A Development of Ethical Leadership, and the Cause-and-Effect Factors of Ethical Leadership Scales for Students at Thailand National Sports University

Jutatip Suangsuwan

Faculty of Education, Thailand National Sports University, Samut Sakhon Campus
E-mail : jutasu5@hotmail.com

Abstract

This secondary research developed student ethical leadership and its causes and effects scales. There were six objectives: 1) To examine the construct validity of the ethical leadership; and its cause and effect; 2) To examine the discriminatory power of the developed scales; 3) To create T-normal criteria for measuring the variables; 4) To examine the slope parameters of each item and the Threshold values of variables; 5) To examine the different functions of items on each variable; and 6) To compare the number of validated items using Classical Testing Theory (CTT) and Item Response Theory (IRT) methods. The secondary data set comprises 1,048 students from the 16 campuses of Thailand National Sports University (TNSU). The research procedures were as follows: 1) Reviewing research documents regarding the meaning components of ethical leadership, its causes and effects in primary research; 2) Carrying out data analysis to answer the research objectives, including analyzing the difference in mean scores using the independent t-test; 3) Checking the structural validity of the variables by confirmatory factor analysis; 4) Creating measurement criteria by finding normalize T-score; 5) Analyzing according to the IRT concept by examining the joint slope parameters of the items and its Threshold values of each answer item in each variable with Multilog program; 6) Checking the different functions of each item with the SIBTEST program; and 7) Comparing the number of items that pass the criteria of CTT and IRT using t-test dependent.

The main research results can be summarized as follows: (1) All 62 variables are validated: the index values of all models were not statistically significant. (2) The developed scales have discriminatory power. The results of testing the difference in the mean scores of the group of people with high, and low score for each variable show that t-values for each variable and its items are statistically significant (p value <.01). (3) The normalized T score for all 62 variable scales in the raw score ranged from 2 to 20 points. The normalized T score ranged from 2.25 to 65.24. Sample characteristics that be measured can be classified into three groups, namely, those with low, medium, and high measurement score. (4) The common slope parameters of each item assessment item and the Threshold value of each answer item for all 211 items of 62 variables have β values distributed over a range of χ^2. The slope parameters of

the item ranged from 2.17 to 6.86, and the Threshold values of each answer item were $\beta_1<\beta_2<\beta_3<\beta_4$. (5) Items in each variable classified according to the background of the sample (e.g., gender, faculty, year of study, region, etc.) reveal that there were 13 items biased toward gender, 19 items biased toward faculty, 16 items biased toward year of study, and 7 questions biased toward region. (6) Questions examined by analysis using CTT and IRT methods indicate that 27 of 62 variables had the same number of questions. There were 33 variables with no statistically-significantly difference, and there were only two variables with a distinct difference in their average number of scales.

Keywords: 1. Ethical leadership; 2. Cause and effect factors of ethical leadership; 3. Scales development

Introduction

Ethical leadership is an issue that has been continuously discussed and given importance to from the past to the present. Various organizations have a variety of ethical problems in the organization. Those problems result in the need for continuous development and promotion of morality and ethics in organizations and society. This is especially important regarding the morality and ethics of executives or people who are leaders in various sectors. Choojeen (2019) reported that there is a lack of ethics among executives (e.g., unfairness in salary promotions, position transfers, etc.), and there are reports of misconduct regarding disciplinary violations and corruption. Hassan et all (2024) revealed the role of ethical leadership in corporate governance through a systematic review. They found the complex connection between ethical leadership and corporate governance, highlighting the crucial role that leaders play in promoting ethical conduct and decision-making in their organizations.

Takong & Jariyawat (2018) said that the world today is facing problems. Many leaders steer their organizations toward success without moral principles, and that makes them unable to be a successful leader. The problem of a lack of morality or a decline in the morality and ethics of administrators or leaders is, therefore, something that is called for. These phenomena show that various societies and organizations face problems with both leaders and personnel in organizations related to lack of adherence to or not prioritizing morality and ethics. This causes ethical problems where the standards of the meaning of '*good*', '*bad*', '*right*', and '*wrong*' have changed from the original (Wongpitakkul, 2019; Panphet, 2022).

A key attribute of an accepted leader is someone who excels in performing tasks, and acts as a paragon of goodness. There is an important basis behind the success of being a good leader, which is the matter of behaving according to the principles of morality and ethics, combined with the knowledge and ability to mold others into being examples of good behavior. People with good intelligence, ability, and good vision can make organizational management successful. These leaders help develop the organization to facilitate progress and constructive change with their management ability. People with morality and ethics are able to motivate

others, and serve as a role model for others to trust and accept. A good leader will motive others and bring out the best in co-workers and underlings. Such leaders usually possess intellectual aptitude along with moral and emotional intelligence (Chaichompu, 2014, and Bunruang, Manokan & Thongngok, 2022).

