Scientax
Jurnal Kajian Ilmiah Perpajakan Indonesia

djp

VOLUME 5 NO. 1 | OKTOBER 2023 Journal Page: ejurnal.pajak.go.id ISSN 2686-5718

Review of the Employee Performance Evaluation System at the Directorate General of Taxes: Its Effectiveness and Employee Satisfaction Level

Anies Said Basalamah[a], Herru Widiatmanti[b]
[a] Leadership and Managerial Training Center, Financial Education and Training Agency, Ministry of Finance, Indonesia.
Email: asbasalamah@yahoo.com
[b] Leadership and Managerial Training Center, Financial Education and Training Agency, Ministry of Finance, Indonesia.
Email: widiatmanti@gmail.com

* Corresponding author: asbasalamah@yahoo.com

ABSTRACT

This quantitative research seeks to determine the effectiveness of the Directorate General of Taxes (DGT) performance evaluation system, measure how satisfied the employees are, and to prove differences among different categories of employees. This research is an extension of the study by Widiatmanti (2020) on 1,587 DGT employees. Analysis of Variance (ANOVA) and Exploratory Factor Analysis (EFA) were used to analyze the data in addition to descriptive data analysis. Generally, respondents feel the system is quite effective, and they are quite satisfied, except for three variables that most respondents strongly disagree with and relate to the performance status of SABCD and the level 1 and level 2 ranking sessions. ANOVA method indicates that there are indeed differences among different categories of employees, while EFA method reveals that despite seven insignificant variables, none of them suggests the need to change DGT Regulation 12/PJ/2018, including its SABCD performance status. However, the need for change or adjustments arises because there are provisions at higher levels of regulation: Government Regulation and two ministerial decrees. Therefore, we suggest that the use of forced rank in the SABCD performance status is abolished and replaced mainly due to excellent, moderate, and poor performances.

Keywords: performance evaluation system; DGT, Exploratory Factor Analysis (EFA)

1. INTRODUCTION

Law of the Republic Indonesia No. 5 of 2014 concerning State Civil Apparatus and Government Regulation No. 11 of 2017 concerning Management of Civil Servants, which has been amended by Government Regulation No. 17 of 2020, have significantly changed the management of civil servants, especially with regard to career management, career development, competency development, career patterns, transfers to other units, and promotions. This change is related to the application of the merit system to improve civil servants' competencies, performance, and professionalism. In implementing the merit system, an objective, measurable, accountable, participatory, and transparent employee performance assessment is needed to achieve

doi: 10.52869/st.v5i1.429
Received: July 19, 2022; Accepted: October 11, 2023; Published: October 31, 2023

Scientax: Jurnal Kajian Ilmiah Perpajakan Indonesia is Sinta 4 Journal (https://sinta.kemdikbud.go.id/journals/profile/9121)
How to Cite:
Basalamah, A. S., & Widiatmanti, H. (2023). Review of the employee performance evaluation system at the Directorate General of Taxes: Its effectiveness and employee satisfaction level. *Scientax: Jurnal Kajian Ilmiah Perpajakan Indonesia*, *5*(1), 87–102. https://doi.org/10.52869/st.v5i1.429

professional, agile, high-integrity, and high-performing civil servants. Moreover, the government has also enacted Government Regulation No. 30 of 2019 concerning Civil Servants Performance Assessment, which regulates, among other things, the substance of performance appraisal consisting of work behavior assessment and performance appraisal, weighting method of employee performance standards, appraisers, and performance appraisal teams, performance information systems, assessment procedures, performance ratings, performance reporting, performance rewards, punishments, and objections.

One of the applications at the ministerial level can be seen in the Directorate General of Taxes (DGT) of the Ministry of Finance (MoF). Its employees receive the highest performance allowance compared to other employees at MoF and all other civil servants. Performance appraisal at DGT as regulated in the Director General of Taxes Regulation No. PER-12/PJ/2018 concerning Performance Management within the DGT uses the e-performance application as the performance measurement applicable to other echelon I units within MoF. However, the employee performance rating (1 to 5) and the employee performance status (SABCD, where S represents the best and D represents the worst) only apply to the DGT employees. Table 1 depicts not only the five categories for each status but also shows the existence of a forced distribution. The performance status of employees at DGT is one of the criteria for providing performance allowances and promotions (Widiatmanti, 2020). According to the DGT Regulation No. 12 of 2018, as indicated by Widiatmanti (2020), in addition to providing categories to the best employees to give awards to them, there are also other purposes such as for employee management, development of conducive and competitive work climate, means of effective communication and the establishment of a harmonious relationship between subordinates and superiors, increasing employee job satisfaction, and developing an effective work culture so that employees are able to make optimum contributions.

The implementation of performance appraisal using forced distribution, which has an impact on employee performance allowance and promotion, raises the reluctance and anxiety of superiors who assess and rank, in addition to being perceived or even expressed as unfair by employees who are assessed and rated. From the authors' initial observations with several employees, especially those who were involved in the Leadership Development Program training activities at the Center for Education, Training and Development of Human Resources, some employees told how difficult they were when facing employees during the ranking process, in addition to the difficulty in motivating employees affected by forced distribution for each office unit. Moreover, according to them, performance status

Table 1 Performance Status Categories of DGT Employees
Source: Widiatmanti (2020)

Employee Percentage	Performance Status	Categories of Employee Performance Status	Conversion Status Employee Performance Achievement (%)
15%	Best Performance	S (15%)	100,0
70%	Average	A (20%)	97.5
		B (30%)	95.0
		C (20%)	92.5
15%	Below Average	D (15%)	90.0

also became one of the criteria for organizations and leaders in transferring employees from one unit to another. These phenomena not only contradictory to the purpose of the DGT Regulation No. 12 of 2018 as indicated earlier, but also lead to employees dissatisfaction, both those who assess and those who are assessed, regarding the determination of the distribution per category of employee performance status. This dissatisfaction inspires the authors to research to answer the following research questions: what is the perception of DGT employees regarding the current performance appraisal model? If there are few things need to be changed, what are they?

2. LITERATURE REVIEW

Performance evaluation is not only a systematic evaluation of the performance of employees but also understands their abilities so that it can be used to plan further career development for the employees concerned. Torrington et al. (2020) define job evaluation as a "formal, systematic process to determine the relative worth of jobs within an organization," and it can be done in several ways or methods: ranking method that is appropriate in small organizations, classification method or job grading that is often used in public sector organizations in which jobs are categorized into groups, and factor-comparison method that combines the ranking and point factor methods using quantitative data. The latter is the most widely used one (Torrington et al., 2020).

