# Influence of Sample Sizes in Extraction of Employability Skills of Engineering Graduates Using Factor Analysis

M. M. G. V. Shyamalee, K. S. Walgama, W. M. V. S. K. Wickramasinghe, and S. Dissanayake

***Abstract*—Identification of employability skills expected by employers in engineering disciplines is vital for fresh engineering graduates who enter to Sri Lankan work force. Factor Analysis was found to be one of the best methods for identifying these skills, and use of proper sample sizes in such an analysis is essential for reliable results. From this view point, this study establishes a method for investigation of applicable sample sizes in extraction of employability skills of civil engineering graduates. For this purpose, a questionnaire survey was carried out among civil engineering organizations to measure employer expectation on thirteen technical and fifteen nontechnical specific attributes identified as important. These attributes were measured using five point Likert scale based on their importance levels. In the factor analysis process, the communality values of each attribute that represent the degree of participation in the analysis was considered as the key indicators to examine their consistency with effective sample sizes. This study used fifty random samples generated for sample sizes ranging from 25 to 100 with 25 increments and the total sample size 122. From the results, it was found that the communality values showed consistent trend above the sample size 75. Moreover, it should be noted that questionnaire should be properly designed and also respondents should be correctly chosen when this sample size is selected for questionnaire surveys. Finally the finding of this study will help in planning and conducting questionnaire surveys effectively and data analysis with a proper sample sizes in order to make reliable predictions in related studies.**

***Index Terms*—Employability skills, factor analysis, random samples, sample size.**

## I. Introduction

Employability skills of engineering graduates can be explained in terms of specific skills essential for them to perform well in their work place. These skills are much sought after these days by employers. The needs of employability skills differ from country to country, from sector to sector, and from time to time. However, in general, certain qualities such as communication skills, interpersonal skills, integrity, right attitude, problem solving skills, decision making and team building skills can be considered as common employability skills. Nowadays, the higher education institutions around the world play significant role in helping the undergraduates to develop required employability skills by introducing new courses, activities, and training. The excellent academic degrees alone are inadequate for the present day employers who are looking for fresh engineering graduates with competencies or capabilities in generic skills [1]. A number of countries such as United Kingdom, Australia, United States, Canada, Japan, Europe and Malaysia had identified the required employability skills and developed a national framework to clarify the specific attributes of these skills for various positions in different sectors/industries. Furthermore, the national frameworks are changed from time to time in order to address the need of the industry [2].

With the continuous changing nature of the present development of Sri Lanka, employers in the country expect certain skills from fresh engineering graduates. Therefore, having good awareness on required employability skills will help engineering undergraduates to develop those skills to have smooth transition from academic environment to working environment and also to enhance their professional career. In addition, this information will be very useful in curriculum revisions and new course developments. From this view point, identification of employability skills of engineering graduates is vital. An employability skill of an engineering graduate can be identified as a single attribute or a combination of attributes which need to make progress at the work place or to get a good employment. Therefore, employability skills can be extracted through a substantial number of measurable attributes of an engineer. In order to extract these skills, *Factor Analysis* (FA) is found to be an appropriate analysis method which provides underline dimensions as latent factors from significant number of variables measured. The communality values of each variable which estimate in the analysis process represent the degree of association of each variable with all other variables. These values are considered as the key elements to examine the participation of variables in the FA [3]. In order to use FA for extraction of employability skills, awareness on the applicable sample size of the data is important in planning data collection process and to have reliable factor solution from data analysis.

Two types of general recommendations were given by past researchers in terms of minimum sample size in factor analysis. Some studies reveal that absolute sample size ($N$) equals to 100, 200, 250, 300 and 500 are important to provide reliable results. Some other studies highlighted that the ratio of sample size $N$ to number of variables are equal to 3, 6, 5, and 10 are important [4]. However, inconsistencies were observed among the recommendations given from past studies for minimum sample sizes required for FA [5]-[11].

Manuscript received August 7, 2016; revised January 16, 2017.

M. M. G. V. Shyamalee is with the Industrial Training and Career Guidance Unit, Faculty of Engineering, University of Peradeniya, Sri Lanka (e-mail: shyamalee@pdn.ac.lk).

K. S. Walgama is with the Department of Engineering Mathematics, Faculty of Engineering, University of Peradeniya, Sri Lanka (e-mail: kswalgama@pdn.ac.lk).

W. M. V. S. K. Wickramasinghe is with the Department of Civil Engineering, Faculty of Engineering, University of Peradeniya, Sri Lanka (e-mail: vskw@pdn.ac.lk).

S. Dissanayake is with the Department of Civil Engineering, Kansas State University, 2118 Fiedler Hall, Manhattan, KS, 66506, USA (e-mail: sunanda@ksu.edu).

doi: 10.18178/ijiet.2017.7.11.981

Later, it was found that general rules of the minimum sample size are not valid and not useful and also it is hard to say whether absolute sample size or ration of sample size to number of variables is important [12], [13]. Further, it was found that sample size depends on communality of the variables which shows association of each variable with all other variables [14]. Therefore, this study focused on the investigation of applicable sample sizes in extraction of employability skills of civil engineering graduates through communality values estimated using factor analysis.