The development of measurement tools is part of the science of education that can be used as a research methodology. In this research, the aim was to develop measurement and evaluation tools based on two main theories: Classical Test Theory (CTT) and Modern Test Theory. This approach can be divided into the following two theories: Generalizability Theory (GT), and Item Response Theory (IRT). The important goal of theories is to develop tools to measure the variables or characteristics which need to be measured. They have the same important processes and steps as research in general. The developed theories aim to develop instruments with important qualities such as validity, reliability, etc. The Modern Test theory aims to enable more accurate measurement results by including general implications of actual scores and test quality according to various conditions, including a description of respondent behavior. However, all three theories affect the analysis of tool quality because each theory has different basic assumptions (Kanchanawasi, 2007).

Classical Test Theory emphasizes checking the validity and reliability of questions or assessment items with the basic operation for developing measurement and evaluation tools, such as content validity analysis based on expert opinions, or checking internal consistency reliability that may change depending on the context and sample size, and uses advanced analytics including Factor Analysis to check construct validity. The concept of tool development based on traditional testing theory is still widely used today. However, the development of conceptual tools still has disadvantages due to the limitations of the theory, i.e., that there is a preliminary agreement that the measurement error score that has a unique error, causes the analysis to determine the reliability of the test to be analyzed under one source of error at a time. That fact is inconsistent with the natural conditions of the measurement to increase accuracy and be more consistent with the actual situation. By contrast, IRT emphasizes the importance of checking reliability, validity of instruments. That it has fewer preliminary assumptions in a measuring action shows that there will be more than one source of error. In addition, It also emphasizes verification of the different functions of the questions to explain the relationship of internal characteristics or abilities that exist in individuals. This refers to test response behavior in the form of a mathematical function called ‘test response model’ or ‘test characteristic function’ (Item Characteristic Function) or ‘test characteristic curve’ which is characterized by an S-shaped curve. These functions can show the relationship between the probability of choosing an answer with the hidden characteristics that influence an answer (Kanchanawasi, 2007; Sakolkijrungroj, 2015)

Establishing normal criteria by finding normal T values is a step taken to standardized questions or assessment items that can be used to interpret scores obtained on a test or

assessment. It informs the level of each person's characteristics, because interpreting measurement results from raw scores cannot provide complete meaning on its own (Supanprakan, 2012). Therefore, in developing measurement tools, it is necessary to create normal T criteria by using data obtained from groups, and using statistical methods to develop normal criteria. The popular methods of normal criteria development are percentiles and normal t-score criteria.

It is important to understand ethical leadership through measurement and evaluation research based on CTT. This involves the study of the elements in order to understand the characteristics of ethical leadership, whose elements necessarily contribute to ethical leadership, or the characteristics of ethical leadership. Edmonson, Fisher & Polnick (2003) stated that the studies of measuring or measuring instruments for ethical leadership create understanding of ethical leadership. The results of these study give us an idea of the status quo of ethical leadership. What is more significant is to gain information for decision-making in development plan, and organizing programs or activities to promote ethical leadership as well. Trivedi (2020) mentioned that theoretically modeled and validated scales using statistical methods result in explaining or evaluating the measured characteristics that have been verified to be consistent with real conditions and has validity in the measurement tools.

There are many studies which focused on the development of measurement tools, measurement models, and ethical leadership components. Most research focuses on the components of ethical leadership, both in the form of quantitative and qualitative research. For quantitative research, most research focuses on component development using Factor Analysis, and check the quality of the tools developed using the process according to the CCT. Supanprakan (2012) showed the development of an ethical leadership scale based on the concept of IRT combined with CTT. That study examined the 20 components of ethical leadership, checking content validity, verification of conditional validity, and reliability using CTT. The study checked the validity of the IRT by finding the joint slope parameter of the questions and the Threshold value of each answer item before checking the different functions of the questions, including creating normal criteria for measurement tools by creating normal criteria with t-scores.

Suangsuwan (2022) studied various variables related to ethical leadership that covers ethical leadership cause and effect variables factors. The research aimed to develop indicators and cause-and-effect models of ethical leadership of the students at TNSU. The research tool in that study was a scale for measuring 21 ethical leadership components, and 28 causes and 13 effects variables of ethical leadership. The model consists of three large groups of characteristics: ethical leadership, its causal factors, and its effects factors. In that research, the researchers developed a measuring instrument themselves, and examined the quality of the instrument in terms of content validity by having three experts check the consistency between the questions or assessment items and the variable definitions. The developed tool was tested

with a sample of 30 people and analyzed for consistent reliability within the Alpha Cronbach test. The results of the analysis revealed that the developed tool had content validity, and accuracy was at an acceptable level.

Suangsuwan (2022) developed the elements to explain ethical leadership and checking the quality of the instruments at the preliminary level, including checking the quality of validity and reliability according to the concept of CTT without examining the construct validity, or carrying out the examination of measuring instruments based on the concepts of IRT and criteria development. This not provide the complete or accurate ethical leadership scales in terms of validity, and standardized measuring tools.

