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

# การพัฒนาเครื่องมือวัดภาวะผู้นำเชิงจริยธรรม และปัจจัยที่เป็นสาเหตุและผลภาวะผู้นำเชิงจริยธรรมของนักศึกษา มหาวิทยาลัยการกีฬาแห่งชาติ

Jutatip Suangsuwan
จุฑาทิพย์ สร้างสุวรรณ

*Article History*
Receive: February 28, 2025
Revised: June 23, 2025
Accepted: June 23, 2025

## บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อตรวจสอบความตรงเชิงโครงสร้างของเครื่องมือวัดภาวะผู้นำเชิงจริยธรรม และปัจจัยที่เป็นสาเหตุและผลภาวะผู้นำเชิงจริยธรรม ตรวจสอบอำนาจจำแนกของเครื่องมือ สร้างเกณฑ์ T ปกติเพื่อการวัดตัวแปร ตรวจสอบค่าพารามิเตอร์ความชันร่วมของแต่ละข้อคำถาม และค่า Threshold ของการตอบแต่ละคำถาม ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม และเปรียบเทียบจำนวนข้อที่ผ่านเกณฑ์ที่วิเคราะห์โดยวิธีทฤษฎีการทดสอบแบบดั้งเดิม (CTT) และวิธีทฤษฎีการตอบสนองข้อสอบ (IRT) กลุ่มตัวอย่างได้แก่นักศึกษามหาวิทยาลัยการกีฬาแห่งชาติ จำนวน 1,048 คน จากข้อมูลทุติยภูมิ การดำเนินการวิเคราะห์ข้อมูล ได้แก่ การวิเคราะห์ความต่างคะแนนเฉลี่ยด้วยวิธีกลุ่มรู้ชัด ตรวจสอบความตรงเชิงโครงสร้างโดยการวิเคราะห์องค์ประกอบเชิงยืนยัน สร้างเกณฑ์โดยหาค่า T ปกติ ตรวจสอบค่าพารามิเตอร์ความชันร่วม ค่า Threshold ของแต่ละข้อคำถาม ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม และเปรียบเทียบจำนวนข้อคำถามที่ผ่านเกณฑ์ที่วิเคราะห์โดย CTT และ IRT โดยใช้ t-test dependent ผลการวิจัยพบว่า 1) ผลการตรวจสอบความตรงเชิงโครงสร้างของเครื่องมือวัดภาวะผู้นำเชิงจริยธรรม และปัจจัยที่เป็นสาเหตุและผลภาวะผู้นำเชิงจริยธรรม พบว่าตัวแปร ทั้ง 62 ตัว มีความตรงเชิงโครงสร้าง 2) ผลการตรวจสอบอำนาจจำแนกของเครื่องมือที่พัฒนาขึ้น พบว่าเครื่องมือมีอำนาจจำแนก 3) ผลการสร้างเกณฑ์ T ปกติ พบว่า คะแนน T ปกติ จำแนกกลุ่มบุคคลเป็นกลุ่มได้ 4) ผลการตรวจสอบค่าพารามิเตอร์ ความชันร่วมและค่า Threshold ของการตอบ พบว่า ค่าพารามิเตอร์ความชันร่วมมีค่า $\beta$ กระจายครอบคลุมช่วงของค่าพารามิเตอร์ ค่า Threshold ของแต่ละรายการคำตอบมีค่า $\beta_1 < \beta_2 < \beta_3 < \beta_4$ 5) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแต่ละตัวแปร พบว่า มีข้อคำถามที่มีความลำเอียงระหว่างกลุ่มย่อย และ 6) ผลการเปรียบเทียบจำนวนข้อที่ผ่านเกณฑ์ที่วิเคราะห์โดย CTT และ IRT พบว่า ข้อคำถามตัวแปร 27 ตัว มีจำนวนข้อคำถามเท่ากัน มีตัวแปร 33 ตัว ที่ข้อคำถามมีความแตกต่างอย่างไม่มีนัยสำคัญทางสถิติ แสดงว่าทฤษฎีการวัดทั้งสองทฤษฎีให้ผลการพัฒนาเครื่องมือวัดที่ใกล้เคียงกัน

**คำสำคัญ** : ภาวะผู้นำเชิงจริยธรรม; ปัจจัยที่เป็นสาเหตุและผลภาวะผู้นำเชิงจริยธรรม; การพัฒนาเครื่องมือวัด; คุณภาพเครื่องมือวัด; วิธีการตามทฤษฎีการวัดดั้งเดิมและทฤษฎีการตอบสนองข้อสอบ