## II. Methodology

### A. Data Collection

Data collection was conducted through a questionnaire survey carried out among employers of civil engineering organizations where the fresh graduates are typically recruited. The questionnaire measured importance of twenty eight attributes covering 13 and 15 attributes under technical and nontechnical skills, respectively as presented in Table I. These attributes were identified using the Accreditation Manual of the Institute of Engineers, Sri Lanka (IESL) which complies with the Washington Accord requirements, and also from relevant literature [15]. Each attribute was measured using a five–point Likert scale ranging from 1 to 5 where 1 and 5 indicate "Not Important" and "Highly Important", respectively. Scales 2, 3, and 4 respectively refer "Fairly Important", "Important", and "Very Important".

TABLE I: Attributes Included in the Questionnaire

| No. | Attribute |
|---|---|
| | **Technical Skills** |
| 1 | Science Fundamentals |
| 2 | Engineering Fundamentals and Application |
| 3 | Probability and Statistics |
| 4 | Computer Science and Technology |
| 5 | Engineering Practice |
| 6 | Measurement Systems |
| 7 | Technical Standards |
| 8 | Specifications and Inspection Standards |
| 9 | Testing Practices |
| 10 | Environmental Constraints |
| 11 | Code of Ethics |
| 12 | Proficiency Standards |
| 13 | Design Skills |
| | **Nontechnical Skills** |
| 14 | Logical Thinking |
| 15 | Problem Solving Skills |
| 16 | Verbal Communication Skills |
| 17 | Written Communication Skills |
| 18 | Presentation Skills |
| 19 | Team Work |
| 20 | Competence |
| 21 | Integrity |
| 22 | Commitment |
| 23 | Tolerance |
| 24 | Flexibility |
| 25 | Reliability |
| 26 | Conscientiousness |
| 27 | Punctuality |
| 28 | Approachability |

### B. Theoretical Background of Factor Analysis

The variation of communality values which is estimated in the process of FA with different sizes of data samples is the main focus of this study. The size of the data sample is the number of respondents $n$ for the questionnaire. With this information, a data set ($\boldsymbol{X}$) considering responses on $p$ attributes of $n$ respondents can be placed in the matrix form as shown in (1) and (2).

$$\boldsymbol{X}_{\boldsymbol{n}\times\boldsymbol{p}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \tag{1}$$

$$\boldsymbol{X}_{\boldsymbol{n}\times\boldsymbol{p}} = [\boldsymbol{X_1} \quad \boldsymbol{X_2} \quad \cdots \quad \boldsymbol{X_j} \quad \cdots \quad \boldsymbol{X_p}] \tag{2}$$

where $x_{ij}$ is the response given by $i^{\text{th}}$ respondent on the $j^{\text{th}}$ attribute and $\boldsymbol{X}_j$ represents $n \times p$ column vector having mean $\mu_j$ that gives $n$ responses on the $j^{th}$ attribute. The concept of the factor model includes that the deviation of $X_j$ from its mean $\mu_j$ is linearly dependent upon $m$ number of unobservable random factors $\boldsymbol{F_1}, \boldsymbol{F_2}, \ldots, \boldsymbol{F_m}$ which are called common factors, and the error terms $\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}, \ldots \boldsymbol{\epsilon_j}, \ldots \boldsymbol{\epsilon_p}$, as shown in (3) [16]. The error $\boldsymbol{\epsilon_p}$ which is due to additional variation is sometimes called as specific factors or unique factors.

$$\begin{aligned} \boldsymbol{X_1} - \mu_1 &= l_{11}\boldsymbol{F_1} + l_{12}\boldsymbol{F_2} + \cdots + l_{1m}\boldsymbol{F_m} + \boldsymbol{\epsilon_1} \\ \boldsymbol{X_2} - \mu_2 &= l_{21}\boldsymbol{F_1} + l_{22}\boldsymbol{F_2} + \cdots + l_{2m}\boldsymbol{F_m} + \boldsymbol{\epsilon_2} \\ &\vdots \qquad\qquad \vdots \\ \boldsymbol{X_j} - \mu_j &= l_{j1}\boldsymbol{F_1} + l_{j2}\boldsymbol{F_2} + \cdots + l_{jm}\boldsymbol{F_m} + \boldsymbol{\epsilon_j} \\ &\vdots \qquad\qquad \vdots \\ \boldsymbol{X_p} - \mu_p &= l_{p1}\boldsymbol{F_1} + l_{p2}\boldsymbol{F_2} + \cdots + l_{pm}\boldsymbol{F_m} + \boldsymbol{\epsilon_p} \end{aligned} \tag{3}$$

This can be written in matrix form as given in (4),

$$\boldsymbol{X}_{(p\times1)} - \boldsymbol{\mu}_{(p\times1)} = \boldsymbol{L}_{(p\times m)}\boldsymbol{F}_{(m\times1)} + \boldsymbol{\epsilon}_{(p\times1)} \tag{4}$$

The coefficient $l_{im}$ is called the loading of the $j^{\text{th}}$ attribute on the $m^{\text{th}}$ factor and the matrix $\boldsymbol{L}$ is the matrix of factor loadings. Both $\boldsymbol{F}$ and $\boldsymbol{\epsilon}$ are unobservable random vectors and satisfy the following three conditions.

