to know self-graded satisfaction by the students. In this regard, an individual instrument for each of the evaluations needs to be developed and validated before its adoption. As such, even one may modify the existing instrument to suit the need of a specific regional environment. Sometimes, if an existing instrument developed elsewhere is being used; one needs to ensure its appropriateness in terms of comparative environment as well as practices. Also, if copyright, due to permission from concerned people regarding its use needs to be collected. As a matter of fact, use of such practices is very minimal in developing countries like India in general, and in humanities in particular [1] [2].

The present paper aimed to develop a questionnaire regarding evaluation of humanities courses (e.g., Commerce/Economics/Law) and assess its validity and reliability. To develop the questionnaire suitable for the proposed setting, a comprehensive review of literature with a focus on target questionnaires was carried out [1] [2]. Taking help from such questionnaires and also incorporating views from the teachers of the related subjects, the relevant items were selected and drafted version of the questionnaire was prepared. The drafted questionnaire was distributed amongst the subject experts; taking into account their feedback, the type items were rearranged with a focus on the clarity of the questions. The focus was on simple language to ensure clarity to the students to facilitate accuracy of the collected data. To assess the validity and reliability of finalized instrument, data were collected on target courses.

2. Stages of Scale Development

In this modern era where concepts are ever evolving, it may not be always possible to find a subject specific, psychometrically sound and culturally appropriate scale. This necessitates development of a sound scale measure. However, the process for developing reliable and valid measurement instruments requires a well-established framework to guide researchers through the various stages of scale development. To simplify, various steps were involved in developing the questionnaire: item generation, expert evaluation and developing the content and structure, pilot testing and evaluation, final testing in the present study.

2.1. Items Generation

The generation of appropriate items to describe the contents of the required scale is the most crucial element of establishing sound measures. Also, called as content validity of the scale, it incorporates every aspect of its construct measure. This brings us to the basic question that how do we proceed with identifying multidimensional nature of construct when the knowledge on any given subject matter may either be abundant or too scanty.

There are 2 primary ways in which domains of construct can be generated, 1) through exploring subjective experiences of the phenomenon; and 2) thorough comprehensive review of existing literature on the items, construct and subsequent pooling of items. All these approaches were used in the development of the present questionnaire. The supervisor and the researcher discussed in length and breadth about the focus on relevant items that can capture the responses appropriately describing desired construct.

A thorough review of literature was performed for selecting the questions. For this, the related publications were searched from the field of assessment, evaluation, education, academics and related disciplines, with a focus on course evaluation surveys under accreditation of higher education. The relevant titles were retrieved and the contents with/ without used questionnaires were managed adequately through a bibliography software. The multiple publications from a single study were merged by adopting the process of filtering, sorting and text management. From the theoretical aspects, review involved thorough review of scales and generation of domain specific items under the planned questionnaire. Subsequently, inappropriate items for the proposed set up were dropped from the list, and a list of drafted items was prepared. Also, few of the items were replaced/ added in view of the need of present set up [3] [4] [5] [6]. The questionnaire was reviewed further for adequacy with respect to friendly language and clarity about the content.

2.2. Developing the Structure of the Questionnaire

Based on the considered relevant domains, questions were divided into various sections/domains. The comprehension level of the domains as well as the items under them was kept to minimal level so that it was easy to understand and administer. Next crucial step was to have consensus on the type of responses. It was being felt after deliberation that the Likert type item response would be befitting the situation the most. Hence, on the lines of predominant use of five points Likert type item responses in the literature, each of the Likert type items was coded as:

1 = Strongly Disagree, means the statement is true very rarely

2 = Disagree, means the statement is true poorly

3 = True Sometimes, means the statement is true about half the time

4 = Agree, means the statement is true most of the time and

5 = Strongly Agree, means the statement is true all or almost all of the time

These steps shaped the questionnaire for an independent evaluation by experts on relevant varying aspects.

2.3. Expert Evaluation of Structure of the Drafted Questionnaire

The review by experts (involving subject experts as well as research methodologists) ensured face validity and appropriateness of contents of questionnaire. In the present study, the experts were the faculties of the college who worked collaboratively with the scholar and the supervisor.

During review of the selected items, due precaution was taken to avoid highly correlated items. For this, two ways were adopted in the present study. First, opinions of experts on the construct of the framed individual questions and their overall observations about the same were recorded. Second, item specific review involved placing the expert's response into either of three forms: "Must be Kept", "Can be Kept" and "May be deleted".

Overall feedback was an open ended response so as to judge the length, language, ease in the comprehension level and appropriateness of domain specific items. The responses of the experts were reviewed and the questionnaire was further modified to incorporate their suggestions.

2.4. Finalization of Scale Structure and Pilot Testing

The revised scale after incorporating changes as suggested in expert review was pilot tested among few volunteer students with a focus on language simplicity, clarity and friendly understanding by the students. The items were explained to the students. Since the students under pilot testing were ensured to resemble the actual study population, it provided true indication of suitability of the questionnaire and scope for further desired modification.

During the process of pilot testing, other than few minor corrections, the questionnaire was found to be framed suitably and easy to administer in the true sense.

Under pilot testing, the researcher emphasized that students should respond with truthfulness so as to ensure data quality. During the course of discussion, it was felt that students wished to respond on some characteristics, the provision for which was not made in the questionnaire.

Since students might have varying opinions on some aspects about the courses, there was need to have *a priori* provisions for this. Therefore, such opinions on part of students were recorded through open ended questions.