Therefore, this secondary research utilized data from research on indicator development, and the cause-and-effect models of ethical leadership among students at TNSU (Suangsuwan, 2022). This research aimed to develop ethical leadership, and identify cause and effect of ethical leadership using knowledge about measurement and evaluation based on the concepts of CTT and IRT. Additionally, the aim was to develop criteria to use in developing ethical leadership and its cause-and-effect factors for the benefit of obtaining a robust tool to measure ethical leadership characteristics and other related variables.

Research objectives

There are six research objective as follows: 1) To examine the construct validity of student ethical leadership and its causes and effects scales to measure at TNSU; 2) To examine the discriminatory power of the developed scales; 3) To create normalized T criteria for measuring theses variables; 4) To check the slope parameters of each question or assessment item and the Threshold value of each answer item in each variable; 5) To examine the different functions of the questions or individual assessment items on each variable; and 6) To compare the number of valid test items analyzed using CTT and IRT method.

Research scope

1. Scope of population research: The target population for this research is students at TNSU, Academic Year 2022. There are approximately 13,000 students from 17 campuses, and the graduates are expected to become professionals related to physical education and sports science in various organizations.

2. This secondary research used data from research on a development of indicators. The cause-effect model of ethical leadership of students at TNSU focuses on 62 variables with totaling 211 items, that were used as a foundation of this scale development.

3. This research is interested in studying the development of instruments using CTT and IRT methods to provide information for further scale development. Therefore, this scale development involved more steps than the development of instruments developed using CTT

or IRT methods. The conclusions only reveal the difference in the number of questions or assessment items available from two theories. It does not include finding a relationship between the analytical results obtained for each item of the two theories.

4. In terms of content, the variables of interest in this research are in accordance with the study framework of Suangsuwan (2022). The variables consist of three main components: 1) Components of ethical leadership, consisting of 21 variables; 2) Causes of ethical leadership factor, consisting of 28 variables; and 3) Effect of ethical leadership factor; consisting of 13 variables.

Research Procedure

Research population and sample. The population is year 1-4 students of TNSU for Academic Year 2022 from the following three faculties: Faculty of Science and Sports, Faculty of Liberal Arts, and Faculty of Education from 16 university campuses. The sample group that provided the secondary data used in this research comprised 1,048 students.

Data analysis. This quantitative research using primary data was conducted through seven main steps that are aligned with the research objectives, as follows:

1. Basic data analysis. This part of the analysis was conducted to analyze the data and groups of informants in order to know the characteristics of such groups of people. The general data analysis regarding the sample of 1,048 people involved creating a data log file, recording data, preparing a sample data file using statistics including, frequency percentages, means, and standard deviation. Scores of 62 variables were analyzed through means and standard deviation.

2. Construct validity of the 62 variable scales was examined by studying its components, meaning of variables from the research concept of Suangsuwan (2022). This step included checks of the quality of the instrument regarding construct validity through Confirmatory Factor Analysis (CFA) with SEM, and using χ^2 and the index to measure the harmony of the model, adjust assumptions of the model to be consistent with empirical data, and report the results of examining the construct validity of all 62-variable models. According to Hair et all (1998), there is no exactly rule for identifying sample size for SEM analysis. SEM analysis is robust and the model is generalizable to another sample from that same population (Hair et all., 2021). This research randomly selected about 100 cases to validate measurement model with statistics package program (Suangsuwan, 2008). The cases random identified 107 sample for this research analysis step.

3. Verification of the discriminatory power of measuring instruments by specifying the sample for analysis. For this stage, the researchers used the Jung Teh Fan's principle of selecting a sample of 27% to analyze the discriminatory power. A sample group of 284 people was obtained, sorted by the scores of those who scored on each variable from the highest to the least, creating a variable for the group of high and low scorers, which can be changed into two values, namely the high- and low-score groups. A value was assigned to 30 respondents on each variable (15 people for high/low score groups), then analyzed to find the discriminatory power

of each variable by testing the mean difference between the high-score group and the low-score group through independent t-test.

4. Creation of normal criteria for measuring instruments for variables measurement. Scores for each variable were analyzed using the Z test, creating a T -score norm by converting raw scores into T-score standards, setting normal criteria, and determining how to interpret normal scores. 1,048 sample were utilized as subject analysis.

5. Checking the slope parameters of each item. The Threshold value of each answer item in each variable was analyzed based on the IRT method. The analysis of checking the common item slope parameter and the Threshold value of each answer item (Category threshold parameter) utilized the Graded-Response Model (GRM) method in the MULTILOG program.

6. Examining the different functions of item of each variable. All scales of 62 variables were analyzed by examining the different functions of the questions (Differential Item Functioning) with the SIBTEST program.

7. Comparison of the number of calibrated items from methods of the two theories. The number of items that passed the test quality criteria of each theory using t-test dependent statistics were retained in the analysis. Data analysis steps are summarized in Table 1.