1) $\boldsymbol{F}$ and $\boldsymbol{\epsilon}$ are independent i.e.: $Cov\,(\boldsymbol{\epsilon}, \boldsymbol{F}) = E(\boldsymbol{\epsilon F}') = 0_{(pxm)}$ where $Cov$ = Covariance matrix
2) $E(\boldsymbol{F}) = 0_{(mx1)}$, $Cov(F) = E[\boldsymbol{FF'}] = \boldsymbol{I}_{(mxm)}$
3) $E(\boldsymbol{\epsilon})=0_{(px1)}$, $Cov(\boldsymbol{\epsilon}) = E[\boldsymbol{\epsilon\epsilon}'] = \boldsymbol{\Psi}_{pxp}$, where $\boldsymbol{\Psi}$ is a diagonal matrix.

The factor model implies the covariance structure of $\boldsymbol{X}$ and ($\boldsymbol{X}$, $\boldsymbol{F}$) [16] as shown in (5a) and 5(b)

Covariance matrix ($\boldsymbol{\Sigma}$) of $\boldsymbol{X}$

$$\begin{aligned} \boldsymbol{\Sigma} &= Cov\,(\boldsymbol{X}) = E(\boldsymbol{X} - \boldsymbol{\mu})\,(\boldsymbol{X} - \boldsymbol{\mu})' \\ &= \boldsymbol{LF}(\boldsymbol{LF})' + \boldsymbol{\epsilon}\,(\boldsymbol{LF})' + \boldsymbol{LF}\,\boldsymbol{\epsilon}' + \boldsymbol{\epsilon\epsilon}'\text{mmmmmmm} \\ &= \boldsymbol{LL}' + \boldsymbol{\Psi} \end{aligned} \tag{5a}$$

$$\begin{aligned} Cov\,(\boldsymbol{X},\boldsymbol{F}) &= E(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{F}' \\ &= E(\boldsymbol{LF}+\boldsymbol{\epsilon})\,\boldsymbol{F}' \\ &= \boldsymbol{L}E(\boldsymbol{FF}') + E(\boldsymbol{\epsilon F}') \end{aligned} \tag{5b}$$

As a summary, Factor model can be written as given in (6)

$$Cov(\boldsymbol{X})=\boldsymbol{LL}'+\boldsymbol{\Psi}, \text{ or } \; Var(\boldsymbol{X_i}) = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \boldsymbol{\varphi_i} \tag{6}$$

$$Cov\,(\boldsymbol{X},\boldsymbol{F}) = \boldsymbol{L} \quad \text{or} \quad Cov(\boldsymbol{X_i},\boldsymbol{F_j})=l_{ij}$$

The portion of variance of $i^{th}$ attribute $Var\,(\boldsymbol{X_i})$ contributed by the $m$ common factors is called the communality of $i^{th}$ attribute and the portion of variance of specific factor is called specific variance. Therefore, total variance of $i^{th}$ variable can be given as in (7)

$$\underbrace{Var(\boldsymbol{X_i})}_{\text{Total variance } (\sigma_{ii})} = \underbrace{l_{i1}^2 + l_{i2}^2 + \cdots + l_{ij}^2 + \cdots + l_{im}^2}_{\text{Communality } (h_i^2)} + \underbrace{\boldsymbol{\varphi_i}}_{\text{Specific variance}} \tag{7}$$

Then, above relationship can be written as in (8)

$$\boldsymbol{\sigma_{ii}} = \boldsymbol{h_i^2} + \boldsymbol{\varphi_i} \tag{8}$$

For this model, correlation relation of attributes which is the standardized version of covariance relation can be used since it is easily interpretable with data than covariance relation. Then, the above relationship can be written as in (9).

$$\boldsymbol{1} = \boldsymbol{h_i^{*2}} + \boldsymbol{\varphi_i^*} \qquad \text{where } \; \boldsymbol{\varphi_i^*} = \boldsymbol{1/r^{ii}} \tag{9}$$

$\boldsymbol{h_i^{*2}}$ - Initial communality estimates
$\boldsymbol{\varphi_i^*}$ - Initial specific variance estimates
$\boldsymbol{r^{ii}}$ - $i^{th}$ diagonal element of inverse of correlation matrix $\boldsymbol{R}$

### C. Data Analyses

Common Factor Analysis (*CFA*) uses reduced correlation matrix while Principal Component Analysis (*PCA*) considers correlation matrix ($\boldsymbol{R}$) without any reduction in the variance as shown in (10). *CFA* is the extension of *PCA* since it uses *PCA* to predefine the number of factors.

$$\boldsymbol{R_{p\times p}} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1i} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2i} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{1i} & r_{2i} & \cdots & 1 & \cdots & r_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & r_{ip} & \cdots & 1 \end{bmatrix} \tag{10}$$

where $r_{ij}$ is correlation coefficient between $i^{th}$ and $j^{th}$ attribute.

Among the different methods available, the eigenvalues of $\boldsymbol{R}$ that represent the amount of variance associated with factors are used to select important factors. For this purpose, $\boldsymbol{R}$ should be positive definite matrix such that

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

Then, it can be expressed in terms of its eigenvalues ($\lambda_i$) and associated normalized eigenvectors ($e_i$). Among the available methods *Latent root criterion* is used to determine the factors. This method uses the variance of an important factor ($\lambda_i$ ) which explains at least unit variance (variance of single variable) to retain in the factor structure. Thus, only the factors having eigenvalue greater than one are considered for the number of significant factors ($m$) and all values less than one are considered to be insignificant factors ($p$-$m$) and subsequently discarded [17].