3. Broad Domains of the Questionnaire

The questionnaire finalized through pilot-testing involved five domains. As described below, the first domain had three items, the second had seven items, the third also had three items, and the fourth had only one item, and the fifth had one item. In addition, there were three open ended items to capture opinion of students on some aspects.

3.1. Specific Domains

The first domain had questions pertaining to the commencement of the courses as described below:

Domain I: At the beginning of the course, I was told about:

1) *The course contents & expected knowledge and skills*

2) *Available resources for studies, including faculty hours and study material*

3) *The doable things to succeed, including steps and criteria for assessment*

The second domain had questions regarding the faculties and had seven items as described below:

Domain II: During this course, my teachers:

4) *Conducted the classes as per the course contents*

5) *Were interested in what they were teaching*

6) *Were available beyond class hours to assist me*

7) *Were fully committed towards teaching* (*e.g.*, *Punctuality*, *clarity in teaching*)

8) *Used updated teaching materials* (*Slides*, *hand-outs*)

9) *Encouraged me to ask questions*

10) *Explained the links between this course and other courses in the subject*

The third domain aimed at capturing the information on available infrastructure through three items as described below:

Domain III: During this course, my college/department:

11) *Allocated sufficient classes for the listed course contents*

12) *Provided the resources* (*textbooks*, *library*, *and computers*)

13) *Provided required technology* (*internet facility* & *programs*) *to support learning*

The fourth domain had a single item for recording students applied knowledge and skills by registering under the respective courses as described below:

Domain IV: This course helped me to:

14) *Understand* & *apply desired knowledge*, *than simply to memorize*

The fifth domain was the "GLOBAL ITEM" targeted to understand the overall satisfaction level on part of students as described below:

Domain V: Overall Evaluation:

15) *I was satisfied with overall quality of this course*

This may further help to assess the consistency of reporting on global item (item 15) with those reported on fourteen individual items. In addition to the fifteen Likert type items described earlier under various domains, there were three open ended questions as described below:

Open Ended Items:

16) *What did you like most about this course*? *Two major aspects on priority*

17) *What did you dislike most about this course*? *Two major aspects on priority*

18) *What suggestion*(*s*) *do you have to improve this course*? *Two major aspects on priority.*

The quantitative analysis of the recorded Likert response relied upon first five domains. The information on open ended questions was not utilized for the quantitative analysis because it has responses that were qualitative in nature.

3.2. Reliability and Validity of the Questionnaire

The pilot testing provides information about the feasibility of scale administration and appropriateness of scale items in terms of time taken to administer, and difficulty level of the items. After the pilot administration, before final use, scale was subjected to psychometric analysis including assessment of theoretical structure and reliability measured through Cronbach's alpha. Cronbach's alpha is a measure of internal consistency, that is, how closely related a set of items are as a group? Thus, it is considered to be a measure of scale reliability. A "high" value for alpha does not imply that the measure is unidimensional. Value above 0.75 is considered acceptable [7] for retention and items/dimensions.

Also, inter-item correlation matrix was plotted (not listed in this article) to provide an assessment of item redundancy: the extent to which items on a scale are assessing the same content [8]. Ideally, the average inter-item correlation for a set of items should be between 0.20 and 0.40, suggesting that while the items are reasonably homogenous, they do contain sufficiently unique variance so as to not be isomorphic with each other. Based on calculation on this measure, retaining of items in the questionnaire was further justified. However, depending on these results, appropriate revision of the scale-items could further be done and the final structure of the scale could be prepared.

3.2.1. Data Used for Validation

For this, the developed questionnaire consisting of fifteen Likert type items including global item was used to collect data among two groups of undergraduate students studying two courses namely Accounts and Law from the Prahladrai Dalmia Lions College of Commerce and Economics, Malad (West), Mumbai, which is headed by the guide of the scholar. The data was collected primarily by the candidate himself. The completed questionnaires in Accounts were 190 and in Law were 90. The reliability and validity analysis was carried out on two sets of data.

For data collection, first of all, liasoning with the school administration was done followed by briefing them with other rationale of the proposed study. Briefly the aims and objectives of the survey were explained in simple manner and their permission was obtained. After this, the scholar enquired about the courses that were run in the school. The approximate strength of the students in these respective courses was enquired. This process of liasoning was facilitated by the guide for the present study who is also the esteemed faculty in the same college. Therefore, prior to data collection, these steps enabled the scholar to proceed for the required data collection.

As to any research, data quality is of utmost importance. Hence, the scholar reiterated the objectives of the present study before the students registered under these two courses. The students were sensitized to the importance of the study. The scholar, to the best of his efforts, explained the queries raised by the students. This aroused some curiosity among the students and few queries were in the form of importance and benefits of the study. Therefore, the scholar explained at length and breadth regarding the rationale of the study. The codes under Likert type items were explained to the students. The scholar emphasized that students should respond with truthfulness so as to ensure data quality. The items presented in simple and clear manner were obviously expected to further help in maintaining the accuracy and reliability of the collected data.

Once the scholar felt that the students got convinced as well as clarity about structured questionnaire and were willing to give responses, he distributed the same to the students. For this, prior permission for the data collection with the concerned teachers of the respective courses was also obtained.

3.2.2. Validation of Questionnaire

While conducting any research, utmost care is needed for obtaining precise data which is valid, reliable and able to provide results with greater generalization. These terms have their relative importance in the field of item response theory (IRT). Among the researchers, these terms are sometimes used interchangeably. However, they all address the same common research question that is; what is being measured, how valid is this measurement and did the researcher measure precisely the attributes which he primarily aimed for.

