ST841 Statistical Practice I

# Consulting TAA

| Author: | Contact Email: |
|---|---|
| Drew Hollis | anhollis@ncsu.edu |
| Xinyu Zhang | xzhang97@ncsu.edu |
| Qiang Heng | qheng@ncsu.edu |

October 20, 2020

## Contents

# 1 Introduction

For this project, we cooperate with Dr. Al Chen in the Department of Accounting to conduct research related to Tax Analytics and Automation Diffusion in Large Companies. In order to stay relevant and competitive, businesses must take advantage of the latest technologies. Many new technologies and tools can automate mundane business administration tasks. Automating administrative tasks leads to greater efficiency and fewer errors and it frees up employees to focus on tasks that add more value to the company. All of these principles hold true for a company's tax and accounting department as well.

The advent of robotic process automation (RPA), the automation of business processes using artificial intelligence and software, has had a particularly significant impact on the accounting profession. [6] The extent to which companies and accounting departments have adopted and integrated automated tax tools into their workplace varies. The aim of our client's study is to determine what factors seem to have the most significant impact on the initiation, adoption, and routinization of tax analytics and automation (TAA) tools in larger companies.

Initiation, adoption, and routinization are explained in [1]. Initiation refers to an initial evaluation of the suitability of a technology by the company, adoption refers to the the point at which a company recognizes a technology as valuable for its interests and begins to integrate that technology into its business practices, and routinization refers to the full-scale deployment of that technology.

In determining what factors most impact the initiation, adoption, and routinization of a technology by a company, our client employed the technology-organization-environment (TOE) framework that was first proposed in [9] and later used by Zhu, et al. [10] to study the diffusion of e-business technology in companies.

The TOE framework indicates that technology diffusion is impacted by the technological sophistication and preparedness of a company, the organizational and managerial attributes of a company, and the the regulatory and competitive context or environment of a company.

Our client is interested in testing five hypotheses regarding TAA tools using the TOE framework:

1. Technology readiness is positively related to TAA initiation/adoption/rou-

tinization.

2. Technology integration is positively related to TAA initiation/adoption/routinization.
3. Managerial obstacles are negatively related to TAA initiation/adoption/routinization.
4. Competition intensity is positively related to TAA initiation/adoption/routinization.
5. A supportive regulatory environment is positively related to TAA initiation/adoption/routinization.

To answer these questions, our client sent surveys to a chief tax officer at each of the Fortune 1000 companies asking about the intiation, adoption and routinization of TAA within that company's tax department and the level of technological readiness, technological integration, managerial obstacles, competition intensity, and regulatory support for TAA for the company.

Several of the variables like initiation, routinization, managerial obstacles, technology integration, competition intensity, and regulatory environment were measured using a number of 5-point Likert scale survey items. For instance to measure initiation, respondents were asked to use a Likert scale on 7 survey items to rate the significance that each of 7 potential TAA benefits had in their initial decision to pursue TAA tools.

Other variables like technological readiness where measured by asking a series of questions relating to the kinds of technology used at the company and the number of tax employees at the company who specialize in the use of TAA tools.

Of the 1000 companies that received surveys, 70 responded. The client has already performed an analysis answering the hypotheses for the routinization phase of the technology diffusion process. He wants us to perform analysis that explores the intitation, adoption, and routinization phases simultaneously.

# 2 Method / Analysis

## 2.1 Survey Details and Preliminary Data Transformation

The data were collected using a survey with 28 questions. The first question which measures the initiation phase, asks what factors managers considered when conducting an initial evaluation of TAA tools. The question consists of 7 Likert scale items. These 7 Likert items are transformed into a composite score using the methodology in 2.2. The second question asks the respondent to select which of four applications of TAA tools their company has adopted. This question is meant to measure the level of TAA adoption in the company. The responses will be combined into a single composite score using a weighting scheme for the different items based on discussions with the client. The third question is a 4-item Likert question meant to measure the degree to which companies have routinized TAA tools. It is transformed in a similar way to the data from question 1.