After defining number of factors, the analysis process uses reduced correlation matrix ($\boldsymbol{R_r}$) shown in (11) in order to estimate the loading of each attribute of selected factors.

$$\boldsymbol{R_r} = \begin{bmatrix} h_1^{*2} & r_{12} & \cdots & r_{1i} & \cdots & r_{1p} \\ r_{12} & h_2^{*2} & \cdots & r_{2i} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{1i} & r_{2i} & \cdots & h_i^{*2} & \cdots & r_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & r_{ip} & \cdots & h_p^{*2} \end{bmatrix} \tag{11}$$

The matrix $\boldsymbol{R_r}$ can be expanded in terms of its $p$ number of eigenvalues and $\boldsymbol{R_r^*}$ consider $m$ number of defined eigenvalues ($\lambda_m^*$) and associated normalized eigenvectors ($e_m^*$) in order to estimate the factor loadings of variables as given in (12).

$$\begin{aligned} \boldsymbol{R^*_{r(p\times p)}} = {} & \lambda_1^* e^*_{1(p\times 1)} e^{*\prime}_{1(1\times p)} + \lambda_2^* e^*_{2(p\times 1)} e^{*\prime}_{2(1\times p)} + \cdots \\ & + \lambda_m^* e^*_{m(p\times 1)} e^{*\prime}_{m(1\times p)} \end{aligned}$$

$$\boldsymbol{R^*_{r(p\times p)}} = \begin{bmatrix} \sqrt{\lambda_1^*} e_1^* & \sqrt{\lambda_2^*} e_2^* & \cdots & \sqrt{\lambda_m^*} e_m^* \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1^*} e_1^{*\prime} \\ \sqrt{\lambda_2^*} e_2^{*\prime} \\ \vdots \\ \sqrt{\lambda_m^*} e_m^{*\prime} \end{bmatrix}$$

$$\boldsymbol{R^*_{r(p\times p)}} = \boldsymbol{L^*_{(p\times m)}}\, \boldsymbol{L^{*\prime}_{(m\times p)}}$$

Thus, loading matrix $\boldsymbol{L^*_{(p\times m)}}$becomes

$$\boldsymbol{L^*_{(p\times m)}} = \begin{bmatrix} \sqrt{\lambda_1^*} e_{11}^* & \sqrt{\lambda_2^*} e_{12}^* & \cdots & \sqrt{\lambda_m^*} e_{1m}^* \\ \sqrt{\lambda_1^*} e_{21}^* & \sqrt{\lambda_2^*} e_{22}^* & \cdots & \sqrt{\lambda_m^*} e_{2m}^* \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\lambda_1^*} e_{p1}^* & \sqrt{\lambda_2^*} e_{p2}^* & \cdots & \sqrt{\lambda_m^*} e_{pm}^* \end{bmatrix}$$

$$\boldsymbol{L^*_{(p\times m)}} = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{bmatrix} \tag{12}$$

Then, communality of each attribute ( $\boldsymbol{h_i^{**2}}$) can be estimated through factor loadings as given in (13).

$$\boldsymbol{h_i^{**2}} = l_{i1}^2 + l_{i2}^2 + \cdots + l_{ij}^2 + \cdots + l_{im}^2 \tag{13}$$

This study uses different sample sizes ($n$) such as 25, 50, 75, 100 and 122. From the total sample 122, fifty (50) random

samples were generated for each sample size in order to conduct the factor analysis based on these samples. Factor Analyses, employing Pearson correlation matrix and assuming that the data are continuous, were conducted for attributes separately under technical and nontechnical skills. In this analysis, the communality values estimated for each random sample within a particular sample size was used to get the mean communality values for each sample size through probability distribution. Variations of mean communality values with sample sizes were used to establish the applicable sample sizes.

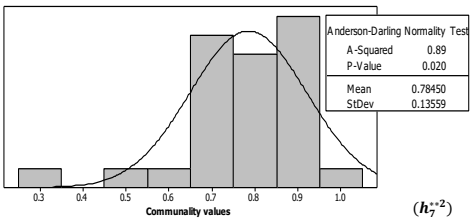
## III. Results and Discussion

### A. Profile of the Data Sample

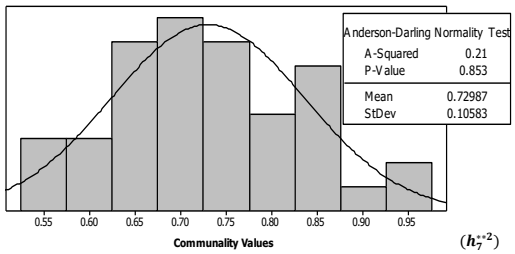
A total of 122 employers of civil engineering organizations responded to the questionnaire. Among them, 89% of respondents were professional engineers (above senior engineer level) who are involved in the recruitment of engineers and the management of their organizations. Further, the organizations which participated in the data collection process were from different work sectors such as Roads (26%), Buildings (38%), Bridges (23%) and others (13%). Among these work sectors, respondents were from different work nature such as construction work (35%), maintenance work (28%), civil engineering design work (32%) and others (5%). The responded organizations were 45%, 20% and 34% respectively from Government, Semi-Government and Non-Government organizations.