There are different sources of errors in measurement that decreases the validity. Such list is exhaustive, however, four important errors that may inadvertently creep during the process of data collection maybe on the part of Respondent, Situational Factors, the Scholar and the Instrument itself. The first three errors are those errors that can be taken care during the process of data collection itself. For example, the respondent error may occur because of lack of motivation on part of students to express unbiased opinions regarding the courses under which they are registered. To minimize this, as explained earlier under the description of study site, the students were explained about aims & objectives of the study; and understanding of questionnaire; and motivated to express their unbiased response.

The ambiance of the survey and the place where the survey is conducted is equally important to avoid errors on part of situational factors. For example, liasoning with school administrator and resulting cooperation goes in a long way to ensure data quality. Further, ensuring required anonymity on the part of respondents also ensures accuracy in the responses obtained by the students. The initiatives on part of the scholar before the data collection ensured comfort of the students and enabled in minimizing these errors. The error on part of the scholar was nullified by administrating structured questionnaire, wherein these three sources of errors were controlled during the process of data collection.

The probable error arising due to the use of the invalidated structured instrument also needed to be controlled. Hence, the need for validation of the questionnaire arises. The validation was done through rigorous application of statistical techniques.

3.2.3. Techniques of Questionnaire Validation

Reliability deals with accuracy and precision of procedures involved in measurements. In true theoretical sense, validation refers to the extent to which a test measures what we actually wish to measure. There are three basic tests of validity, namely Content Validity, Criterion Related Validity and Construct Validity [8].

Content Validity

The content validity refers to the extent to which a particular instrument covers the topic under study. In the present study, the primary aim was to measure the satisfaction level among college students regarding varying items under various domains of each course; hence the selected items on the basis of extensive review literature and discussions were expected to be adequate enough to fulfill the requirement under the content validity.

Criterion Validity

The criterion validity refers to our ability to accurately predict some outcomes or estimate the existence of some current conditions in the present study. This was fulfilled by providing the description of total items under various domains which will be discussed under subsequent sections.

Construct Validity

The construct validity is regarded as one of the most complex and abstract situation. In a nutshell, a measure is said to possess construct validity if it confirms to predicted correlations with other theoretical propositions. To make it more clearer, the construct validity measures the degree to which responses obtained through particular instrument can be accounted for by the explanatory constructs of sound theory. In the present situation, explanatory constructs of sound theory refer to broad domains under the questionnaire:

- Domain I: At the beginning of the course, I was told about:
- Domain II: During this course, my teachers:
- Domain III: During this course, my college/department:
- Domain IV: This course helped me to:
- Domain V: Overall evaluation

As a matter of fact, if global item (item 15) is found consistent with those on remaining fourteen items, data/ results are expected to be more accurate and reliable. Further, if measurements on our devised scale closely correlate in a predictive way with these domains, we can conclude that there is construct validity. The care on first part was ascertained by running exhaustively the responses of global item (15th item). So we are referring to situation in which the outcome that is overall satisfaction of the courses is predicted precisely in an unbiased manner through preceding domains and their respective items cumulatively.

Statistically two analytical methods were applied for questionnaire validation. The first one is Karl Pearson Correlation Coefficient and the second one is Exploratory Factor Analysis.

1) Karl Pearson Correlation Coefficient

This statistic measures the strength of linear relationship between two quantitative attributes (e.g., x and y) collected on same individuals and is expressed mathematically as [9]:

$$\rho(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)*V(Y)}} \tag{1}$$

$$Cov(X,Y)=\frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right] \tag{2}$$

$$\bar{x} = \frac{1}{n}\left[\sum_{i=1}^{n} x_i\right] \quad (3)$$

$$\bar{y} = \frac{1}{n}\left[\sum_{i=1}^{n} y_i\right] \quad (4)$$

$$Var(X) = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \quad (5)$$

$$Var(Y) = \frac{1}{n}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] \quad (6)$$

The required assumptions involve normality of data and linear relationship between *x* and *y*. The range of this correlation coefficient is ±1. A value of +1 is considered as perfect positive correlation and −1 is considered as perfect negative correlation.

As a rule of thumb, in item response theory, the correlation value of 0.30 or higher in either positive or negative direction is considered as a minimal threshold. However, this value may be expected to be higher for other scientific disciplines like engineering, and management.

The correlation coefficient was estimated by considering data collected for the purpose of reliability and validity on each of the two subjects (Accounts and Law) under the academic program BCOM.

2) Cronbach's Alpha Coefficient

This measures the internal consistency of items under a particular construct. This is considered as a very strong measure of internal consistency in item response theory and is expressed mathematically as [7]:

$$\alpha = \frac{k}{k-1}\left\{1 - \frac{\sum_{i=1}^{n}\sigma_{y_i}^2}{\sigma_x^2}\right\} \quad (7)$$

where,

k = sum of the measure of a quantity

$x = y_1 + y_2 + \dots + y_k$

$\sigma_{y_i}^2$ = variance of the observed total test score

σ_x^2 = the variance of component *i* for the current sample of persons

The theoretical value of this statistic varies from 0 to 1 because it is the ratio of two variances. As a rule of thumb, $\alpha \geq 0.9$ is considered excellent. We can summarize the values of this statistic as follows:

Cronbach's Alpha	Internal Consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Very Good
$0.8 > \alpha \geq 0.7$	Good
$0.7 > \alpha \geq 0.6$	Fair
$0.6 > \alpha \geq 0.5$	Low
$0.5 > \alpha$	Unacceptable

3) Factor Analysis

It is considered as a dimension deduction technique and works on the fundamental statistical principle of Principle Component Analysis. It is a useful technique for investigating variable relationship for complex characteristics such as socio-economic status, dietary patterns, and psychological scales like the structured questionnaire for Likert type item responses under the present study. It enables the researcher to investigate concepts that are not easily measured directly by condensing large number of items into few interpretable underlying factors consisting of sub-sets of considered Likert type items.