The next 7 questions are used to measure the technology readiness of the company. There is a question asking about the database technologies used by the company, the responses from this question are transformed into a binary variable which is 1 if the company is using an advanced database technology and a 0 if they are using no databases or simple database technology like Excel or Access. The next question asks if the company has a manager for TAA tools. This is coded as a binary variable that is 1 if the company does have such a manager and 0 otherwise. The next question asks about the total number of IT individuals in the company. It is coded as a categorical variable taking on 0 for none, 1 for less than 20, and 2 for between 20 and 200, and 3 for more than 200. There is also a question asking how many IT professionals are dedicated to the tax department and the use of TAA tools. This question is coded as 0 for none, 1 for between 1 and 10, 2 for between 10 and 20, and 3 for more than 20. The next question asks for the total number of tax professionals in the company; this raw count is included as a variable in the dataset. Following this, there is a question about how many tax professionals specialize in TAA tools. This response is also included in the dataset as a raw count. The final technology readiness question asks the respondent to select from a list of 9 TAA tools which ones are employed by their company. This data will be condensed into a single score by consulting with the client to establish weights for the different technologies. The composite score will be a sum of the weights associated with the technologies claimed by each company.

The next question measures technology readiness using a two item Likert scale question. The data from this question is transformed into a composite score using the methodology described 2.2.

The next set of questions seeks to measure the overall size of the company. The first question asks for the total number of company employees. This data is incorporated as a categorical variable with 1 for less than 100, 2 for 100-300, 3 for 300-500, 4 for 500-1500, 5 for 1500-5000, and 6 for more than 5000. The next question asks about the geographic extent of the company. It is incorporated as a binary variable which is 0 if the company has only sites in its home country and 1 if the company has sites in multiple countries. The next question asks about the number of foreign subsidiaries of the company and is incorporated into the data as a categorical variable with one level corresponding to 0 subsidiaries, one level to 1-20 subsidiaries, and one level to more than 20 subsidiaries. The last two questions ask for the percentage of foreign sales and the percentage of foreign purchases. Both variables are incorporated as categorical variables with levels corresponding to $0 - 5\%$, $5.1 - 15\%$, $15.1 - 25\%$, $25.1 - 35\%$, $35.1 - 45\%$, and more than 45%.

The next three questions seek to measure the managerial obstacles to TAA, the level of competition faced by the company, and the regulatory/governmental attitude to TAA use in the company's home country. All three of these questions are multi-item Likert scale questions and are transformed and represented using the composite score method detailed in 2.2.

The last few questions are general demographic questions about the respondent and are not expected to be of interest for the data analysis.

## 2.2 Creating Composite Scores from Multiple Likert Items

One challenge presented by the TAA dataset is the fact that several response variables of interest like TAA evaluation and routinization and several explanatory variables of interest like technology integration and managerial obstacles are measured using multiple Likert scale items.

When several survey items are used to measure a single underlying variable, it is often desirable to have some way of combining the information from the several survey items into a single composite score representing the variable of interest. A common approach for deriving such composite scores is to obtain the principal components of the several survey items and use these principal components, especially the first principal component as the composite score [3][7].

Our difficulty is that we are dealing with Likert items for which principal components analysis is not really well-defined. One solution to this problem that we employed for this data analysis is nonlinear principal components analysis with optimal scaling as described in [5] and [2]. This approach proceeds by finding an optimal transformation of the Likert data into a continuous representation and then performing standard PCA on this transformed representation.

Similar to standard principal components analysis, the nonlinear principal components analysis problem can be represented as a loss minimization problem. We minimize the loss function using a technique called alternating least squares.

Let $\mathbf{X}_{n\times p}$ be a matrix of $n$ observations of $p$ Likert scale items. We assume that $x_{ij}$, the $i$-th observation of the $j$-th item is 1 for the lowest rank item on the Likert scale, 2 if it is the second lowest rank item on the Likert scale, and so on. This represents the raw Likert scale data we would like to condense into a composite score. Let $\mathbf{Q}_{n\times p}$ represent the transformed version of $\mathbf{X}$ under optimal scaling. Let $\mathbf{H}_{n\times q}$ be the matrix of component scores where $q \leq p$. In our application, since we want a single composite score, we have $q = 1$. Finally, let $\mathbf{A}_{q\times p}$ be the matrix of component loadings. $\mathbf{HA}$ can be interpreted as a rank $q$ approximation to $\mathbf{X}$. We wish to find the optimal approximation, so we minimize the loss function:

$$L(\mathbf{Q}, \mathbf{H}, \mathbf{A}) = n^{-1} \sum_{j=1}^{p} (\mathbf{q}_j - \mathbf{H}\mathbf{a}_j)^T (\mathbf{q}_j - \mathbf{H}\mathbf{a}_j) \tag{1}$$

where $\mathbf{q}_j$ and $\mathbf{a}_j$ are the $j$-th columns of $\mathbf{Q}$ and $\mathbf{A}$ respectively. Alternating least squares works by iteratively minimizing the loss function with respect to one of $\mathbf{Q}$, $\mathbf{H}$, or $\mathbf{A}$ while holding the others constant until some convergence criterion is met. Thus, the optimal scaling problem (finding $\mathbf{Q}$) is solved simultaneously with the principal components problem.