### B. Communality Estimates

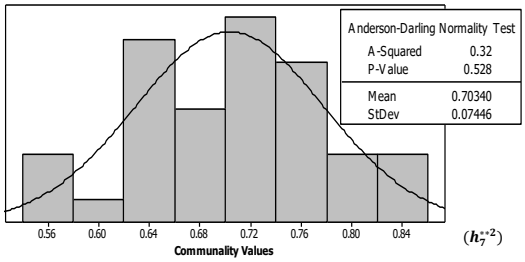
The communality values of each attribute were obtained for 50 random samples generated for each sample size. A better representative data sample provides communality values ranging from 0 to 1. Mean communality values and its standard deviation for respective sample sizes were obtained through probability distribution of the computed communality values of random samples and this mean value was considered as the communality value for the particular sample size. The standard deviation indicates that the amount of deviation of the communality values computed from random samples from the mean communality. Anderson-Darling Test was performed to determine whether these communality values are normally distributed. This behavior is indicated by its statistic value and *p* values. According to its hypothesis, if *p* value is greater than 0.05, communality values are normally distributed and the mean communality value represents the sample size. The probability distribution plots obtained for the sample sizes 25, 50, 75 and 100 for attribute 7 "Technical Standards" listed under Technical skills are presented as a representative case as shown in Fig. 1.



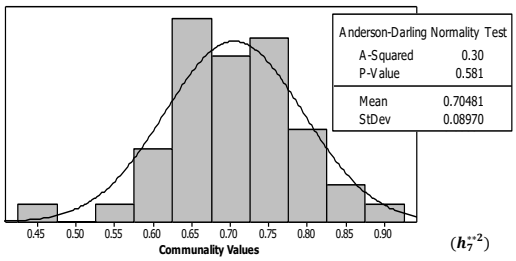

(a) Sample Size *n* = 25




(b) Sample Size *n* = 50




(c) Sample Size *n* = 75




(d) Sample Size *n* = 100

Fig. 1. Distribution of communality values for the attribute of "Technical Standards" for different sample sizes.

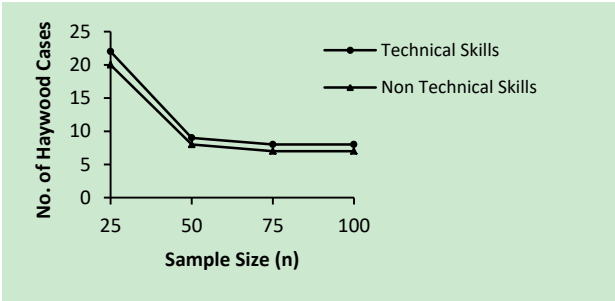



Fig. 2. Variation of Haywood cases with different sample sizes.

These plots indicate that the mean communality values (with standard deviation) for the sample sizes 25, 50, 75 and 100 respectively are 0.78(0.13), 0.73(0.11), 0.71(0.07) and 071(0.09). Since Anderson Darling test indicates that the *p* value is less than 0.05, mean communality value of sample size 25 is not reliable. The *p* value of the sample size 50 is higher than that of other samples. The standard deviations of sample sizes 75 and 100 indicate narrow spread of communality values from its mean than the sample sizes of 25 and 50. Considering these results, the mean communality value computed for sample size 75 and 100 are much reliable than those of the sample sizes of 25 and 50. Further, it is clear that the communality values become consistent with increasing sample size.

When factor analysis is conducted for 50 random samples, in some random samples the values of communalities did not converge. This happened due to one or more communality values of some attributes are nearly equal to one or above one (over-determination). This is called *Haywood Case* and this happens because of the presence of highly correlated attributes. This can be avoided by conducting appropriate

tests prior to factor analysis. With the increase of sample size, this situation was gradually decreased as shown in Fig. 2.

The communality values obtained for attributes under technical and nontechnical skills for all the sample sizes considered are presented in Table II and Table III. In addition, the communality values of sample size 122 (total sample) were also estimated and included in the Table II and Table III to compare the values with the other sample sizes. From these results, it was found that the communality values gradually decrease and show consistent pattern with the increase of sample size.