Table 1: Analysis steps, number of samples, statistics, and the obtained result

Steps	Sample Number	Statistics	Obtained result
1. Preliminary analysis	Sample with 1,048 students	Frequency, Percentage, Means, Standard deviation, and Statistics for distribution analysis	-Characteristics of sample - 62 variables Distributions
2. Checking construct validity	group of 107 students randomly selected	Developing measurements model with Confirmatory Factor Analysis (CFA).	- Results of the 62 variables scales validity
3. Verifying the discriminant power of scales	15 students with the lowest score and 15 with the highest score	Comparing the means between the group of students with highest-scoring and lowest-scoring by t-test independent	- Results of the discriminant power of each item, and each of 62 variables

Table 1: (Cont.)

Steps	Sample Number	Statistics	Obtained result
4. Developing of normal T criteria for each variable measurement	All sample with 1,048 students	- Developing normal criteria using T-score normalization.	-Criteria for interpretin scores from 62 variable measuremen scales
5. Checking the common slope	All sample with 1,048 students	- Analyzing the common slope and Threshold value of each item using	- Items with reliability and discriminant

parameters of each item		Graded-Response Model analyzed with MULTILOG	power of 62 variables measurement scales
6. Checking the different function item	All sample with 1,048 students	-Analyzing and examining the different functions of the item using the SIBTEST program	- Item bias that should be concerned or eliminated
7. Comparison of validated item number deriving from CCT and IRT methods	All sample with 1,048 students	-Comparing the means of validated item number deriving from CCT and IRT methods through t-test dependent	-Information for using measurement of 62 variable scales for student evaluation

Research results

The results of the data analysis are divided into seven parts, as follows:

1. Results of data analysis of the sample group. The majority of students were male (68.4%). Over one in four (26.8%) were Year 1 students, 18.5% were Year 2, 24.0% were Year 3, and 30.7% were Year 4 students. Just over one in three (34.4%) were in the Faculty of Sports Science, 10.4% were in the Faculty of Liberal Arts, and 55.2% were in the Faculty of Education. Results of basic statistical analysis of 62 variables found that the mean scores for each variable were at a 'moderate' to 'high' level. All 62 variables were used in developing the measurement tool ($\bar{X}$=3.88 to 4.15, S.D.=0.63-1.04). The minimum value ranged from 1.00 to 2.00, while the maximum value was 5.00, with a range from 3.25 to 4.00.

2. Results of examining the construct validity of scales found that all 62 variables had construct validity, even though the number of items in each variable was 2 to 4 items. The consistency index values of all models were not statistically significant. The value of χ^2 ranged from 0.006 to 7.978. The df ranged from 1 to 4, the P-value ranged from 0.123 to 0.999, the GFI ranged from 0.977 to 1.000, the AGFI ranged from 0.887-1.000, and the RMR ranged from of 0.001 to 0.084.

3. Results of examining the discriminatory power of the ethical leadership instrument. and factors that cause and result in students' ethical leadership at TNSU. The total of 62 variables was created by combining the summated scores and weighted scores. It was found that the mean scores of the group with low ethical leadership scores and the group with low ethical leadership scores obtained by combining both methods comprised all 62 variables. There is a statistically-significant difference. The t values for each item are <1-31.00, with df value of 14 -28. Comparing the average values for each variable reveals that the t value is <1 -424.281, df of 14-28, indicating that all 62 developed measuring instruments have discriminant power.

4. Results of normalizing the instrument to measure variables found that the scores were in the raw score range of 2 to 20 points. The normal T score ranged from 2.25 to 65.24 in measuring the characteristics that were intended to be measured. The low measurement score group, which is a group with a T score less than or equal to 30, had a T value ranging from 2.25 to 34.98. The moderate measurement score group, which is a group with a T score greater than 30 but less than 50, had a T value ranging from 35.06 to 49.99. The group with a high measurement score, which is the group with a T score greater than or equal to 50, had a T value ranging from 50.03 to 67.23.

5. Results of checking the common slope parameters of each item, and the Threshold value of each answer item showed that the β values spread across the range of θ, The common slope parameter of the questions or individual assessment items ranged from 2.17 to 6.86, and the Threshold value of each answer item had a value of $\beta_1<\beta_2<\beta_3<\beta_4$, with the value of β_1 ranging from -6.17 to - 2.89, β_2 ranging from -2.99 to -2.13, β_3 ranging from -1.17 to -0.48, and β_4 ranging from 0.22 to 0.99. It was also found that every question had the same answer selection curve, that is, people with high θ values had a probability value of selecting response Items 4 and 5 higher than response Items 1, 2, and 3.

6. Results of examining the different functions of items found that a number of exams were biased in favor of the reference group and focal group when analyzing the different functions of the questions for respondents of different genders, faculties, years, and regions. The number of items with bias are presented in Table 2.

Table 2: Number of test items in which test function was detected using IRT method.