All that remains is to speak briefly about the form of $\mathbf{Q}$. In general for each of $j = 1, .., p$, we have

$$q_{ij} = \sum_{k=1}^{b} \alpha_k \phi_k(x_{ij}) \tag{2}$$

where $q_{ij}$ is the $i$-th entry of $\mathbf{q}_j$ and $\phi_k(\cdot)$ is a basis function representation of $x_{ij}$. After some preliminary experimentation, a second degree B-spline was determined to do the best job of capturing the structure in the Likert items. If $x_{ij}$ is a missing value, $\phi_k(\cdot)$ is set to $1/b$. This is one of three methods for handling missing values suggested in [2]. It is called the averaging method. After experimentation, it was found that the other two methods produced imputations for the missing values that were too extreme relative to the non-missing values. The optimization of the loss function with respect to $\mathbf{Q}$ involves finding for each $j = 1, ...p$, the set of $\alpha_1, ..., \alpha_k$ that minimizes the loss function.

## 2.3 Missing Value Imputation

As introduced before, there are 28 survey questions in total. However, based on subject knowledge of the client and other related reference, only 23 questions by now have enough evidence or interpretability to be extracted into the cleaned dataset for further analysis. As the following figure 1 shows, the 23 questions that involves 56 variables should map into three response variables and eleven predictor variables.

Thus, we conduct an exploratory data analysis on this cleaned datset with 68 observations and 56 original variables in total. To have some brief overview, we check the proportion of missing values that all the variables have in advance, and have observed that there are fifteen variables that have a proportion of missing values larger than fifty percent; ten variables have the $10\% - 50\%$ proportion of missing values; and only the five categorical variables corresponding to Q5 have no missing value. Since the sample size is only 68, which is too small to throw any records with missing values, and the proportion of missing values in the datasets is large in a number of variables as figure 2 show, we need find some better way to make full use of the data and conduct appropriate missing value imputation for those variables.

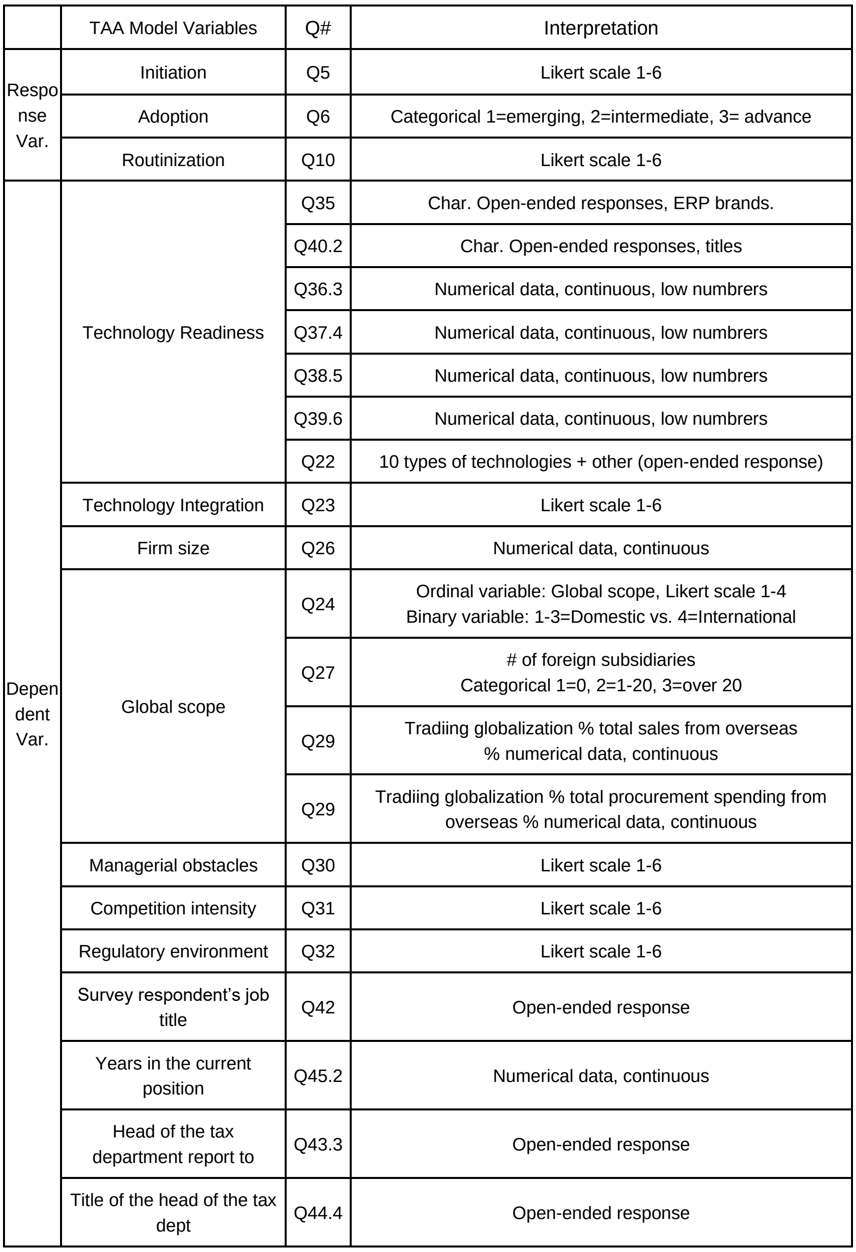
As for the missing value imputation, we adopt the method named AMELIA. [4] Here the method assumes the first assumption that the complete data $\mathcal{D}_{n\times k} = (\mathcal{D}^{obs}, \mathcal{D}^{mis})$ follows the multivariate normal distribution, which is