According to Mathis et al. (2017), there are many methods that can be used to evaluate performance, either one method for all jobs and employees, different methods for different groups of employees, or a combination method. Rating scale, comparative methods which compare the performance levels of employees against one another that include ranking and forced distribution, and narrative methods such as critical events or other essay methods are tools for evaluating performance. Another tool is to use psychologists or assessment center to assess the potentials of employees (Mathis et al., 2017; Dessler, 2020; Torrington et al., 2020).

Research on the effectiveness of performance evaluation is inconclusive. Sharma et al. (2016) found that accuracy and fairness are two factors that contribute performance management systems (PMS) effectiveness. Dewettinck & van Dijk (2012) also found that fairness mediates the relationship between characteristics of employee PMS and their effectiveness. Torrington et al. (2020) indicate that "performance management must have credibility with employees," while Blackman et al. (2018) argue that to be effective, PMS have to be operated as part of the organization's core business.

The research results are also inconclusive. Mathis et al. (2017) quotes a report that indicates "86% of firms actively use performance management, around 70% of those surveyed thought that the process was not positive, and 29% believed that is was unfair. A mere 3% of companies plan to alter their current approaches." Evita et al. (2017) indicates that the Graphic Rating Scale and employee daily work reports are considered ineffective because many employees consider both methods are merely a formality, subjective, and there are no clear and measurable standards as well as feedback on employee performance achievements. As a result, employees feel uncomfortable and unmotivated at work.

Research by Harbi et al. (2017) links performance appraisals with Saudi Arabian culture in the form of wasta personal relationships (literally, wasta can be interpreted as connections or favoritism) which cause people who do not have close relationships to feel unfairly treated and lead to negative outcomes. They also show how employees begin to reject this wasta norms and adopt alternative values related to the notion of organizational justice and individual egalitarianism (Harbi et al., 2017). Bias by raters is also indicated by Mathis et al. (2017) as one of the major sources of error in the performance appraisal process.

3. RESEARCH METHODOLOGY

This research is an extension of the authors' quantitative and qualitative research (2022) using questionnaires as attached. Although the results show that in general respondents are satisfied with

the system with most of the modes are 3 indicating respondents are agree, and the system is perceived to be quite effective, respondents who chose Strongly Disagree and Disagree for the negative tones of the questions were quite a lot. The interview also reveals that all reviewees hoped that the SABCD performance status is abolished since it causes injustice, reduces employee motivation, reduces collaboration efforts, and creates organizational uproar (Widiatmanti, 2020).

To know which factors from X1 to X38 need to replace or abolish as suggested by interviewees, this study was done using the theoretical framework shown in Figure 1. The analysis used to achieve the first research objective is the exploratory factor analysis (EFA), while the ANOVA test was used to fulfill the second objective.

The population in this study was 45,948 all DGT employees (Biro SDM, 2020) that spread over the work units of the Head Office, 34 Regional Tax

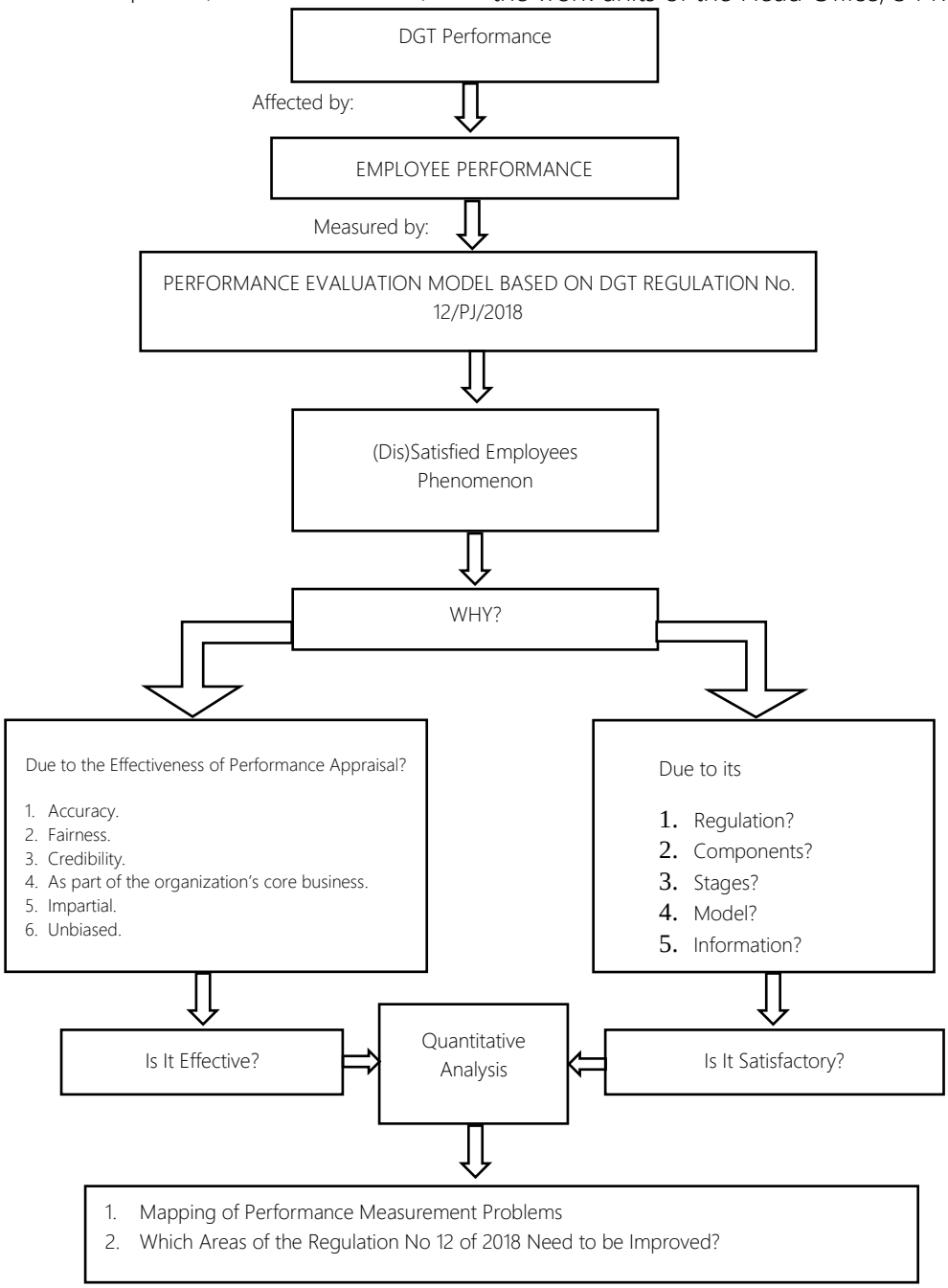

Figure 1. Research Theoretical Framework
Source: Adapted from Widiatmanti (2020)