Classification Variable	List of bias items (Item number of bias items)		
	Ethical leadership	Cause of ethical leadership Factor	Effect of ethical leadership Factor
Gender	(3 items)	(7 items)	(1 item)
Female	14	80, 81, 82, 85, 89, 98, 143, 151	-
Male	34, 39	157	177
Faculty	(7 items)	(10 items)	(1 item)
F. of Education	10, 26, 52, 54, 57	62, 64, 66, 72, 75, 103, 115, 120	205
Not F. of Education	33, 58	93, 97	-
Year of Study	(5 items)	(11 items)	(1 item)
Higher Year (3st and 4st)	34, 37	69 , 90, 104, 114, 130, 143, 145	205
Lower Year	9,18, 20	75, 122, 152, 153	-

Classification Variable	List of bias items (Item number of bias items)		
	Ethical leadership	Cause of ethical leadership Factor	Effect of ethical leadership Factor
(1st and 2nd)			
Region	(4 items)	(2 items)	(1 item)
Central	37, 42, 54	155	177
Not Central	23	87	-

6. The results of comparing the number of exams that pass the valid criteria analyzed using CTT and IRT showed that, of the 62 variables, there were 10 variable scales for which the number of scales calibrated by CTT and IRT were the same. There were seven variables in which the mean number of questions was not significantly different. In other words, most of the questions in the variables obtained from the development of the instrument using CTT and IRT had a number of items that were not statistically-significantly different. There were only two variables that are clearly different. The numbers of each item validated by CTT and ITR and the number of item comparative are presented in Table 3.

Table 3: The numbers of each item validated by CTT and ITR, and the number of item comparisons

	Components/variables	Measurement Theory				Number of Item Comparisons		
		CTT		IRT				
		No. item	$\bar{x}$ (s.d.)	No. item	$\bar{x}$ (s.d.)	t	df	p
	Ethical Leadership							
	Personnel Ethics							
1.	Diligent	2	1(0)	2	1(0)	Not able to compare		
2.	Patient	3	1(0)	3	1(0)	Not able to compare		
3.	Responsibility	3	1(0)	3	1(0)	Not able to compare		
4.	Work Commitment	2	1(0)	0	0(0)	Not able to compare**		

Table 3: (Cont.)

	Components/variables	Measurement Theory				Number of Item Comparisons		
		CTT		IRT				
		No. item	$\bar{x}$ (s.d.)	No. item	$\bar{x}$ (s.d.)	t	df	p
5.	Modesty	3	1(0)	3	1(0)	Not able to compare		
6.	Integrity	3	1(0)	2	.667 (.577)	1.000	2	0.423
7.	Ethical Persistent	3	1(0)	2	.667 (.577)	1.000	2	0.423

8.	Region Adherence	3	1(0)	2	.667 (.577)	1.000	2	0.423
	Interpersonal Ethics							
9.	Ethical Communication	3	1(0)	2	.667 (.577)	1.000	2	0.423
10.	Building trustworthiness	3	1(0)	2	.667 (.577)	1.000	2	0.423
11.	Mercy	2	1(0)	2	1(0)	Not able to compare		
12.	Caring	3	1(0)	2	.667 (.577)	1.000	2	0.423
13.	Justice	2	1(0)	1	.500 (.707)	1.000	1	0.500
14.	Modest	3	1(0)	2	.667 (.577)	1.000	2	0.423
15.	Respect	3	1(0)	2	.667 (.577)	1.000	2	0.423
	Societal Ethics							
16.	Good Family Member	3	1(0)	2	.667 (.577)	1.000	2	0.423
17.	Good Group Member	2	1(0)	2	1(0)	Not able to compare		
18.	Building Unity	4	1(0)	4	1(0)	Not able to compare		
19.	Regulation Respect	4	1(0)	2	.500 (.577)	1.732	3	0.182
20.	Social Responsibility	3	1(0)	2	.667 (.577)	1.000	2	0.423
21.	Public mind	3	1(0)	2	.667 (.577)	1.000	2	0.423
	Cause of Ethical Leadership							
	Ethical Cultivation							
1.	Ethical Model	4	1(0)	3	.750 (.500)	1.000	3	0.391
2.	Family Ethical Cultivation	4	1(0)	3	.750 (.500)	1.000	3	0.391
3.	Societal Ethical Cultivation	4	1(0)	2	.500 (.577)	1.732	3	0.182
4.	Ethical Experience	3	1(0)	2	.667 (.577)	1.000	2	0.423
5.	Ethical Culture	4	1(0)	4	1(0)	Not able to compare		
	Ethical Development							
6.	Ethical Developing	3	1(0)	0	0(0)	Not able to compare**		
7.	Ethical Learning	4	1(0)	2	.500 (.577)	1.732	3	0.182
8.	Ethical Training	3	1(0)	1	.333 (.577)	2.000	2	0.184
9.	Ethical Literacy	4	1(0)	3	.750 (.500)	1.000	3	0.391

Table 3: (Cont.)