The key concept under factor analysis is that multiple observed variables have more or less uniform patterns of responses because they are all associated with a latent construct or factors which cannot be measured directly. The major statistical measures obtained under the Factor Analysis are Factor Loadings, Sample Adequacy, Communalities, Scree plot and the Rotated component matrix [7] [8].

Factor Loadings

The terminology used to express the relationship of each item to the underlying factor is factor loadings. Higher the factor loading the better is the factor. In statistical sense, factor loadings are interpreted like standardized regression coefficients which are obtained by running a regression analysis by considering the items as the response variable and factor as a dependent variable. Although, under explanatory factor analysis, there is no theoretical threshold to label the items under a particular factor good or bad, but in social sciences a measure of 0.30 or high is considered adequate.

Sample Adequacy

Field [10] advocated that in general over 300 respondents for sampling analysis is probably adequate to explore the probability of factor analysis. However, sometimes among the researchers there is a disparity as to precisely answer the question "How many respondents need to be considered for running a factor analysis?". This is a methodological difficulty and is likely to differ from situation to situation.

To overcome the above problem, as an alternative, another statistic for assessing the sample adequacy is the Bartlett's Test of Sphericity, which is also provided under The Factor Analysis. This statistic tests the Null Hypothesis that the correlation matrix is an identity matrix.

An identity matrix is a matrix in which all of the diagonal elements are 1 and all the off diagonal elements are close to 0. We wish to reject this Null Hypothesis because the variable considered for factor analysis is expected to correlate to a reasonable level among them and therefore should be different from 0. KMO and Bartlett's Test measures the adequacy of sample for determining whether or not the responses obtained on a sample are adequate for interpretable findings under factor analysis. Theoretically, Kaiser [11] recommended a value close to 0.5 as minimum, values between 0.7 - 0.8 as acceptable and values above 0.9 as excellent.

Communalities

The table of Communalities exhibits the proportion of variance that can be explained by a concerned item and a value close to 0.5 or higher is considered for further analysis. This is a very exhaustive output under the factor analysis and contains ten headers under the section. These ten headers are component, initial Eigen values total, initial Eigen values % of variance, initial Eigen values cumulative %, extraction sums of squared loadings % of variance, extraction sums of squared cumulative %, rotation of sums of squared loadings total, rotation of sums of squared loadings % of variance, rotation of sums of squared loadings cumulative %.

Scree Plot

It is a graph of Eigen values against all the factors. This graph acts as a guiding principle for determining the number of factors to be retained under factor analysis. The point of flattening of the curve is indicative for number of factors to be retained.

Rotated Component Matrix

The rationale behind consideration of rotated component matrix is to reduce the number of factors on which the variables under the investigation have high loadings. Actually, rotation does not cause distortion in the results under factor analysis, but it helps in better interpretation of results. The rotated component matrix illustrates the factor loadings of all items considered in the analysis and theoretically then can be considered as the values of the regression coefficient as obtained when we run the linear regression analysis [12]. Theoretically, the value of factor loadings should be greater than 0.30 and below this level, the items should be dropped from the analysis and we need to apply Factor Analysis, again to the dataset. Statistically, these factor loadings explain how much proportion of variability in the latent construct of the factor can be accounted by the considered item or variable.

3.3. Material and Methods

The structured questionnaire measuring five domains as explained in earlier section enabled to get raw-data for subsequent validation on three data sets in relation to courses Accounts, Law and both taken together (Accounts + Law). It was screened further through consistency checks on an excel spreadsheet.

The spreadsheet consisted of relevant fields for each of the surveyed students like Serial number, Academic Program Code, Course Code, Specification of Semesters, Courses Codes and Responses on Likert type 15 items. After running the consistency checks and frequency distributions, after ensuring clean data sets, they were subsequently used in required analyses.

Generally, in such a situation there are different analytical techniques that can be applied. Broadly these techniques can be classified broadly under two categories, namely Descriptive Statistics and Inferential Statistics. The Descriptive Statistics focus in summarizing the data through descriptive measures like Mean, Median, Mode, Range, Standard Deviation and Interquartile Range. The Inferential Statistics involve the techniques and analytical methods necessary for deriving desired inferences using the required analytical results. To be more specific, descriptive statistics help in understanding data as well as its distribution and inferential statistics from analytical results help in drawing inferences leading to important clues for policy planners.

In the present study, being focus on validation of the developed questionnaire, the major focus is on inferential statistics using analytical results of above-described validation data sets regarding questionnaire validation. Its objective is to examine how well the items included in the questionnaire confirm expectations regarding the psychometric properties of the tool. The analysis typically begins with analyzing the proportion of students with missing response on each item, and those with complete response to each item. Further, score distribution is explored by calculating the observed range of scores and the proportion of students with worst and best possible score (floor and ceiling effect) on each dimension as an indicator of the extent to which scale captures the range of the underlying dimension.