Offices, 352 Tax Offices and 204 Tax Counseling and Consulting Offices. The questionnaire was sent using Google Forms. However, the number of respondents who filled out the questionnaires was only 1,684 or 3.665% response rate which might be small compared to the average internet survey. Nevertheless, using the Slovin formula, the level of achieved tolerable error is 2.392% which is more conservative than 5% commonly used. In other words, 1,684 samples returned is higher than the minimum sample according to Slovin formula for 5% tolerable error. However, since there are few respondents who filled the questionnaire twice or even three times, only 1,587 respondents were processed.

4. ANALYSIS AND DISCUSSIONS

4.1 Descriptive Data

Of the 1,587 respondents, 1,068 are male (67.30%) and 519 are female (32.70%); 1,247 respondents are married (78,58%), 307 respondents are single (19.34%) and 33 respondents are either widows or widowers (2.08%). Meanwhile, based on their age, 56 respondents are over 55 years old (3.53%); 422 respondents are over 45 up to 55 years (26.59%), 453 respondents are over 35 up to 45 years (28.54%), 460 respondents are over 25 up to 35 years (28.99%), 186 respondents are over 20 to 25 years (11.72%), and 10 respondents (0.63%) aged up to 20 years.

As for the respondents' positions, 882 respondents are staffs (55.58%), 193 respondents are functional officers (12.16%), 444 respondents are echelon IV or supervisors (27.98%), 63 respondents are echelon III or administrators (3.97%), and 5 respondents are echelon II (0.32%). Based on their educational level, 508 respondents are high school graduates or equivalent (32.01%), 209 respondents are Diploma I (13.17%), 194 respondents are Diploma III graduates (12.22%), 623 respondents are S1 or Diploma IV graduates (39.26%), 43 respondents are S2 graduates (2.71%), and 10 respondents are S3 graduates (0.36%).

In terms of the length of time the respondents worked, 208 respondents have up to five years of service (13.11%), 277 respondents have five to ten years of service (17.45%), 270 respondents have more than ten to 15 years of service (17.01%), 537 respondents have more than 15 to 20 years of service (33.84%), 214 respondents have more than 20 to 30 years of service (13.48%), and 81 respondents have more than 30 years of service (5.01%). Combining this characteristic with the respondents' age that only ten respondents whose age are up to 20 years, most respondents are representative enough in a sense that they are able to distinguish between the old DGT performance evaluation method and the new one.

Few respondents seem to hide their identities, such as echelon III officer with the length of service of less than ten years which is impossible to achieve at DGT, or with age of up to 20 years but having years of service of more than 5 years. Other respondent chose 3 on the age meaning his or her age is 25 – 35 years, yet he or she chose 6 for tenure, meaning he or she has already work for more than 30 years. However, since EFA method does not distinguish whether the respondents are male or female, officials or staff and so on, the 1,587 respondents were continued to be processed since all of them filled out the questionnaire completely. However, for the purpose of distinguishing among the categories of respondents, these weird respondents will be dropped from the calculation.

To achieve the objectives of this research, the questionnaire sent was divided into four groups. The first group relates to the performance appraisal variables, the second one relates to the performance appraisal model variables, the third group relates to the effectiveness of the performance appraisal model variables, and the fourth group relates to the performance appraisal regulatory variables. Appendix 1 details the answers to the questionnaire along with each median and mode. As can be seen in Appendix 1, all variables were responded by both those who strongly agreed (4) and those who strongly disagreed (1). This shows that statistically their answers support the author's initial observations because the questions in the questionnaire are

Table 2 Criteria for Performance Appraisal Variables
Source: Widiatmanti (2020)

Total Score of Group A	Total Score of Group B	Total Score of Group C	Criteria	Total Score of Group D	Criteria
11 – 19.3	6 – 10.5	11 – 19.3	Unsatisfactory	10 – 17.5	Ineffective
19.4 – 27.7	10.6 – 15.1	19.4 – 27.7	Less	17.6 – 25.1	Less Effective
27.8 – 36.1	15.2 – 19.7	27.8 – 36.1	Quite	25.2 – 32.5	Quite Effective
36.2 –	19.8 – 24.3	36.2 –	Satisfactory	32.6 – 40.1	Effective

$$Average = \frac{Total\ Answers}{1.587} = \frac{\sum choices\ 1 + \sum choices\ 2 + \sum choices\ 3 + \sum choices\ 4}{1.587}$$

partly derived from these initial observations. For example, for question A2 that asks whether respondents are satisfied with the results of the performance appraisal they get, although the majority of respondents chose satisfied (3 or agree) and very satisfied (4 or strongly agree), 84 respondents chose very disagree (1) and 220 chose disagree (2). Similarly, although most respondents chose agree (3) and strongly agree (4) in almost all questions, there are still many respondents who chose disagree (2) and strongly disagree (1).

Three questions that the majority of respondents chose strongly disagree (1) and disagree (2) are the accuracy of using the normal curve for the performance status of SABCD employees (question C8), level 1 and level 2 ranking sessions are publicly informed to employees (question C11), and the performance status of SABCD motivates employees to improve group performance (question D8) with median of 2, respectively and mode of 2, 2, and 3, respectively. As with question D8 that performance status (performance categories SABCD) motivates employee to improve group performance, even though the median is also 2 which means disagree, the number who chose Agree is still more than those who chose Strongly Disagree, Disagree and Strongly Agree with the mode 3 which means agree. For all other variables, the modes and medians are 3 which means Agree. The only variable that both median and mode are 4 is question B3 namely performance is assessed within a certain period of time (see in Appendix 1). As shown in Appendix 1, all variables were responded, both by those who chose strongly agree (4) and those who chose strongly disagree (1) with almost all medians and modes are 3.