	Components/variables	Measurement Theory				Number of Item Comparisons		
		CTT		IRT				
		No. item	$\bar{x}$ (s.d.)	No. item	$\bar{x}$ (s.d.)	t	df	p
10.	Ethical Understanding	4	1(0)	2	.500 (.577)	1.732	3	0.182
	Ethical Characteristics							
11.	Ethical Inquiry	4	1(0)	4	1(0)	Not able to compare		
12.	Ethical Sensitivity	4	1(0)	2	.500 (.577)	1.732	3	0.182

13.	Ethical Concept	4	1(0)	4	1(0)	Not able to compare		
14.	Ethical Concerns	4	1(0)	3	.750 (.500)	1.000	3	0.391
15.	Ethical Reasoning	4	1(0)	3	.750 (.500)	1.000	3	0.391
16.	Ethical Decision Making	4	1(0)	2	.500 (.577)	1.732	3	0.182
17.	Ethical Confident	4	1(0)	4	1(0)	Not able to compare		
	Psychology							
18.	Self-Efficacy	4	1(0)	3	.750 (.500)	1.000	3	0.391
19.	Self-Confident	4	1(0)	4	1(0)	Not able to compare		
20.	Optimistic	4	1(0)	4	1(0)	Not able to compare		
21.	Hope	4	1(0)	4	1(0)	Not able to compare		
	Human Relationship							
22.	Openness	3	1(0)	1	.333 (.577)	2.000	2	0.184
23.	Relationship Efficacy	3	1(0)	3	1(0)	Not able to compare		
24.	Agreeableness	3	1(0)	2	.667 (.577)	1.000	2	0.423
25.	Emotional Flexibility	4	1(0)	1	.250 (.500)	3.000	3	0.058
	Working Depositions							
26.	Performance Commitment	3	1(0)	2	.667 (.577)	1.000	2	0.423
27.	Self-Regulation	3	1(0)	3	1(0)	Not able to compare		
28.	Working Standard Setting	3	1(0)	3	1(0)	Not able to compare		
	Performance Output							
1.	Work Advancement	3	1(0)	3	1(1)	Not able to compare		
2.	Work Performance	3	1(0)	2	.667 (.577)	1.000	2	0.423
3.	Performance Effectiveness	4	1(0)	4	1(1)	Not able to compare		
	Group Level Result (Dynamic result)							
4.	Group Commitment	4	1(0)	3	.750 (.500)	1.000	3	0.391
5.	Good Membership	4	1(0)	4	1(0)	Not able to compare		
6.	Group Member maintain	4	1(0)	4	1(0)	Not able to compare		

Table 3: (Cont.)

	Components/variables	Measurement Theory				Number of Item Comparisons		
		CTT		IRT				
		No. item	$\bar{x}$ (s.d.)	No. item	$\bar{x}$ (s.d.)	t	df	p
	Leader Related Result							
7.	Leader Trust	4	1(0)	4	1(0)	Not able to compare		
8.	Leader Role	4	1(0)	4	1(0)	Not able to compare		
9.	Leadership Effectiveness	4	1(0)	4	1(0)	Not able to compare		
	Group Related Result							

10.	Voice	4	1(0)	4	1(0)	Not able to compare		
11.	Group Justice Perception	3	1(0)	2	.667 (.577)	1.000	2	0.423
12.	Effective Communication	3	1(0)	3	1(0)	Not able to compare		
13.	Group Ethical Climate	3	1(0)	3	1(0)	Not able to compare		

Note: ** means the result of the comparison is clearly different. The T value was not able to be analyzed due to zero standard deviation of the two means variables.

Discussion

The results consistent with the six research objectives are as follows:

1) Results of checking the construct validity of the 62 variables scales show that the measurement models are validated, the concordance index values of all models are not statistically significant. This shows that the developed instrument is valid and intended to measure the construct. This also means that previous research developed measurement tools and validated the measurement tools based on the basic CTT method are valid (e.g., checking for validity, and reliability). This research supports findings from previous research on quality research tools. All variables have construct validity when examining construct validity by confirmatory factor analysis or measurement model development. This result is consistent with Wiratchai (1999) who suggested that Factor Analysis and Confirmatory Factor Analysis are methods that help researchers create components from many variables, grouped together into a single component. If the model is valid, it shows that the scale is validated.

2) Results of examining the discriminatory power indicate that the scale and items had discriminatory power. This is due to the development of measurement tools in the primary research that followed the criteria and methods for developing measurement tools from defining variables, verification by experts, and checking for reliability with the Alpha coefficient. In terms of the discriminatory power of the scales developed, the verification results are at a level that passes the specified criteria, indicating that the scale has a construct validity, and discriminatory power. So, it can be used to collect data. The results of this research are in line with the idea of Saenglertuthai (2017) who presented the discriminatory power of the questionnaire by dividing the respondents into high groups and low groups and then finding the value of the t -test. If the average scores are different, it shows that the developed tool has the power to classify groups. That is, the measurement results can tell that those with high total scores indicate high levels of the trait measured by the instrument, while those with low total scores indicate low levels of the trait measured by the instrument.