ผู้ช่วยศาสตราจารย์ คณะศึกษาศาสตร์ มหาวิทยาลัยการกีฬาแห่งชาติ วิทยาเขตสมุทรสาคร, Assistant Professor, Faculty of Education, Thailand National Sports University, Samut Sakhon Campus
*Corresponding author Email: jutasu5@hotmail.com

NPU JOURNAL

**ABSTACT**

This research aimed to examine the construct validity of the ethical leadership and its causes and effects; examine the discriminatory power of scales; create T-normal criteria for measuring variables; check the slope parameters of each item, and the threshold values of selecting each item; examine the differential item functioning (DIF) of the items; and compare the number of validated items analyzed using Classical Testing Theory (CTT) and Item Response Theory (IRT) methods. The sample comprises 1,048 Thailand National Sports University (TNSU) students from secondary data. The research procedures consisted of comparing means analysis with known group technique, validating construct validity with confirmatory factor analysis, developing normalized T-scores, examining the joint slope parameters and their threshold values of each item, checking the different functions of each item, and comparing the number of validated items between CTT and IRT methods using the dependent t-test. The results were as follows: 1) The results of examining the construct validity of student ethical leadership and its causes and effects scales indicated that all 62 variables were validated. 2) The results of examining the discriminatory power of the developed scales showed that the scales had discriminatory power. 3) The results of creating normalized T criteria revealed that the normalized T-scores were able to classify individuals into differentiate groups. 4) The results of checking the slope parameters of each item, and the threshold value of each answer item pointed out that the common slope parameters and the threshold had β values distributed over a range of parameters, the Threshold values of each answer item were $\beta_1 < \beta_2 < \beta_3 < \beta_4$. 5) The results of examining the different functions of the items on each variable notified that there were items biased toward subgroup. 6) The results of comparing the number of valid test items analyzed using CTT and IRT methods verified that 27 variables had the same number of items, there were 33 variables with no statistically significant difference in item number, which indicated that both theories provide similar results for scales developing.

**Keywords :** Ethical Leadership; Cause and Effect Factors of Ethical Leadership; Scales Development; Item Quality; Classical Testing Theory (CTT) and Item Response Theory (IRT) Methods

## 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. Lacking of ethics leading to corruption and follower untrustworthy (Phramaha Samack Atibhaddho et al., 2025). Leaders play a crucial role in promoting ethical conduct and decision-making in their organizations (Hassan et al., 2024). The problem of a lack of morality or a decline in the ethics of leaders is, therefore, something that is called for. International corruption indicators show that only 28 of the 180 countries decrease their corruption levels over the last twelve years which indicates the lack of countries justice and effective rule (Transparency International Secretariat, 2024). These phenomena show that various societies and organizations face problems with both leaders and personnel 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 (Panphet, 2022).

It is important to understand ethical leadership through measurement and evaluation based on CTT. This involves the elements study in order to understand ethical leadership, whose elements necessarily contribute to it, or its characteristics. The studies of measuring ethical leadership are significant as they create

NPU JOURNAL

understanding of ethical leadership by displaying ethical behavior (Andrich & Marais, 2019). The results of these studies 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 programs or activities to promote ethical leadership as well (Shakeel et al., 2024). 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 (Trivedi, 2020).

There are many studies which focused on the development of ethical leadership measurement tools. 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. The development of a social skill scales based on the concept of IRT combined with CTT examined the 20 items, 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 (normalized t-scores) for measurement scales (Samart et al., 2022). In addition, 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 (Phatthiyathani, 2019). 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.

In international context, the ethical leadership scales (ETL) were both based on CTT and IRT, and emphasized on administrators. Most of the research was not interested in the caparison between the scale developing based on CTT and IRT which provides information for understanding the theories relationship. Moreover, the sample for those researches were administrators, not the Physical Education students whose aim and need are to be leaders in the PE communities. Some researchers test the ETL measurement scale for Public Servant leaders using CFA based on CTT (Shakeel et al., 2024) while others might want to validate principals ETL using IRT and CTT (Sen & Gocen, 2021). These researches emphasized administrators ETL scale development and promoted trends in combining complementary methods from both classical test theory and modern test theory in validating instruments. However, the PE students, who are prospective leaders or administrators, need to be educated for their ETL to reduce misconduct, and transfer personal and social development to their future charges through PE activities (Opstoel, et al., 2020 ; Cardinal, 2023).