Descriptive statistics also shows similar results (Widiatmanti, 2020). Using the criteria as shown in Table 2, for group A a total value of 51,107 was obtained, namely the total who answered 1 multiplied by 1 plus the total who answered 2 multiplied by 2 plus the total who answered 3 multiplied by 3 plus the total who answered 4 multiplied by 4, or (1,308 X 1) + (3.447 X 2) + (7.903 X 3) + (4.799 X 4). With a total of 1,587 respondents, an average value of 32.20 (i.e., 51,107:1,587) is obtained. Referring to Table 2, the value of 32.20 is in the criteria of "Quite Satisfactory." Meanwhile for group B a total score of 29,913 was obtained so that an average value of 18.85 was also included in the criteria of "Quite Satisfactory." Meanwhile for group C a total value of 48,716 was obtained so that an average value of 30.70 was also included in the criteria of "Quite Satisfactory." As for group D, the total score is 42,878 so that an average value of 27.02 is obtained which is included in the criteria of "Quite Effective." This indicates that there is still room for improvement for DGT so that employee dissatisfaction can be reduced or even eliminated as well as increasing the effectiveness of performance appraisal at DGT.

Another descriptive analysis that can be explored is to compare whether or not the population means differ among various types of DGT employees who participated in the survey. As indicated earlier, few respondents seem to hide

Table 3 ANOVA for Different Types of Respondents
Source: Survey results.

Categories	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Gender	11857.45	38	312.04	434.03	0.000	1.405027
Marital Status	7710.60	38	202.91	278.17	0.000	1.405027
Positions	7161.43	38	188.46	255.17	0.000	1.405027
Educational Levels	5991.38	38	157.67	211.74	0.000	1.405027
Places of Work	4840.38	38	127.38	152.27	0.000	1.405027

their identities. As a result, categories of tenure and age were not be processed in differentiating population means among various types of DGT employees. Using ANOVA, it is found that the lowest and highest average variables are 2.268 and 3.487, respectively. It can be seen from Table 3 that the p-value is 0, so it can be concluded that there are significant differences in the population average among the five types of DGT employees, i.e., staffs, functional officers, echelon IV, echelon III and echelon II. This means that on average the five types of employees at DGT gave different answers.

4.2 Validity and Reliability Tests

To assure that exploratory factor analysis (EFA) is the correct measuring instrument, several tests should be carried out before this method is used as a "requirement" that EFA is a valid tool. The first validity test is using Pearson correlation test to assess whether or not data is correlated with one another. Of the 1,587 respondents who completed the questionnaire, all variables have significant correlations because each has a p-value of < .001. Therefore, in further testing, no variables need to be eliminated as a treatment (Hair et al., 2019).

Normality test in multivariate analysis is a basic assumption where if the variation from the normal distribution is large enough, then all statistical methods used will be invalid (Hair et al., 2019). In this study we u sed the Shapiro-Wilks test which is one of the two most popular methods (Hair et al., 2019). Using Jamovi program, the lowest Shapiro-Wilks value is 0.718 and the highest is 0.908 but each variable has a low p value of < .001. Therefore, following the opinion of Zaiontz (2017) that testing whether the data follows a normal distribution multivariate analysis is difficult, and for large samples such as in this study, it usually relies on the multivariate central limit theorem, i.e., for a certain set of "random vectors X1, X2, ..., Xk that are independent and identically distributed, then the sample mean vector, $\bar{X}$, is approximately multivariate normally distributed for sufficiently large samples" (Zaiontz, 2017). Accordingly, all of the 38 variables remain being processed.

Reliability testing carried out on 38 variables results in a Cronbach alpha value of 0.975 with the lowest value of 0.974 and the highest of 0.976, all of which is greater than the lowest acceptable reliability limit of 0.6 or 0.7 (Hair et al., 2019).

An autocorrelation test was conducted using the Durbin–Watson test to determine whether there is a relationship between the elements of a series of observations in order to eliminate the effect of standard errors. From the respondents' data, the d value is 1.97. Since tables in statistical text books generally contain Durbin-Watson tables for k of 1 to 5 while this study employs 38 variables, we use the rule of thumb that if the value of the Durbin–Watson test is between 1.5 and 2.5 (Hutcheson & Sofroniou, 1999) or close to 2 (McClave et al., 2018), it means there is no autocorrelation. As stated by McClave et al. (2018), autocorrelation exists when the d value is less than 2 (positive correlation) or greater than 2 (negative correlation), close to 0 (very strong positive correlation) or close to 4 (very strong negative correlation). Thus, it can be concluded that this study shows there is no autocorrelation since d statistic 2.5 > 1.97 > 1.5 and tends to close to 2 (Hutcheson & Sofroniou, 1999; McClave et al., 2018).

Multicollinearity test was used to determine whether or not there is a linear relationship among variables in the regression model that influence each other. One of the two methods usually used is variance inflation factor (VIF) or the tolerance value (Hair et al., 2019) where FIV = (1 / tolerance). The results show that the highest VIF value is 2.02 which is far from 10 as the limit, and the lowest tolerance value is 0.495 which is much greater than 0.1 of the lowest limit value that according to Hair et al. (2019) are considered as common cutoff thresholds (see Appendix 2 for the detail results). A VIF value of more than 5 indicates a high multicollinearity, a VIF value of up to 5 means moderate multicollinearity, and a VIF value of 1 means there is no multicollinearity (Hair et al., 2019).

Heteroscedasticity test was conducted to determine whether there are symptoms where the probability distribution of deviation is not the same for all observations, and to determine the fulfillment of the homoscedasticity assumption (Hair et al., 2019). Using Jamovi we conducted the Levene test as one the most common tests (Hair et al., 2019; Navarro & Foxcroft, 2019). The results showed small p-values as predicted by Grace-Martin (2021) and Navarro & Foxcroft (2019) for large samples as with this research. Out of 38 variables, only 17 variables have p-values of 0.05 or above as the cutoff threshold. Even when we conducted Welch one-way test as a remedy (Navarro & Foxcroft, 2019), the results are quite the same, only 19 variables have p-values of 0.05 or above. Since it will be 50% of the variables that need to be excluded from the calculation, we tend to neglect the Levene test for large samples as suggested by Grace-Martin (2021).