3) The results of creating the normal criteria for the measurement show that the developed normal criteria could be used to classify groups of people because the sample group used in developing the criteria was large enough. This made the distribution of scores similar to a normal curve. The criteria can be used to group people according to the variables or tools you want to measure, and makes the interpretation of scores clear and able to compare individual

characteristics. This finding is consistent with the concept of Pradujprom, Pantong & Kitiyanusan (2021) who suggested that the normal criteria for measuring variables developed by researchers which is a measure that can be used to measure and evaluate students. This supports the idea of having students evaluate themselves, and realize to what extent their measurement scores are effective both in the overall picture and in each sub-aspect. The measurement results can provide suggestions for self-improvement, and make students understand their development to reach to a satisfactory level, and relevant individuals and agencies can jointly plan, set policies, or design activity curricula that will cultivate learners to develop themselves to flourish. This is consistent with the concept of Supanprakan (2012) and Phatthiyani (2003) who suggested that the normal criteria (Norm) is an important component of standardized tests used for interpreting scores obtained on standardized tests, and the level of characteristics of each person. Interpreting measurement scores from a scale or raw score does not provide complete meaning in itself. It must be considered together with related things such as the number of questions, time of measurement, precision, accuracy, and standard deviation to comparing each person's raw scores or to compare between various abilities. It cannot provide any information to determine what the measurement results reflect. Therefore, creating a score that will help interpret the meaning of the score or interpret the results obtained from the measurement will yield information that is consistent with the purpose of the measurement and is in a condition that can be useful information.

4) The results of examining the common slope parameters of questions or individual assessment items reveal that the common slope parameters of all 62 question variables have β values distributed over the range of Threshold values of each answer item. The results shows that β_1 values $< \beta_2 < \beta_3 < \beta_4$ indicates that the scales have the power to discriminate through the criteria set according to the concept of IRT. Hence, the scales can be used to collect research data or measure and evaluate results. This is consistent with the work of Supanprakan (2012), which is the development of a direct leadership measurement tool, and the quality of the measurement tool is checked using analysis based on the concept of IRT together with CTT. This is consistent with the idea of Marungruang (2012) who developed a measurement tool using the concept of test response theory. The results of the research found that, when examining the slope parameters of each question or evaluation item and the Threshold value of each answer item, it was found that if the β value spreads over the range of the joint slope parameters of the questions, and the Threshold values of each answer item are $\beta_1<\beta_2<\beta_3<\beta_4$, this indicates that the developed questions are appropriate and can be used.

5) The results of examining the different functions of 211 questions (or individual assessment items) show that the different functions of the exams are classified according to gender, faculty, Academic Year level, and region. These biased items should be improved or excluded from use in developing measurement tools because it may cause unfairness. The

results of this research are consistent with Chaiyapornpattana (2011) who suggested that if most of the questions had no difference in function between the subgroups in the sample, the questions could be appropriately used to measure students in various groups. For the questions that are biased, Marungruang (2012), proposed that we should improve or eliminate those items to ensure fairness in measurement and evaluation.

6) The results of comparing the number of items passing the criteria of CTT and IRT methods indicate that some variables had a different number of questions. This can happen because, firstly, the main concepts of two methods are similar since IRT is an additional part from CTT. Therefore, some of the results of the analysis are different, especially the analysis of test bias according to IRT that causes the number of questions on each variable to be adjusted or eliminated because they are biased toward a particular group of respondents. According to Kanchanawasi (2007), testing theory comes from the fields of education and psychology which are interested in the elements that affect measurement in various situations in order to propose measures to solve or reduce problems of measurement. The main aim of studying testing theory is to use it as a source of knowledge for understanding the measurement model, basic agreement tool development, results analysis, and application of the knowledge and understanding to help evaluators to create and develop quality tests, and be able to accurately interpret measurement results so that the findings can be used as information for appropriate educational and psychological decision-making.

Suggestions

Based on the findings of this research, the author offers the following recommendations:

1. Suggestions for improving educational policy. A purpose of this research was to develop scales for evaluating the characteristics of learners as leaders in order to improve personnel development. The evaluation results can be used to develop students in terms of the success of student development. The results in the development of scales based on the concept of IRT showed that some items of the 62 scales were biased toward gender, Academic Year, faculty, and region of the educational institution. This shows that responses to the same question item may have different scores if variables that cause bias are taken into account. Therefore, in developing other measurement and evaluation tools, we need to consider many issues for accurate assessment results as and consistent with the respondent's condition as possible, especially in evaluation measures or in research that collects data with different subgroups.