The major evidence used to explore appropriateness of the considered construct in the proposed questionnaire consists of statistical assessments of three Components-Dimensionality, Reliability and Validity. The dimensionality is mainly concerned with the homogeneity of items. This is assessed using exploratory Factor Analysis (FA) which analyses scores on several items to see if they can be reduced to underlying dimensions. In addition, the FA explains a substantial amount of the variance in the scores. Based on the factor loadings, the researcher needs to decide which items from the scale may be retained or deleted. As such, Principle Component Analysis (PCA) or Exploratory FA are conventionally used approaches in this regard. Further, confirmatory FA can also be used to verify the obtained factor structure of the set of observed items. The reliability of the new questionnaire is tested by examining the internal consistency using Cronbach's α. Further, if follow up data on same scale is obtained, test-retest ability may be analyzed using the Intra-class Correlation Coefficient (ICC). For both Cronbach's α and the ICC, values over 0.75 are considered acceptable. Regarding *Validity*, construct validity is assessed using convergent validity. It refers to the degree to which different items of the same construct are correlated.

Although each item is observed on a Likert type item (1 = "*strongly disagree*"; 2 = "*disagree*"; 3 = "*true sometimes*"; 4 = "*agree*"; 5 = "*strongly agree*"), for the exploratory analysis, it was considered to be measured as on interval scale [9]. Further, it was assumed that the responses to items moderately follow normal distribution. The observed higher magnitude of factor loadings as well as higher communalities under factor analysis may pave the way for optimum sample size required for appropriate factor analysis [13]. To be more specific, under such circumstances, exploratory analysis of CES data even on 90 to 280 students [10] [13] may serve the purpose. Further, value of Kaiser-Meyer-Olkin measure (KMO) closer to 1 may also indicate sampling adequacy regarding distinct and reliable factors under factor analysis. Earlier studies [11] [14] have further recommended that value of KMO below 0.5 suggests either collecting more data or reviewing the inclusion of additional variables. Acceptance of its value greater than 0.5 may further be categorized as: between 0.5 and 0.7 as moderate, between 0.7 and 0.8 as good, between 0.8 and 0.9 as great and above 0.9 as superb.

To begin with, for each data set, the correlation matrix dealing with simple correlation of each item with all other items included in CES questionnaire was examined. From this, one can also visualize the extent of relationship of each item with the respective overall satisfaction item. Practically, the items have to be inter-correlated, but the presence of too high correlation between items may cause difficulties in determining the unique contribution of an item to a factor [2] [10]. The pair of items involving a positive and high to very high correlation may indicate possible presence of multi-collinearity. This may be observed in two steps. First, there will be inter-correlation between the items in case of a significant Bartlett's test of sphercity. Second, the determinant of correlation matrix less than 0.00001 will also reveal that there is multi-collinearity [10]. To overcome this problem of collinearity, as a rule of thumb, items involving a correlation of 0.70 and more may be considered collinear [4]. Under such circumstances, only one of the two collinear items needs to be considered. Keeping in view of the presence of more number of such pairs, a series of exploratory subsets of items need to be generated to carry out further exploratory analysis.

Under each of the considered three data sets, separately for original set of items, as well as each of the possible various exploratory subsets of items, factor analysis involving principal component analysis as extraction method and orthogonal varimax (with Kaiser normalization) as rotation method [15] had to be carried out. Such exploratory analysis guides regarding appropriateness of the items in questionnaire. Every time, as reliability coefficient, Cronbach's alpha for questionnaire was worked out. Likewise, Cronbach's alpha in case of dropping item(s) was also examined. Under factor analysis, different underlying factors (*i.e.*, constructs or components) and their associated items were also identified. One of the possible interpretations of obtaining more or less number of constructs may be that CES questionnaire failed to completely measure quality of course, but did measure some related constructs [10]. Also, each of the observed constructs needs to involve a different set of related items in order to consider them as sub-components of quality of course [10]. Its contradiction may simply be attributed to varying scales of measurements of involved domains in the questionnaire. The reflection of relative merits in terms of reduced problem of colinearity among items; retaining relationship among them; involving fewer extracted factors (*i.e.*, constructs), increased independence of extracted factors; negligible change in overall reliability of questionnaire, and the proportion of total variance being explained by the underlying constructs [2] [4] [10] [11] [13] [14] could be assessed through comparative appraisal of results under exploratory analysis of considered three data sets.

3.4. Results

The assessment of the developed CES questionnaire was carried out using three data sets on two courses: Accounts = 190 students, Law = 90 students, and Accounts + Law = 280 students. The questionnaire consists of 15 items (I1 to I15) involving five domains (D1, D2, D3, D4, and D5). Under presentation as well as description of results, the suffix in the form of Item_Domain (example, I1_D1) refers to the first item under domain 1.

As a first strategy under questionnaire validation, specification of item numbers along with domains was used to depict all possible pair wise correlation coefficients (not listed here) between items across all the three groups. From the array of correlation matrices, as required, it is quite clear that none of the pairs had colinearity problem. For majority of combinations, the correlation coefficient remained mild but significant. More or less, these results remained consistent across three data sets.

To carry out correlation analysis at domain levels, instead of item levels, cumulative scores of the items under specific domains were considered. In other words, domains involving two and more items could generate data on Likert scale. For example, scores under the first dimension involving three Likert type items were likely to be in the range of 3 to 15; scores under the second dimension involving seven Likert type items were likely to be in the range of 7 to 35; and scores under the third dimension involving three Likert type items were likely to be in the range of 3 to 15. However, domains four and five involving only one Likert type item each remained to be Likert type items. The domains level correlation analyses (not listed here) also revealed similar results as those under item specific correlation analysis, except few exceptions of obtaining moderate level correlations.