To provide certainty that the methods used in this study will provide consistent results, testing of the questionnaire was carried out using the expert panel method (Schindler, 2019) that added and improved the questions in the questionnaire (Widiatmanti, 2020). This study did not test the questionnaire because it is the extension of the study of Widiatmanti (2020).

4.3 Exploratory Factor Analysis (EFA)

Based on validity and reliability tests discussed earlier, none of the variables were deleted since all of them meet the criteria except for the low p-values in normality and Levene tests that can be neglected for big samples such as in this research (Zaiontz, 2017; Grace-Martin (2021). As such, the EFA testing can be carried out. In addition, with the large number of variables (38 variables) and big samples (1,587 respondents) in this study, the minimum ratio of 20 respondents for every 1 variable (Hair et al., 2019) is also met. Using Jamovi program, testing of these 38 variables was conducted using the following steps:

1. Determining which variables significantly correlate with each other. In EFA, this kind of correlation is extremely important, and only variables with significant correlation will be processed in the next stage.

   Bartlett test of sphericity and Kaiser-Meyer-Olkin test measure of sampling adequacy (MSA) were performed, resulting in the overall MSA of 0.978 with the smallest MSA of 0.940 and the largest 0.992 (see in Appendix 3). According to Hair et al. (2019), an MSA of 0.8 or more is very good, 0.7 or more is good, 0.6 or more is mediocre, 0.5 or more is bad, and less than 0.5 is unacceptable. Since there are no variables with individual MSA below 0.5, no further variables were dropped for further testing stage. As for Bartlett's test for sphericity, the p-value is < .001, indicating that in general this test is adequate.
2. Determining the value of communalities which indicate whether or not a variable is dominant in a set of variables.

   As pointed out by Hair et al. (2019), "high communality values indicate that a large amount of the variance in a variable has been extracted by the factor solution. Small communalities show that a substantial portion of the variable's variance is not accounted for by the factors." The most reliable criteria

according to Hair et al. (2019) for variables between 20 to 50 is when the communality values are above 0.40.

Although Jamovi does not provide a feature to determine communality values, the program calculates the uniqueness values which when subtracted from 1 will result in communalities since uniqueness equals to "1 – communality" (Navarro & Foxcroft, 2019). Using the criteria mentioned above, several variables that are included in the factor loading but have a communalities value of 0.4 or less will be excluded from EFA. As a result, variables B4 and B5 will be dropped for further testing since it has communality value of 0.123 and 0.106, respectively (see the bold and italic numbers in Appendix 4).

3. Determining what factors are considered dominant from the series of variables under study.

   The dominance of these factors is related to the magnitude of the eigenvalue of each variable, which in Jamovi program is set to be 1. The number of components formed in the Jamovi program are sorted from those with the largest eigenvalue to the smallest value, and the number of factors stopped at the eigenvalue of 1. From the data being processed, this research acquires three factors (see in the Appendix 4) since the fourth factor has an eigenvalue value of 0.64650 which is lower than 1.

4. Determining the amount of the loading factor for each variable, which indicates the level of correspondence between the variable and the factor. The greater the loading, the more representative the factor will be.

   In Jamovi program, the loading factor is automatically set to 0.3. In this study, the correlation is significant when the loading is 0.5 or more. It is therefore none of the values in "Factors" columns in Appendix 4 are less than 0.5.

From Appendix 4 we can conclude that five variables are not significant since their loading factors are below 0.5, namely B6 regarding guidance from supervisors, C4 regarding performance appraisal information from supervisors, C6 regarding coaching from supervisors on aspects need to improve, D1 regarding alignment between KPI and organizational goals, and D4 regarding employee development. As for B4 and B5, these two variables were already be excluded since their communality values are less than 0.40 as the threshold for variables between 20 to 50 (Hair et al., 2019).

Appendix 4 shows that if DGT employees were not considered male or female, married or not, echelon officers or not, or whatever their ages and educations are, there are three dominant factors affecting performance evaluation with 31 variables that significantly dominant. The first factor is related to variables A1 to B2, D2, D5 and D6. The eleven variables that relate to the performance appraisal in the A group are all significant variables. Combine with the results described in section 4.1. Descriptive Data, these eleven variables are considered effective and satisfactory with the average of 32.2 indicating quite effective, and the medians and modes of 3 respectively indicating respondents generally Agree.

As for the performance appraisal model variables in group B, only three variables are significant. Two variables, namely B1 (feedback from supervisor regarding performance) and B2 (performance assessment using a rating scale), are in the first factor and B3 in the third factor. As with variables B4 and B5, EFA method considers these variables are insignificant since their communality values are below 0.4 (Hair et al., 2019). It also means that although most respondents chose 3 (agree) and 4 (very agree) for this B4 (see in the Appendix 1), EFA method points out that performance assessment does not have to be done using technical and managerial ability assessment in order to be considered fair by employees or as indicated by Sharma et al. (2016) and by Dewettinck & van Dijk (2012). DGT Regulation No. 12/PJ/2018 is not as bad as wasta in Harbi et al. (2017) research that create rejection from employees. This is also true for the B5 variable, which means that performance appraisal does not necessarily have to be done using the

self-assessment method, although most respondents considered so (see in the Appendix 1).

On the other hand, since Minister of State Apparatus and Bureaucratic Reform has enacted regulation No. 38 of 2017 regarding Competency Standards for State Civil Apparatus both for technical and managerial skills, despite its insignificance according to EFA method, DGT Regulation No. 12/PJ/2018 with regard to technical competency evaluation needs to be changed so that performance evaluation would be better and considered fair by employees if it is conducted based on, among other things, technical and managerial competencies of every employee. This is also true for the variable B6 (providing guidance, i.e., coaching, mentoring and counseling on the results of employee performance). Although EFA method considers this variable insignificant (see in Appendix 1), since coaching for performance is mandatory for every manager at MoF as regulated in the ministerial decree No. 590/KMK.01/2016, then every manager is still obliged to do so, especially for employees whose performance is not include in the S and/or A categories. However, this would be very difficult for every manager if the employee's original performance is good but is forced to be excluded from category S, A, B and/or C due to a forced ranked system. As a result, this forced ranked system needs to be reconsidered. If this forced rank mechanism relates to allowance or annual budget, then the revised category of SABCD could be mainly due to excellent, moderate and poor performances, while the allowance depends on the availability of budget. These two practices would be similar to other units within the MoF.