2. Suggestions for further research. This research focuses on the development of measurement tools. It does not focus on the information obtained from the measurement. In using information from measurements for further research, the items that passed the quality criteria from this research can be used in other research, such as a model for measuring the components of ethical leadership, or components of factors that are causes and effect factors

of ethical leadership. The variables score in the analysis may be a calculation of each respondent's score based on the concepts of CTT and IRT, along with comparing differences in the models. It is possible to study the model developed from the two concepts, and compare model differences. The developed model was also checked to see if there were differences between the sample groups that differed in gender, Academic Year, faculty, and region of the educational institution using multiple group strategy. In the same way, the score of the respondent or a measurement tool can be developed as a research tool that can be developed into a standard measurement, and examine the tools in various aspects based on other concepts of CTT, such as examining the relationship between the developed tools and standard tool. In addition, the research results deriving form scale development according to the concepts of CTT or IRT might be examined, as well as to provide suggestions for researchers to develop measurement, evaluation, and gain accurate research results in the future.

References

Boonrueang, K., Manokarn, M. & T. Thongngok (2017). Ethical leadership among school administraotors in Lampang Primary Educational Service Area 3. *Veridian E-Journal, Silapakorn University, 10*(3), 1409-14260.

Chaichompu, S. (2014). Ethical leadership in school administration. *Educational Administration, Burapha University, 8*(2), 1-15.

Chaiyapornpattana, N. (2011). *Development of a Multidimensional Thinking Styles Scale Based on Theory of Mental Self-Government for Sixth Grade Students.* Doctor of Philosophy Thesis, Educational Measurement and Evaluation, Education: Chulalongkorn University.

Choojeen, K. (2019). *Factors Affecting Ethical Leadership of Basic Education School Administrators Under the Office of the Basic Education Commission*. Doctor of Philosophy Thesis, Educational Administration, Education: Sukhothai Thammathirat Open University.

Edmonson, S., Fisher, A. & B, Polnick. (2003). Portrait of an ethical administrator, paper presented at *the Annual Conference of the American Association of School Administrators*, New Orleans, LA, April 21-25.

Hair, J. T., and others. (1998). Multivariate Data Analysis. 5th ed. Englewood Cliffs, NJ: Prentice-Hall.

Hair, J.F., Hult, G.T.M., Ringle, C.M., Sarstedt, M., Danks, N.P., Ray, S. (2021). An Introduction to Structural Equation Modeling. In: Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R. Classroom Companion: Business. Springer, Cham.

Hassan, S. S., Kaur, N., George, A. J., Singh, S., & David, R. (2024). Role of ethical leadership in corporate governance: A systematic review. *IUP Journal of Corporate Governance*, 23(1), 51–71.

Kanchanawasi. S. (2007). *New Testing Theory*. 3rd edition, additional revised edition. Bangkok: Faculty of Education, Chulalongkorn University.

Marungruang, P. (2012). *Development of Parenting Quality Indicators: An Analysis of Psychometric Properties Using Multigroup SEM*. Doctor of Philosophy Thesis, Educational Measurement and Evaluation, Education: Chulalongkorn University.

Panphet K. (2022). Ethical leadership based on Vuddhidhammas for school administrators in the 4.0 era. *Journal of Education Review*, *9*(1), 491-498.

Phatthiyathani., S. (2019). *Educational Measurement.* 12th ed. Karalasin: Prasan Publishing.

Pradujprom, P., Pantong, K. & R. Kitiyanusan. (2021). Development of normal T criteria for measuring growth for students. High school level. *Research and Development Institute Journal Bansomdejchaopraya Rajabhat University*, *6* (1), 147-158.

Saenglertuthai, J. (2017). Quality of tools used in research. *Journal of Curriculum Research and Development*, *7*(1), 1-15.

Sakolkijrungroj, S. (2015). *Applying Item Response Model for Developing the Thai Happiness Scale: Computer-adaptive testing*. Doctoral dissertation, Measurement and Cognitive Technology College of Research and Cognitive Sciences: Burapha University.

Suangsuwan, J. (2008). *A Development of the Indicators and the Cause and Effect Model of Collaboration of Primary School Teachers in Phra Nakhon Si Ayutthaya Province.* Doctor of Philosophy Thesis, Educational Research and Psychology, Education: Chulalongkorn University.

Suangsuwan, J. (2022). *Development of the Indicators and the Cause-and-Effect Model of Student Ethical leadership at Thailand National Sports University*. Samut Sakhon: Thailand National Sports University, Samut Sakhon Campus.

Supanprakan. S. (2012). *Development of a Moral Leadership Measurement Instrument for School Administrators*. Doctoral Degree Thesis. Educational Measurement and Evaluation, Faculty of Education, Chulalongkorn University.

Takong, S. & S. Jariyawat. (2018). Indicators of moral leadership for administrators in basic educational schools: A structural relationship model. *Nakhon Phanom University Journal*, *8*(1), 49-58.

Trivedi, C. (2020). *Measurement Model: Understanding the Error*. Derived from https://conceptshacked.com/measurement-model.

Wiratchai, N. (1999). LISREL *Model: Analysis Statistics for Research*. 3rd ed. Bangkok: Chulalongkorn University.

Wongpitakkul, C. (2019). *Strategies for Ethical Leadership Development for Police Trainees in the Upper-Northern Thailand*. Doctorate Thesis, Educational Leadership and Human Resource Development, Graduate School: Chiang Mai University.