For each of the three data sets, the comparative results of factor analysis for CES questionnaire are presented (**Tables 1-3** & **Figures 1-3**). These tables embody the respective results [2] [4] [10] [11] [13] [14] related to characteristics desired to assess adequacy of exploratory analysis and also the merits of a questionnaire. They are as: Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy, determinant of correlation matrix (status of colinearity), significance level for Bartlet's test of sphericity (interrelationship of items), number of extracted components (Eigen value > 1) as indicative evidence towards uni-dimensionality of the questionnaire, and % of total variance accounted by the extracted factors.

Under analysis of each data set (**Tables 1-3**), the magnitude of factor loadings (≥0.40) for each item on each extracted factor, and the related extracted communalities reflecting the amount of total variance in each item that can be explained by the retained factor [10] in an analysis were consistently higher. The Factor Loading exhibited that all the fifteen items considered under the questionnaire loaded at least in any one of the four factors in all the three data sets and these values were either close to 0.5 or higher. All the three data sets revealed similar observations. To be more specific, item4_D2, item5_D2, item6_D2, item7_D2, item8_D2 and item9_D2 was loaded on factor one in two data sets (Accounts + Law, Accounts). Item10_D2, item11_D3, item12_D3, item12_D3, item13_D3 and item14_D4 loaded on factor two in the two data sets

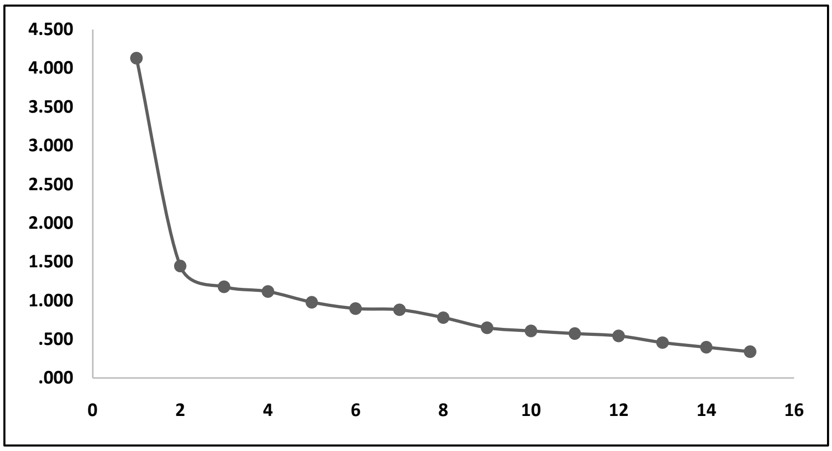

Figure 1. Scree plot for the course Accounts (x-axis component and y-axis Eigen value).

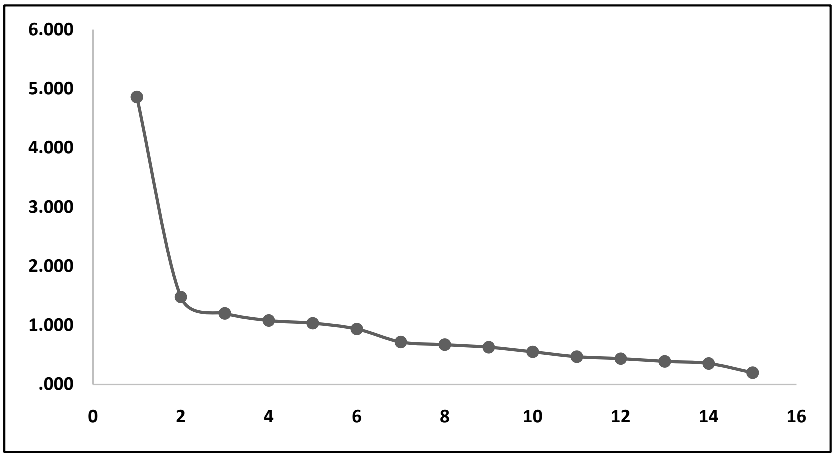

Figure 2. Scree plot for course Law (x-axis component and y-axis Eigen value).

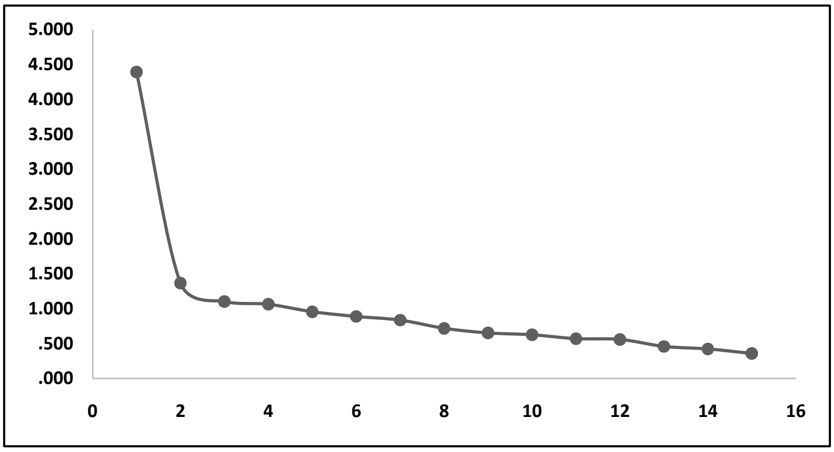

Figure 3. Scree plot for the combined courses Accounts + Law (x-axis component and y-axis Eigen value).