As for the effectiveness of the performance appraisal model represented by C variables, only variables C4 (informing performance appraisal by supervisor) and C6 (coaching by supervisor for aspects need to improve) are considered insignificant. Variable C4 was proposed by Widiatmanti (2020) in order to measure the fairness of DGT PMS and therefore is not enacted yet in the DGT Regulation No. 12/PJ/2018. Since according to EFA this variable is considered insignificant, we do not propose to be included in the revision of DGT Regulation No. 12/PJ/2018. As for variable C6, although insignificant according to EFA method, most manager have already coached their subordinates as indicated in Appendix 1 who chose 3 (agree) and 4 (very agree) with median and mode of 3, respectively (see in the Appendix 1). In addition, as regulated in the ministerial decree No. 590/KMK.01/2016, this is an obligation for every manager in the MoF, especially for employees whose performance are not S nor A categories.

Lastly, for the performance appraisal regulatory variables represented by D variables, only variables D1 (aligning KPI with organizational goals) and D4 (aligning development that employees get with their competency gap) are considered insignificant. Variable D1 and D4 were proposed by Widiatmanti (2020) in accordance with Blackman et al. (2018) who argue that PMS have to be operated as part of the organization's core business to be effective, and with the Government Regulation No. 17 of 2020 that link the merit system with the competency development. As with variable C4, these two variables are also not enacted yet in the DGT Regulation No. 12/PJ/2018. Since according to EFA these variables are considered insignificant, we do not propose variable D1 to be included in the revision of DGT Regulation No. 12/PJ/2018. Yet, in order to comply with Government Regulation No. 17 of 2020, we propose that the revision to DGT Regulation No. 12/PJ/2018 includes this D4 variable.

The second factor relates to variables A9 (performance appraisal with rating scale according to the role in the organization is objective), C7 to C11 (performance status SABCD, and Stage 1 and Stage 2 Rating Sessions), D3 (performance appraisal is able to distinguish between competent and incompetent employees), and D5 to D10 (the performance appraisal regulatory variables). Variable A9 can be included in either first factor or the second one. However, in the first sector it has bigger loading factor, i.e., 0.613 vs. 0.551. This means that correlation value of variable A9 with factor 1 is bigger than that of variable A9 with factor 2, indicating that the sample variation in factor 1 can be better explained by using variable A9 in the straight-line model than that in factor 2.

As described earlier, the effectiveness of the performance appraisal model represented by C variables, all variables are considered significant except variables C4 (informing performance appraisal by supervisor) and C6 (coaching by supervisor for aspects need to improve). While variable C4 can be neglected since it was proposed by Widiatmanti (2020) and it is not enacted yet in the DGT Regulation No. 12/PJ/2018, variable C6 relates to ministerial decree No. 590/KMK. 01/2016 that instructs every manager in the MoF to coach his or her subordinates. As such, this variable still important at DGT, especially for managers whose subordinates' performances are not S nor A categories.

The last factor relates to variable B3 (performance is assessed quarterly, semiannually or annually), C1 to C3 and C5 (effectiveness of the performance appraisal model). Other variables in group B are considered insignificant according to EFA method, i.e., B4, B5, and B6 described earlier. Only variable B6 (providing guidance, i.e., coaching, mentoring and counseling on the results of employee performance) that should be kept since it relates to ministerial decree No. 590/KMK. 01/2016 which every manager must obey, especially whose employees performances are not S nor A categories.

The exploratory factor analysis calculated above shows that statistically the DGT Regulation No. 12/PJ/2018 is effective and satisfactory so that it does not need to be changed or replaced. However, due to Regulations No. 38 of 2017 of the Minister of State Apparatus and Bureaucratic Reform that asks government units to cover both technical and managerial skills, Government Regulation No. 17 of 2020 that link the merit system with the competency development, as well as ministerial decree No. 590/KMK. 01/2016, we suggest that DGT Regulation No. 12/PJ/2018 need to be adjusted. Employees who are categorized as Cs and Ds in the SABCD performance status are obliged to participate in the training program. This adjustment can also be made by issuing a Circular Letter or other directive from the Director General of Tax that asks every supervisor to assign his or her subordinates participate in the training program without changing or replacing the original DGT Regulation No. 12/PJ/2018. However, since employees who are categorized as Cs or Ds are not necessarily due to bad performances but because of forced ranked, we suggest that forced rank in the DGT Regulation No. 12/PJ/2018 is abolished. The revised category of SABCD could be mainly due to excellent, moderate and poor performances, and the annual allowance depends on the availability of budget. These two practices would be similar to other units within the MoF.

5. CONCLUSIONS AND RECOMMENDATIONS

This research shows that most DGT employees participated in this research generally agree and strongly agree with DGT Regulation No. 12/PJ/2018 regarding performance evaluation system. Although there are some employees who chose disagree and very disagree in every question in the questionnaire, in general this regulation is considered quite effective and quite satisfactory. However, the ANOVA method used in this research reveals that there are differences among the different categories of DGT employees indicating that its effectiveness and satisfaction are different between men and women category, between married and unmarried employees, among different educational levels, and among different ranks and positions.

Out of 38 questions (variables), only three variables that the majority of respondents chose strongly disagree (1) and disagree (2). Two variables relate to performance status of SABCD, and the other one relates to the level 1 and level 2 ranking sessions. Using EFA method, these three variables are considered significant. This indicate that DGT should reconsider the use of forced rank in the SABCD performance status since this method although motivates employees to work better than the others (variable C7), the majority of respondents disagree and very disagree that this this method motivates them improving group performance (variable D8). Therefore, we suggest that the use of forced rank in the SABCD performance status is abolished and be replaced mainly due to excellent, moderate and poor performances, similar to other units within the MoF

that has three categories, namely excellent (green), moderate (yellow), and poor (red). If the forced rank mechanism is due to the limited annual budget, then we suggest that the provision of additional performance allowances is adjusted to the availability of the budget, also similar to other units within the MoF.