TABLE II: COMUNALITY VALUES OF ATTRIBUTES (TECHNICAL SKILLS) FOR DIFFERENT SAMPLE SIZES

| Attributes No. | Mean Communality values and related Parameters | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sample Size 25 | | | Sample Size 50 | | | Sample Size 75 | | | Sample Size 100 | | | Sample Size 122 |
| | $h_{im}^{**2}$ | *SD* | *p* | $h_{im}^{**2}$ | *SD* | *p* | $h_{im}^{**2}$ | *SD* | *p* | $h_{im}^{**2}$ | *SD* | *p* | $h_i^{**2}$ |
| 1 | 0.55 | 0.21 | 0.59 | 0.47 | 0.18 | 0.02 | 0.45 | 0.19 | 0.66 | 0.44 | 0.20 | 0.01 | 0.35 |
| 2 | 0.66 | 0.16 | 0.77 | 0.60 | 0.17 | 0.01 | 0.55 | 0.16 | 0.77 | 0.52 | 0.15 | 0.03 | 0.60 |
| 3 | 0.53 | 0.14 | 0.51 | 0.54 | 0.17 | 0.32 | 0.53 | 0.16 | 0.93 | 0.50 | 0.14 | 0.55 | 0.50 |
| 4 | 0.48 | 0.16 | 0.54 | 0.44 | 0.16 | 0.01 | 0.41 | 0.13 | 0.29 | 0.37 | 0.11 | 0.80 | 0.36 |
| 5 | 0.57 | 0.17 | 0.65 | 0.47 | 0.14 | 0.50 | 0.48 | 0.16 | 0.01 | 0.43 | 0.12 | 0.01 | 0.35 |
| 6 | 0.70 | 0.15 | 0.88 | 0.61 | 0.11 | 0.18 | 0.54 | 0.10 | 0.72 | 0.52 | 0.10 | 0.02 | 0.48 |
| 7 | 0.78 | 0.14 | 0.89 | 0.73 | 0.11 | 0.85 | 0.70 | 0.07 | 0.53 | 0.71 | 0.09 | 0.58 | 0.71 |
| 8 | 0.77 | 0.15 | 0.06 | 0.72 | 0.10 | 0.28 | 0.73 | 0.10 | 0.19 | 0.70 | 0.09 | 0.17 | 0.65 |
| 9 | 0.72 | 0.14 | 0.15 | 0.66 | 0.13 | 0.69 | 0.69 | 0.10 | 0.66 | 0.67 | 0.11 | 0.63 | 0.56 |
| 10 | 0.70 | 0.14 | 0.59 | 0.67 | 0.15 | 0.43 | 0.62 | 0.10 | 0.36 | 0.60 | 0.09 | 0.04 | 0.64 |
| 11 | 0.66 | 0.11 | 0.09 | 0.59 | 0.12 | 0.88 | 0.59 | 0.11 | 0.26 | 0.54 | 0.10 | 0.63 | 0.50 |
| 12 | 0.69 | 0.17 | 0.06 | 0.64 | 0.13 | 0.38 | 0.59 | 0.11 | 0.87 | 0.58 | 0.10 | 0.36 | 0.58 |
| 13 | 0.62 | 0.15 | 0.47 | 0.51 | 0.17 | 0.54 | 0.50 | 0.17 | 0.09 | 0.50 | 0.14 | 0.47 | 0.45 |

$h_{im}^{**2}$ – Mean Communality value estimated from probability distribution using communalities of 50 random samples
$h_i^{**2}$ – Communality value estimated from single sample
*SD* – Standard Deviation
*p* – *p*-value at 95% confidence interval

TABLE III: COMUNALITY VALUES OF ATTRIBUTES (NONTECHNICAL SKILLS) FOR DIFFERENT SAMPLE SIZES

| Attributes No. | Mean Communality values and related Parameters | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sample Size 25 | | | Sample Size 50 | | | Sample Size 75 | | | Sample Size 100 | | | Sample Size 122 |
| | $h_{im}^{**2}$ | *SD* | *p* | $h_{im}^{**2}$ | *SD* | *p* | $h_{im}^{**2}$ | *SD* | *p* | $h_{im}^{**2}$ | *SD* | *p* | $h_i^{**2}$ |
| 15 | 0.67 | 0.16 | 0.64 | 0.66 | 0.12 | 0.77 | 0.64 | 0.10 | 0.27 | 0.61 | 0.10 | 0.36 | 0.60 |
| 16 | 0.67 | 0.13 | 0.55 | 0.64 | 0.11 | 0.50 | 0.59 | 0.10 | 0.56 | 0.61 | 0.09 | 0.55 | 0.57 |
| 17 | 0.83 | 0.11 | 0.01 | 0.79 | 0.10 | 0.12 | 0.78 | 0.10 | 0.95 | 0.79 | 0.10 | 0.81 | 0.81 |
| 18 | 0.70 | 0.15 | 0.66 | 0.65 | 0.11 | 0.22 | 0.62 | 0.10 | 0.19 | 0.61 | 0.10 | 0.88 | 0.60 |
| 19 | 0.63 | 0.14 | 0.70 | 0.53 | 0.11 | 0.51 | 0.53 | 0.06 | 0.74 | 0.49 | 0.07 | 0.75 | 0.49 |
| 20 | 0.76 | 0.10 | 0.73 | 0.69 | 0.08 | 0.42 | 0.69 | 0.08 | 0.37 | 0.66 | 0.06 | 0.47 | 0.65 |
| 21 | 0.77 | 0.14 | 0.35 | 0.67 | 0.10 | 0.03 | 0.68 | 0.09 | 0.50 | 0.65 | 0.11 | 0.20 | 0.57 |
| 22 | 0.70 | 0.14 | 0.92 | 0.59 | 0.12 | 0.98 | 0.64 | 0.11 | 0.19 | 0.60 | 0.09 | 0.54 | 0.50 |
| 23 | 0.75 | 0.12 | 0.82 | 0.66 | 0.09 | 0.54 | 0.64 | 0.07 | 0.87 | 0.64 | 0.07 | 0.97 | 0.60 |
| 24 | 0.72 | 0.10 | 0.27 | 0.68 | 0.11 | 0.21 | 0.66 | 0.07 | 0.16 | 0.64 | 0.07 | 0.79 | 0.57 |
| 25 | 0.77 | 0.10 | 0.13 | 0.63 | 0.13 | 0.06 | 0.60 | 0.10 | 0.03 | 0.59 | 0.10 | 0.03 | 0.50 |
| 26 | 0.77 | 0.09 | 0.12 | 0.75 | 0.08 | 0.73 | 0.76 | 0.05 | 0.69 | 0.77 | 0.06 | 0.53 | 0.78 |
| 27 | 0.65 | 0.15 | 0.26 | 0.57 | 0.14 | 0.68 | 0.53 | 0.11 | 0.13 | 0.57 | 0.12 | 0.04 | 0.48 |
| 28 | 0.67 | 0.09 | 0.84 | 0.59 | 0.12 | 0.05 | 0.54 | 0.08 | 0.94 | 0.52 | 0.08 | 0.51 | 0.49 |