Table 1. Results of the factor analysis for the validation dataset involving the course Accounts.

KMO	0.786			
Bartlett's Test of Sphericity (Value, df, p)	603.044 (105, < 0.001)			
Variance Explained	27.553	9.65	7.871	7.454
Cumulative Variance Explained	27.553	37.202	45.073	52.527
	Factor Loadings			
Items_Domains	Factor1	Factor2	Factor3	Factor4
item1_D1			0.728	
item2_D1	0.33			0.608
item3_D1				0.806
item4_D2	0.537			0.481
item5_D2	0.708			0.412
item6_D2	0.715			
item7_D2	0.657		0.445	
item8_D2	0.604			
item9_D2	0.598	0.312		
item10_D2	0.358	0.653		
item11_D3		0.648	0.521	
item12_D3		0.747		
item13_D3		0.555		0.476
item14_D4	0.341	0.558		0.334
item15_D5	0.518	0.429		0.583

Table 2. Results of the factor analysis for the validation dataset involving the course Law.

KMO	0.749			
Bartlett's Test of Sphericity (Value, df, p)	398.387 (105, < 0.001)			
Variance Explained	32.4	9.853	8.004	7.198
Cumulative Variance Explained	32.4	42.253	50.258	57.456
	Factor Loadings			
Items_Domains	Factor1	Factor2	Factor3	Factor4
item1_D1				0.644
item2_D1	0.429		0.484	0.661
item3_D1	0.499		0.46	0.507
item4_D2			0.843	
item5_D2	0.756			
item6_D2	0.57		0.693	
item7_D2	0.591			0.559
item8_D2	0.367	0.404	0.443	
item9_D2	0.605			0.507
item10_D2	0.713			
item11_D3		0.36		0.71
item12_D3		0.79		0.366
item13_D3		0.861		
item14_D4	0.529	0.448		0.619
item15_D5	0.660	0.467		0.402

Table 3. Results of the factor analysis for the validation dataset involving the combined courses (Accounts + Law).

KMO	0.826			
Bartlett's Test of Sphericity (Value, df, p)	924.803 (105, < 0.001)			
Variance Explained	29.289	9.127	7.366	7.101
Cumulative Variance Explained	29.289	38.416	45.782	52.883
	Factor Loadings			
Items_Domains	Factor1	Factor2	Factor3	Factor4
item1_D1			0.751	
item2_D1	0.489		0.462	0.375
item3_D1				0.865
item4_D2	0.601		0.308	
item5_D2	0.753			
item6_D2	0.639			
item7_D2	0.622		0.521	
item8_D2	0.552			
item9_D2	0.609			
item10_D2	0.501	0.512		
item11_D3		0.521	0.664	
item12_D3		0.75		
item13_D3		0.696		0.322
item14_D4	0.479	0.535	0.339	
item15_D5	0.656	0.447		

that were utilized for questionnaire validation. The global item loaded on factor one (factor loading = 0.656) under Accounts + Law; under factor four for Accounts (factor loading = 0.583); and in factor one for Law (factor loading = 0.660). There were few cross loadings observed. However, their values were relatively lower with respect to factors for which they loaded too highly. This supports the validity of this exploratory analysis [10] [13], even with the available sample size. Also, consistency in Kaiser-meyer-olkin (KMO) measure of sampling adequacy for each data set further strengthens this view. The results across the considered data sets reveal that the questionnaire retained the required attributes. For example, as described earlier, it retained the required interrelationship among the items; it does not involve colinearity problem; and it comparatively involved fewer factors. Also, these factors comparatively involved more or less distinct set of items. The proportion of total variance being explained by the underlying factors was at acceptable range. The Scree plots (**Figures 1-3**) exhibit the tapering of the curves almost when four-factor solutions were considered. Hence, four factor models seemed consistently adequate under the questionnaire validation for each of the three data sets. Furthermore, consideration of four-factor solution also helped in explaining cumulatively more than 50% of the total variance under the considered analysis. The total cumulative variance, explained by considering four factors as a solution, was 52.52, 57.45 and 52.88, respectively.

Cronbach's Alpha, as explained earlier, this helps in assessing the internal consistency of items. In other words, it measures over all reliability of a questionnaire with specific considered items. Also, a sequential approach was adopted while estimating this statistic (not listed here) in each of the three data sets used for validation. Subsequently, this enabled to estimate Cronbach's Alpha considering all items as well as considering remaining items after deletions of individual items one by one. As observed from the results, the minimum and maximum estimates of Cronbach's Alpha were 0.784 - 0.825 for Accounts, 0.822 - 0.841 for Law, and 0.801 - 0.821 for Accounts + Law. Thus, it can easily be deduced that even the lower range was close to 0.8, higher than acceptable level, in each of the three data sets.

To estimate internal consistency of domains, as under correlation analysis, cumulative scores of the items under specific domains were considered. As described earlier, first three domains were on Likert Scale and the remaining two domains as Likert type items. The estimates of Cronbach's Alpha for the questionnaire using its five domains (not listed here) had lower values across three data sets as compared to those considering individual items. To be more specific, considering all domains as well as considering remaining domains after deletions of individual domains one by one, they (minimum and maximum) remained consistently lower across three data sets: Accounts (0.535 - 0.630); Law (0.587 - 0.677) and Accounts + Law (0.552 - 0.641). Intuitively, as obvious, this may theoretically be attributed to changing measurement scales of the domains.