The EFA method also shows that seven variables are insignificant in the DGT performance evaluation system according to DGT Regulation No. 12/PJ/ 2018, namely variables B4 (performance assessment must be carried out using technical and managerial ability assessment), B5 (performance appraisal is done using self-appraisal method), B6 (supervisor always provides guidance (coaching, mentoring and counseling) on the results of my performance), C4 (the results of the performance appraisal are always informed by superior), C6 (superior coaches and tells what aspects need to improve), D1 (performance appraisal at DGT is aligned between the KPI and the organization's goals) and D4 (the development that employee gets from the institution is always in accordance with the competency gap). This indicates that the DGT Regulation No. 12/PJ/2018 does not need to be changed, especially with regards to the above variables. However, there are three regulations which according to the authors make the DGT Regulation No. 12/PJ/2018 needs to change, namely regulations No. 38 of 2017 of the Minister of State Apparatus and Bureaucratic Reform that asks government units to cover both technical and managerial skills, Government Regulation No. 17 of 2020 that link the merit system with the competency development, and the ministerial decree No. 590/KMK. 01/2016 that asks every manager in the MoF to coach his or her subordinates with regard to performance. As such, we suggest that DGT Regulation No. 12/PJ/2018 be amended or modified to suit the three regulations as well as to abolish the forced rank mechanism.

6. LIMITATION OF THIS RESEARCH

This research has several limitations. Although the number of samples is considered big and above the minimum samples according to Slovin formula for 5% tolerable error, the proportion is relatively low compared to the total DGT employees of 45,948 (Biro SDM, 2020). Had the sample been more than 1,587, the results might have been different. In addition, this research cut the category of employees with regard to tenure and age since there are some respondents who filled the questionnaire suspiciously. Had these respondents were singled out instead the two categories, the results might have been different.

REFERENCES

[1] Biro Sumber Daya Manusia Kementerian Keuangan. (2020). *Data Pegawai per 1 Juni 2020*. www.sdm.kemenkeu.go.id.

[2] Blackman, D., Buick, F. & O'Donnell, M. (2018). Why performance management should not be like dieting. *Australian Journal of Public Administration*, *76*, 524–280.

[3] Dessler, G. (2020). *Human resource management*. Pearson Education, Inc.

[4] Dewettinck, K. & van Dijk, H. (2012). Linking Belgian employee performance management system characteristics with performance management system effectiveness: Exploring the mediating role of fairness. *International Journal of Human Resource Management*, *24*, 806–25.

[5] Evita, S. N., Muizu, M. O. Z. & Atmodjo, R. T.W. (2017). Penilaian kinerja karyawan dengan menggunakan metode behaviorally anchor rating scale dan management by objectives (Studi kasus pada PT. Qwords Company International). *Pekbis Jurnal*, *9*(1), 18-32.

[6] Grace-Martin, K., (2021). *The problem with using tests for statistical assumptions*. The Analysis Factor. https://www.theanalysisfactor.com/the-problem-with-tests-for-statistical-assumptions/.

[7] Hair, J. F. J., Black, W. C., Babin, B. J., Anderson, R. E., Black, W. C., & Anderson, R. E. (2019). *Multivariate data analysis. cencage learning.*

[8] Harbi, S. A., Thursfield, D., & Bright, D. (2017). Culture, wasta and perceptions of performance appraisal in Saudi Arabia. *The International Journal of Human Resource Management*, *28*(19), 2792–2810.

[9] Hutcheson, G.D. & Sofroniou, N., 1999. *The multivariate social scientist: introductory statistics using generalized linear Models*. Sage Publications Ltd.

[10] Law of the Republic of Indonesia Number 5 of 2014 concerning State Civil Apparatus.

[11] Mathis, R. L., Jackson, J. H., Valentine, S. R., &

Meglich, P. (2017). *Human resource management*. Cengage Learning.

[12] McClave, J.T., Benson, P.G. and Sincich, T., (2018). *Statistics for business and economics* (13th ed.) Pearson Education, Inc.

[13] Navarro, D.J. & Foxcroft, D.R., (2019). *Learning statistics with Jamovi: A tutorial for psychology students and other beginners*. Learn Stats with Jamovi. https://www.learnstatswith-jamovi.com/.

[14] Regulation of the Government of the Republic of Indonesia Number 11 of 2017 concerning Management of the State Civil Apparatus.

[15] Regulation of the Government of the Republic of Indonesia Number 17 of 2020 concerning Replacement of the Regulation of the Government of the Republic of Indonesia Number 11 of 2017 concerning Management of the State Civil Apparatus.

[16] Regulation of the Minister of State Apparatus and Bureaucratic Reform Number 38 of 2017 concerning Competency Standards for State Civil Apparatus Positions.

[17] Schindler, P. S. (2019). *Business research methods*. McGraw-Hill Education.

[18] Sharma, N. P., Sharma, T., & Agarwal, M. N. (2016). Measuring employee perception of performance management system effectiveness. *Employee Relations*, *38*(2), 224–247.

[19] Torrington, D., Hall, L., Taylor, S., & Atkinson, C. (2020). *Human resource management*. Pearson.

[20] Widiatmanti, H. (2020). Model sistem penilaian kinerja pegawai DJP. *Prosiding kajian akademis pusdiklat pengembangan sumber daya manusia Kementerian Keuangan* (pp. 89).

[21] Zaiontz**,** C., 2017. *Multivariate central limit theorem*. Real Statistics. https://www.real-statistics.com/multivariate-statistics/multivariate-normal-distribution/ multivariate-central-limit-theorem/.

APPENDICES

Appendix 1 Questionnaire Results and their Medians and Modes
Source: Widiatmanti (2020)