$$\mathcal{D} \sim N_k(\mu, \Sigma) \tag{3}$$

and a second assumption that missing data are missing at random (MAR), which means the probability of missingness could be fully explained by observed data, defined as

$$p(M \mid \mathcal{D}) = p(M \mid \mathcal{D}^{obs}) \tag{4}$$

where M is the missing matrix with $M_{ij} = I(\mathcal{D}_{ij} \in \mathcal{D}^{mis})$.



| | TAA Model Variables | Q# | Interpretation |
|---|---|---|---|
| Response Var. | Initiation | Q5 | Likert scale 1-6 |
| | Adoption | Q6 | Categorical 1=emerging, 2=intermediate, 3= advance |
| | Routinization | Q10 | Likert scale 1-6 |
| Dependent Var. | Technology Readiness | Q35 | Char. Open-ended responses, ERP brands. |
| | | Q40.2 | Char. Open-ended responses, titles |
| | | Q36.3 | Numerical data, continuous, low numbrers |
| | | Q37.4 | Numerical data, continuous, low numbrers |
| | | Q38.5 | Numerical data, continuous, low numbrers |
| | | Q39.6 | Numerical data, continuous, low numbrers |
| | | Q22 | 10 types of technologies + other (open-ended response) |
| | Technology Integration | Q23 | Likert scale 1-6 |
| | Firm size | Q26 | Numerical data, continuous |
| | Global scope | Q24 | Ordinal variable: Global scope, Likert scale 1-4<br>Binary variable: 1-3=Domestic vs. 4=International |
| | | Q27 | # of foreign subsidiaries<br>Categorical 1=0, 2=1-20, 3=over 20 |
| | | Q29 | Tradiing globalization % total sales from overseas % numerical data, continuous |
| | | Q29 | Tradiing globalization % total procurement spending from overseas % numerical data, continuous |
| | Managerial obstacles | Q30 | Likert scale 1-6 |
| | Competition intensity | Q31 | Likert scale 1-6 |
| | Regulatory environment | Q32 | Likert scale 1-6 |
| | Survey respondent's job title | Q42 | Open-ended response |
| | Years in the current position | Q45.2 | Numerical data, continuous |
| | Head of the tax department report to | Q43.3 | Open-ended response |
| | Title of the head of the tax dept | Q44.4 | Open-ended response |