$h_{im}^{**2}$ – Mean Communality value estimated from probability distribution using communalities of 50 random samples
$h_i^{**2}$ – Communality value estimated from single sample
*SD* – Standard Deviation
*p*– *p*-value at 95% confidence interval

From Fig. 3 and Fig. 4 above, it was found that the communality values of attributes for both technical and nontechnical skills show a consistent trend above the sample size of 75. In addition, one-way analysis of variance (ANOVA) conducted for communality values of sample sizes 75, 100 and 122 showed no significant differences. Therefore, the analysis should be performed for sample sizes greater than 75 to have reliable and consistent prediction of

expected employability skills of engineering graduates.

In addition, number of important factors that are explained by each random sample of the fifty random samples generated for each sample size was examined. As a result, factor structures containing 2, 3, 4 and 5 factors were found for sample sizes of 25 and 50 in both technical and nontechnical skills. From these factor structures, 2- and 5-factor structures were found in less than 10% of the samples and majority of the samples gave 4-(64%) and 3-(60%) factor structures respectively for sample size 25 and 50. In sample sizes 75 and 100, 3- and 4- factor structures were present and the percentages of samples having 3- and 4- factor structures were same in both sample sizes. For technical skills, 62% of samples showed 3-factor structures while the remaining 38% were 4-factor structures. Similarly for nontechnical skills, 68% of samples showed 3-factor structure and 32% was observed 4-factor structure. Further, a 3- factor structure was also observed in sample size of 122 for both skills. Hence, it is clear that the majority of samples showed 3-factor structures and it was observed that the percentage of samples having 3-factor structures was consistent above sample size of 75. Further, the effect of percentage of variance which shows the strength of 3-factor structures was examined with the increase of sample size. These strengths were 72%, 67%, 63% and 62% respectively for sample size 25, 50, 75 and 100 in technical skills. For total sample size 122, this strength was 61%. Similarly for nontechnical skills, these figures were 76%, 72%, 69% and 69% for same sample sizes and it was 68% for total sample size 122. Hence, it is clear that the strength of 3- factor structure showed a consistent pattern above sample size 75. Considering these results, in general, the minimum sample size for extraction of employability skills can be suggested as five times the number of attributes for having reliable results.

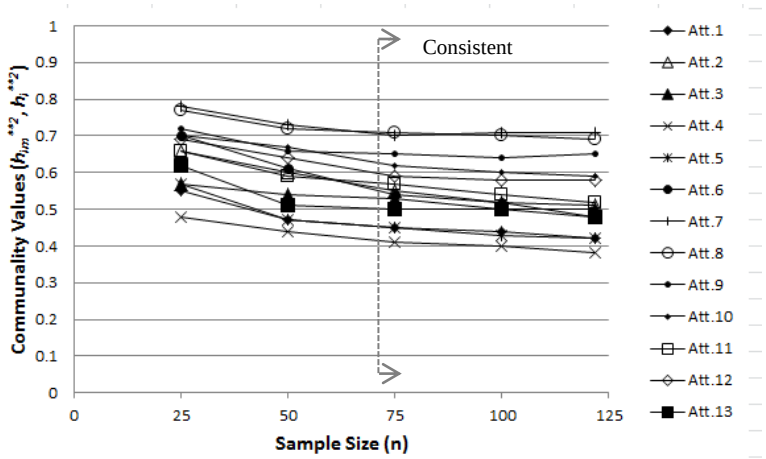



Fig. 3. Variation of communality values of attributes (technical skills) for different sample sizes.

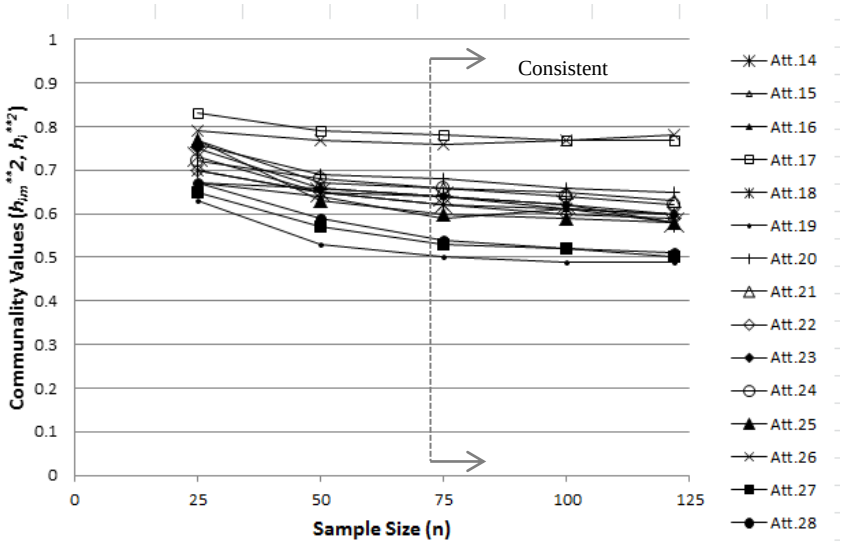



Fig. 4. Variation of communality values of attributes (nontechnical skills) for different sample sizes.