Vari-ables	Statements	Respondents Choices					
		1	2	3	4	Medians	Modes
A1	My performance has been appraised based on the work I have done	85	209	768	525	3	3
A2	I am satisfied with the results of my performance appraisal	84	220	742	541	3	3
A3	Evaluation of employee performance has shown the quantity of work/activities objectively	116	318	745	408	3	3
A4	Evaluation of employee performance has shown the quality of work/activities objectively	101	330	748	408	3	3
A5	Performance appraisal is carried out according to the work plan that has been agreed with superior	67	196	773	551	3	3
A6	Employee performance appraisal has been used as a basis for competency development	124	339	702	422	3	3
A7	Employee performance appraisal has been used as a basis for career development	119	352	697	419	3	3
A8	Performance appraisal has become an objective performance criterion in accordance with organizational goals	131	391	682	383	3	3
A9	Performance appraisal with rating scale according to the role in the organization is objective	223	410	618	336	3	3
A10	Performance appraisal is carried out by comparing the work results with minimum work standards using clear parameters	110	371	732	374	3	3
A11	My supervisor has transparently provided information on the results of the performance appraisal	149	309	687	442	3	3
B1	Performance is assessed based on the work plan, and is always given feedback by superior	65	257	824	441	3	3
B2	Currently, performance is assessed using a rating scale from the lowest to the highest (scoring scale 0 to 100)	41	134	780	632	3	3
B3	Performance is assessed within a certain period of time (quarterly, semiannually or annually)	16	50	666	855	4	4
B4	Performance assessment must be carried out using technical and managerial ability assessment	93	271	694	529	3	3
B5	I will be happy if the performance appraisal is done by self-appraisal.	92	329	731	435	3	3
B6	My supervisor always provides guidance (coaching, mentoring and counseling) on the results of my performance	57	273	752	505	3	3
C1	My performance planning set out in the KPI is in line with my main role in achieving organizational goals	31	186	910	460	3	3
C2	I can easily get information about the results of the KPI achievements through the e-performance application	47	163	780	597	3	3
C3	The SIKKA application for inputting Work Targets is very easy and practical	25	154	807	601	3	3
C4	The results of the performance appraisal are always informed by superior	119	378	701	389	3	3
C5	Performance appraisal guidelines and parameters are clear and understandable	91	347	754	395	3	3
C6	My superior coaches and tells what aspects need to improve	85	343	740	419	3	3
C7	The categorization of performance status S, A, B, C, D motivates me to work better than others	0	368	537	332	3	3
C8	Using normal curve for ranking employee performance status S, A, B, C, D is correct	450	477	445	215	2	2
C9	The Stage 1 Rating Session by the Chairperson and the Raters held in the office always goes well and does not cause turmoil	189	487	667	244	3	3
C10	The Stage 2 rating session by the Chairperson and the Raters held in echelon 2 units always goes well and does not cause turmoil	171	486	698	232	3	3
C11	The results of the Stage 1 and Stage 2 Rating Sessions have been publicly informed to employees	346	539	499	203	2	2
D1	Performance appraisal at DGT is aligned between the KPI and the organization's goals	66	250	820	451	3	3
D2	The performance appraisal results improve knowledge and skills to be even better at work	75	236	730	546	3	3
D3	Performance appraisal is able to distinguish between competent and incompetent employees at work	216	509	639	223	3	3
D4	So far, the development that I get from the institution is always in accordance with the competency gap	106	444	785	252	3	3
D5	Performance appraisal has objective criteria and shows high consistency in its assessment	134	444	726	283	3	3
D6	Performance appraisal system results are accurate and reliable	177	514	650	246	3	3
D7	Performance appraisal are accepted by all employees (superiors and subordinates)	152	462	712	261	3	3
D8	Employee performance status (performance categories S, A, B, C, D) motivates me to improve group performance	437	407	477	266	2	3
D9	Performance appraisal is easy to understand, simple and uncomplicated in its application	197	457	663	270	3	3
D10	Performance appraisal is transparent and informative for employees	226	484	618	259	3	3

Appendix 2 VIF and Tolerance Values
Source: Survey Results

Variables	VIF	Tolerance	Variables	VIF	Tolerance	Variables	VIF	Tolerance	Variables	VIF	Tolerance
A1	1.90	0.526	A11	1.45	0.690	C4	1.45	0.690	D2	1.44	0.694
A2	1.84	0.543	B1	1.57	0.637	C5	1.49	0.671	D3	1.50	0.667
A3	1.74	0.575	B2	1.44	0.694	C6	1.51	0.662	D4	1.48	0.676
A4	1.80	0.556	B3	1.36	0.735	C7	1.52	0.658	D5	1.72	0.581
A5	1.56	0.641	B4	1.14	0.877	C8	1.54	0.649	D6	1.81	0.552
A6	1.62	0.617	B5	1.13	0.885	C9	2.00	0.500	D7	1.59	0.629
A7	1.59	0.629	B6	1.38	0.725	C10	2.02	0.495	D8	1.49	0.671
A8	1.68	0.595	C1	1.37	0.730	C11	1.46	0.685	D9	1.54	0.649
A9	1.54	0.649	C2	1.35	0.741	D1	1.42	0.704	D10	1.60	0.625
A10	1.57	0.637	C3	1.32	0.758						

Appendix 3 MSA Test Results
Source: Survey Results

Variables	MSA	Variables	MSA	Variables	MSA	Variables	MSA	Variables	MSA
A1	0.968	A9	0.990	B6	0.968	C8	0.969	D5	0.986
A2	0.967	A10	0.991	C1	0.979	C9	0.944	D6	0.983
A3	0.983	A11	0.983	C2	0.962	C10	0.940	D7	0.992
A4	0.981	B1	0.987	C3	0.966	C11	0.983	D8	0.969
A5	0.989	B2	0.974	C4	0.975	D1	0.986	D9	0.980
A6	0.984	B3	0.960	C5	0.983	D2	0.989	D10	0.980
A7	0.977	B4	0.955	C6	0.969	D3	0.987		
A8	0.986	B5	0.951	C7	0.960	D4	0.987		

Appendix 4 Component Matrix of Exploratory Factor Analysis
Source: Survey Results

Variables	Factors			Communalities	Variables	Factors			Communalities
	1	2	3			1	2	3	
A1	0.794			0.773	C3			0.671	0.508
A2	0.767			0.725	C4				0.533
A3	0.745			0.755	C5			0.574	0.656
A4	0.770			0.772	C6				0.554
A5	0.709			0.689	C7		0.679		0.568
A6	0.687			0.693	C8		0.768		0.675
A7	0.618			0.624	C9		0.679		0.625
A8	0.677			0.738	C10		0.687		0.630
A9	0.613	0.551		0.706	C11		0.683		0.603
A10	0.658			0.699	D1				0.569
A11	0.562			0.587	D2	0.538			0.587
B1	0.610			0.662	D3		0.570		0.624
B2	0.549			0.518	D4				0.579
B3			0.501	0.422	D5	0.544	0.560		0.719
B4				*0.123*	D6	0.539	0.60		0.751
B5				*0.106*	D7		0.594		0.698
B6				0.454	D8		0.728		0.627
C1			0.615	0.534	D9		0.579		0.619
C2			0.685	0.533	D10		0.619		0.678

Notes: Extraction Method: Principal Axis Factoring; Rotation Method: Varimax