4. Conclusion

The items in the considered CES questionnaire have the overall scientific merits. Under the present study, three different statistical techniques were adopted for questionnaire validation namely Karl Pearson Correlation Coefficient, assessment of internal efficiency with Cronbach's Alpha coefficient and four factor solutions under factor analysis in three data sets. The overall results were quite consistent among them and exhibit good psychometric properties. As such, each of the five aspects of the questionnaire remains intact. As a matter of fact, as obvious, consistency in the results across three data sets suggests clarity of items at the level of students. Hence, the collected data using this questionnaire is more likely to maintain its accuracy and reliability. It suggests that the questionnaire is valid, reliable and measures what it is intended to measure. It may thus be concluded that the considered CES questionnaire (**Appendix**) may be used with its optimal structure and analytical power.

Acknowledgements

The authors are thankful to the editorial board members and reviewers for their

valuable observations to strengthen the article further.

Conflicts of Interest

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

References

[1] Al Rubaish, A., Wosornu, L. and Dwivedi, S.N. (2011) Using Deduction Form Assessment Studies towards Furtherance of the Academic Program: An Empirical Appraisal of Institutional Student Course Evakuation. *iBusiness*, **3**, 220-228. https://doi.org/10.4236/ib.2011.32029

[2] Al Rubaish, A., Wosornu, L. and Dwivedi, S.N. (2012) Item Reduction in "Course Evaluation Survey" Questionnaire: Some Exploratory Analysis and Empirical Evidence. *IJBNST*, **2**, 1-11.

[3] Marsh, H.W. and Roche, L.A. (1997) Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility. *American Psychologist*, **52**, 1187-1197. https://doi.org/10.1037/0003-066X.52.11.1187

[4] Meyers, L.S., Guarino, A. and Gamst, G. (2006) Applied Multivariate Research: Design and Interpretation. Sage Publications Inc., Thousand Oaks, 180-182.

[5] Gravestock, P. and Gregor-Greenleaf, E. (2008) Student Course Evaluations: Research, Models and Trends. Higher Education Quality Council of Ontario, Toronto.

[6] Meads, D.M. and Bentall, R.P. (2008) Rasch Analysis and Item Reduction of the Hypomanic Personality Scale. *Personality and Individual Differences*, **44**, 1772-1783. https://doi.org/10.1016/j.paid.2008.02.009

[7] Nunnally, J.C. (1978) Psychometric Theory. 2nd Edition, McGraw Hill, New York.

[8] Cohen, R.J. and Swerdlik, M.E. (2005) Psychological Testing and Assessment. 6th Edition, McGraw Hill, New York.

[9] Sundaram, K.R., Dwivedi, S.N. and Sreenivas, V. (2015) Medical Statistics: Principles & Methods. Second Edition, Wolters and Kluwer (Health) Pvt. Ltd., New Delhi.

[10] Field, A.P. (2005) Discovering Statistics Using SPSS. 2nd Edition, Sage Publications, London.

[11] Kaiser, H.F. (1974) An Index of Factorial Simplicity. *Psychometrika*, **39**, 31-36. https://doi.org/10.1007/BF02291575

[12] Trotter, R.T. II, Anne, M.B. and Heather, H. (1996) A Method for Systematic Reduction of the Number of Questions in a Network Matrix Questionnaire. *Journal of Quantitative Anthropology*, **6**, 35-47.

[13] Habing, B. (2003) Exploratory Factor Analysis. University of South Carolina, Columbia. http://www.stat.sc.edu/~habing/courses/530EFA.pdf

[14] Hutcheson, G. and Nick, S. (1999) The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models. Sage Publications, Thousand Oaks. https://doi.org/10.4135/9780857028075

[15] Cohen, J., Cohen, P., West, S.G. and Aiken, L.S. (2003) Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. 3rd Edition, Routledge, Abingdon-on-Thames.

Appendix

Academic Program:	**Academic year:**	**Student Year:**
Confidential	**Course Evaluation Survey**	**Do Not Write Your Name/ID**
Semester:	**Department:**	**Course:**

Your feedback is very important to improve the quality of courses further.

Please answer each question as "One Number Only" using the following keys:

1 = Strongly disagree, means the statement is true very rarely

2 = Disagree, means the statement is true poorly

3 = True sometimes, means the statement is true about half the time.

4 = Agree, means the statement is true most of the time

5 = Strongly agree, means the statement is true all or almost all of the time

Questions **Answers**

At the Beginning of the Course, I was told about:

1. The course contents & expected knowledge and skills.

2. Available resources for studies, including faculty-hours and study material.

3. The doable things to succeed, including steps and criteria for assessment.

During this course, my teachers:

4. Conducted the classes as per the course contents.

5. Were interested in what they were teaching.

6. Were available beyond class hours to assist me.

7. Were fully committed towards teaching. (e.g., punctuality, clarity in teaching).

8. Used updated teaching materials. (slides, handouts).

9. Encouraged me to ask questions.

10. Explained the links between this course and other courses in the subject.

During this course, my college/department:

11. Allocated sufficient classes for the listed course contents.

12. Provided the resources (textbooks, library, and computers).

13. Provided required technology (internet facility & programs) to support learning.

This course helped me to:

14. Understand & apply desired knowledge, than simply to memorize.

Overall Evaluation

15. I was satisfied with overall quality of this course

Open Ended Items

16. What did you like most about this course? Two major aspects on priority:

17. What did you dislike most about this course? Two major aspects on priority:

18. What suggestion(s) do you have to improve this course? Two major aspects on priority:

Thank you very much for your contribution.