Figure 1: The mapping of useful survey questions to the variables for analysis

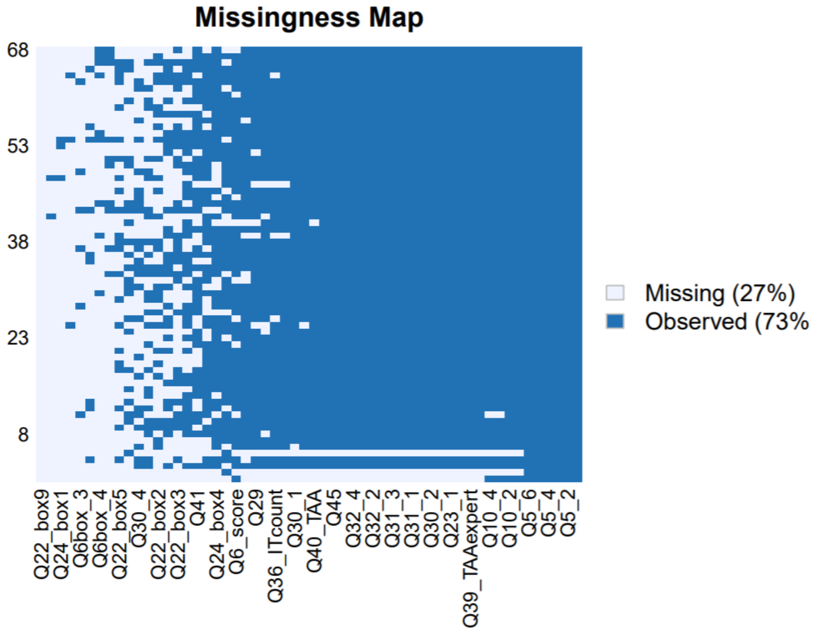


Figure 2: The visualization of missingness in the data. The y-axis is the id of the observation, and the x-axis is the variables with decreasing proportion of missing values from the left to the right

Hence, the normal assumption as well as the MAR assumption should be checked after the data are fully prepared and before the conduction of imputation.

Remembering the normality assumption Eq. 3 and the missing at random assumption Eq. 4, we can obain the following likelihood for the observed data $\mathcal{D}^{\text{obs}}$:

$$p\left(\mathcal{D}^{\text{obs}}, M \mid \theta\right) = p\left(M \mid \mathcal{D}^{\text{obs}}\right) p\left(\mathcal{D}^{\text{obs}} \mid \theta\right) \tag{5}$$

where $\theta = (\mu, \Sigma)$ is the parameter of interest. Thus the likelihood of $\theta$ based on observed data can be expressed as:

$$L\left(\theta \mid \mathcal{D}^{\text{obs}}\right) \propto p\left(\mathcal{D}^{\text{obs}} \mid \theta\right) = \int p(\mathcal{D} \mid \theta) d\mathcal{D}^{\text{mis}} \tag{6}$$

and with a uniform prior of $\theta$, the posterior could be

$$p\left(\theta \mid \mathcal{D}^{\text{obs}}\right) \propto \int p(\mathcal{D} \mid \theta) d\mathcal{D}^{\text{mis}} \tag{7}$$

Since the $\mathcal{D}^{\text{mis}}$ is missing, Expectation-Maximization (EM) algorithm combined with a bootstrap approach is applied here to obtain the estimated modes of the posterior with its variance estimation. By assuming a start point of $\theta_0$, we define the Q function as

$$Q(\theta \mid \theta^v) = E_{\theta^v}(ln(p\left(\mathcal{D} \mid \theta\right)) \mid \mathcal{D}^{\text{obs}}) \tag{8}$$

where $p\left(\mathcal{D} \mid \theta\right)$ is the complete data likelihood. The updates of $\theta^{v+1}$ given $\theta^v$ as

$$\theta^{v+1} = \arg\max_{\theta} Q(\theta \mid \theta^v) \tag{9}$$

Thus, we have the estimations of $\theta$, and can make imputation of the missing data based on the observed data as well as the estimated parameters for the complete data likelihood.

Besides, $m$ simultaneous imputations can be combined together to produce a more stable average imputaion:

$$\bar{T} = \frac{1}{m}\sum_{j=1}^{m} T_j \tag{10}$$

where $j$ represents the $j^{th}$ datasets, and the $\bar{T}$ has the following standard error:

$$SE(T) = \sqrt{\frac{1}{m}\sum_{j=1}^{m} SE\left(T_j\right)^2 + S_T^2(1 + 1/m)} \tag{11}$$

where, $S_T^2 = \Sigma_{j=1}^{m}\left(T_j - \bar{T}\right)^2/(m{-}1)$ is the sample variance across the m estimations, and $SE\left(T_j\right)^2$ is the estimated variance of $T_j$ based on the $j^{th}$ dataset.