## IV. Conclusion

The extraction of important employability skills can be done using Factor Analysis. In this study, the effect of sample size on the reliability of Factor Analysis results was investigated by examining the consistence of communality values of attributes. The number of attributes considered here was 28 and the total sample size was 122. From the total sample, 50 random samples having sample sizes of 25, 50, 75, and 100 were generated for the analysis. The analyses conducted for each and every sample size including 122 total sample size revealed that consistence community values could be observed from samples with size 75 and above. That means, the minimum sample size to have reliable communality values is 75. The consistency depends on the factors such as the minimum number of Haywood cases and

their consistency with the increase of sample sizes, and the consistent number of important factors and their strength in the factor structure (percentage of variance). Therefore, it is important to design questionnaires properly to collect required data. For example, similar type of questions should not include in the questionnaire in order to avoid receiving similar responses. And also, respondents should be selected within the target population to avoid mislead responses and to have consistent result on number of important factors and their contribution to the factor structure. Then, minimum number of sample size can be selected as five times the number of attributes to plan and initiate the data collection process in similar studies. During the data collection period, this methodology can be used to decide the appropriate sample size to minimize the cost and time involved in further data collection. Therefore, the results and discussion presented here will help in planning and conducting questionnaire surveys and analyzing the data with a proper sample size in order to make reliable predictions in related studies.

## REFERENCES


[1] A. Zaharim, Y. M. Yusoff, M. Z. Omar, A. Mohamed, and N. Muhamad, "The comparison on priority engineering employability skills," *International Journal of Engineering and Technology*, vol. 7, no. 2, pp. 61-74, 2010.

[2] (2002). *Employability Skills For Australian Industry: Literature Review And Framework Development, a report by the Australian Council for Educational Research to the Business Council of Australia and the Australian Chamber of Commerce and Industry*. ACER. Department of Education, Science and Training, Canberra. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5351&rep=rep1&type=pdf

[3] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 7th ed. New Jersey, USA: Prentice Hall, 2005.

[4] P. Kline, *Psychometrics and Psychology*, London: Acaderric Press, 1979.

[5] L. Hatcher, *A Step-by-Step Approach to Using the SAS® System for Factor Analysis and Structural Equation Modeling*, Cary, NC: SAS Institute, Inc., 1994.

[6] G. Hutcheson and N. Sofroniou, *The Multivariate Social Scientist: Introductory Statistics using Generalized Linear Models*, Thousand Oaks, CA: Sage Publications, 1999.

[7] D. G. Garson, "Factor analysis: Statnotes," *From North Carolina State University Public Administration Program, 2008*.

[8] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan, "Evaluating the use of exploratory factor analysis in psychological research," *Psychological Methods*, vol. 4, pp. 272-299, 1999.

[9] W. A. Arrindell and J. Ende, "An empirical test of the utility of the observations-to-variables ratio in factor and components analysis," *Applied Psychological Measurement*, vol. 9, pp. 165-178, 1985.

[10] R. B. Cattell, *The Scientific Use of Factor Analysis*, New York: Plenum, 1978.

[11] R. L. Gorsuch, *Factor Analysis*, 2nd ed. Hillsdale, NJ: Erlbaum, 1983.

[12] B. S. Everitt, "Multivariate analysis: The need for data, and other problems," *British Journal of Psychiatry*, vol. 126, pp. 237-240, 1975.

[13] R. C. MacCallum, K. F. Widaman, S. Zhang, and S. Hong, "Sample size in factor analysis," *Psychological Methods*, vol. 4, pp. 84-99, 1999.

[14] K. J. Preacher and R. C. MacCallum, "Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes," *Behavior Genetics*, vol. 32, pp. 153-161, 2002.

[15] *Engineering Program Accreditation Manual*, Institute of Engineers, Sri Lanka, 2014.

[16] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6rd ed. New Jersey, USA: Prentice Hall, 2007.

[17] S. Green and N. Salkind, *Using SPSS for Windows and Machintos: Analyzing and Understanding Data*, Upper Saddle River, NJ: Prentice Hall, 2005

**M. M. G. V. Shyamalee** received her master of engineering degree in the field of traffic engineering from Nagoya University, Japan in 2002. At present, she is a PhD candidate and a lecturer attached to the Industrial Training and Career Guidance Unit, Faculty of Engineering, University of Peradeniya, Sri Lanka. Her research interest is to establish a methodology for evaluation of industrial training of engineering undergraduates. Ms. Shyamalee is an associate member of Institute of Engineers, Sri Lanka.