Besides, it's also common that to treat the missing value of a categorical variable as a new variable; for the questions related to Likert scales where there's always a level that indicating the customer don't know the answer, we can directly change the missing value into that case; some times a mean imputation would also be appropriate if sufficient subject knowledge are grounded. Thus, the missing value imputation of this project can be quite customized to the specific variable with more subject knowledge. However, the cleaner data should be get in the first step.

## 2.4 Structural Equation Modelling

Traditional multiple regression is handy when we only have one dependent variable, but it can't handle the case where we have multiple dependent variables. Structural equation modeling (SEM) is popular methodology in social sciences for representing, estimating, and testing a network of relationships between variables (measured variables and latent constructs) [8]. There are two main advantages of choosing this method. First, we employ the technology-organization-environment (TOE) framework but the three attributes are not directly measurable. SEM handles this issue naturally by encoding them as latent constructs. Second, we have three response variables, namely TAA initiation/adoption/routinization, which form a complex relationship network with our dependent variables that is well handled by SEM. However, covariance-based methods like SEM require a large sample size.

## 2.5 Multivariate linear regression

Another technique for handling multiple responses is multivariate linear regression. Suppose we have $m$ responses with $p$ features and $n$ observations, the multivariate linear regression model can be written in matrix form

$$Y = XB + E,$$

where $Y \in \mathbb{R}^{n\times m}$ is are the dependent variables, $X \in \mathbb{R}^{n\times p}$ are the independent variables, $B \in \mathbb{R}^{p\times m}$ are the coefficients and $E \in \mathbb{R}^{n\times m}$ are the errors. To account for the possible correlation of different columns of $Y$, we assume that each row $E_{i.}$ of $E$ has a covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$. If put in vectorized form,

$$\operatorname{vec}(Y)|X \sim \mathcal{N}([B' \otimes I_n]\operatorname{vec}(X), \Sigma \otimes I_n).$$

It can be shown that the maximum likelihood estimate of $B$ is invariant to $\Sigma$, which can be computed using ordinary least square

$$\hat{B} = (X'X)^{-1}X'Y.$$

Then an unbiased estimate of error covariance matrix $\Sigma$ will be

$$\hat{\Sigma} = \frac{Y'(I_n - X(X'X)^{-1}X')Y}{n-p-1},$$

Also it can be seen that

$$\operatorname{vec}(\hat{B}) \sim \mathcal{N}(\operatorname{vec}(B), \Sigma \otimes (X'X)^{-1})$$

Thus if we want to test $B_{jk} = 0$ against $B_{jk} \neq 0$, under the null hypothesis

$$\frac{\hat{B}_{jk}}{\hat{\Sigma}_{kk}(X'X)^{-1}_{jj}} \sim t_{n-p-1}$$

# References

[1] Randolph B Cooper and Robert W Zmud. Information technology implementation research: a technological diffusion approach. Management science, 36(2):123–139, 1990.

[2] Jan De Leeuw. Multivariate analysis with optimal scaling. 2011.

[3] William G Henderson, Susan G Fisher, Noel Cohen, Susan Waltzman, Laura Weber, VA Cooperative Study Group on Cochlear, et al. Use of principal components analysis to develop a composite score as a primary outcome variable in a clinical trial. Controlled Clinical Trials, 11(3):199–214, 1990.

[4] James Honaker, Gary King, Matthew Blackwell, et al. Amelia ii: A program for missing data. Journal of statistical software, 45(7):1–47, 2011.

[5] Mariëlle Linting, Jacqueline J Meulman, Patrick JF Groenen, and Anita J van der Koojj. Nonlinear principal components analysis: introduction and application. Psychological methods, 12(3):336, 2007.

[6] Steven Mezzio. Robotic process automation for tax: Tax is in the vanguard of the fourth industrial revolution. Journal of Accountancy, 228(6):18, 2019.

[7] Reza Motallebzadeh, J Martin Bland, Hugh S Markus, Juan Carlos Kaski, and Marjan Jahangiri. Neurocognitive function and cerebral emboli: randomized study of on-pump versus off-pump coronary artery bypass surgery. The Annals of thoracic surgery, 83(2):475–482, 2007.

[8] Diana Suhr. The basics of structural equation modeling.

[9] L Tornatzky and Mitchell Fleischer. The process of technology innovation. Lexington, MA: Lexington Books, 165, 1990.

[10] Kevin Zhu, Kenneth L Kraemer, and Sean Xu. The process of innovation assimilation by firms in different countries: a technology diffusion perspective on e-business. Management science, 52(10):1557–1576, 2006.