Previous research studied variables related to ethical leadership, its cause- and- effect factors. The research aimed to develop indicators and models of SEL at TNSU. The research tool in that study was a scale for measuring 21 ethical leadership components, and its 28 causes and 13 effects. The researcher developed a measuring instrument herself, 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. The results revealed that the developed tool had content validity, and accuracy was at an acceptable level (Suangsuwan, 2022). However, that research 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 instruments test based on IRT and criteria development. This did not provide complete ETS 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

NPU JOURNAL

leadership among students at TNSU. This research aimed to develop SEL, and cause and effect of ethical leadership scales using the concepts of CTT and IRT. Additionally, the aim was to develop criteria for 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. Therefore, this study not only demonstrates the process of instrument development grounded in established theories, but also provides evidence of the interrelationships among these theories, reaffirming their robustness and continued relevance in item development.

**Objectives**

The research objectives are 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 scales.
3. To create normalized T criteria for measuring the variables.
4. To check the slope parameters of each item and the Threshold value of each answer item.
5. To examine the different functions of the questions or individual assessment items.
6. To compare the number of valid test items analyzed using CTT and IRT methods.

## Literature Review

**Scale Development Theories**

The development of measurement tools is part of education and research methodology. In this research, the aim was to develop evaluation tools based on two main two theories: Classical Test Theory (CTT) and Item Response Theory (IRT). The important goal of theories is to develop tools to measure the characteristics which need to be measured. They have the same important processes and steps as research in general such as defining construct, planning and item developing, scale constructing, test scoring and norming, and test specificizing and implementing (Kanchanawasi, 2007 ; Irwing & Hughes, 2018). The theories aim to develop instruments with important qualities such as validity, reliability, etc.

CTT 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. It is fundamentally a test-level theory and it based on the true score (T) which is expected value score deriving from the estimates observed score (X) many times, and the measurement errors for each time are the same for all examinees. The concept of tool development based on CTT is still widely used today (Demars, 2018). 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 (Kanchanwasi, 2007)

To enhance accuracy and better align with real-world conditions, some assessment specialists have developed new testing theories that address the initial weaknesses of traditional testing theories. These can be divided into two concepts: Generalizability Theory (G theory) and Item Response Theory (IRT). Both continue to emphasize the examination of the reliability and validity of measurement tools but relax some initial assumptions. Specifically, the measurement tools are understood to involve more than one source of error.

NPU JOURNAL

IRT focuses on examining the differential functioning of test items to explain the relationship between an individual's internal trait or ability and their test responses, represented by a mathematical function called the Item Characteristic Curve (ICC). This curve typically takes the shape of an S-curve. These functions illustrate the relationship between the probability of selecting a particular answer and the latent trait that influences the response behavior (Kanchanawasi, 2007). It provides precise information on item difficulty reliability across range of scale values, and provides the basis for short test with good reliability (Irwing & Hughes, 2018). It is an item-level theory rely on a latent trait or ability proficiency, symbolized $\theta$, is posited to underlie the observed responses (Demars, 2018).

In brief, for scale development, CTT is easier to apply, and works well, but IRT offers greater precision, flexibility, and insight because CTT focuses on the overall test and assumes all items contribute equally, while IRT focuses on individual items and their difficulty and discrimination, allowing for more nuanced and precise ability estimates. However, measurement tools can be developed using different theories. To illustrate, the psychometric properties of a tool developed based on CTT depend on the group which the tool is applied to. Only one standard error value can be obtained for a whole group in the measurement tools developed using CTT. On the other hand, in IRT, item parameters are independent of the respondent group, and group characteristics are independent of the item sample. A unique standard error estimation is possible for each participant when IRT is used (Toraman & Korkmaz, 2023). IRT 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. Therefore, item analysis based on IRT report the slope parameters of each item to indicate how well an item differentiates with different latent trait levels, and the Threshold value ($\beta$) which is the parameter reflects the respondent with given trait has an equal probability of endorsing of each answer item. However, two theories affect the analysis of tool quality because each theory has different basic assumptions (Kanchanawasi, 2007)

**Conceptual Framework**

According to the research objectives and literature review, the research framework was presented in Figure 1 which showed that the scales for measuring 21 ethical leadership components, and its 28 causes and 13 effects were developed and tested qualities based on CTT and IRT methods. Finally, the items of each variable derived from CTT and IRT methods were compared to investigate their differences.

NPU JOURNAL

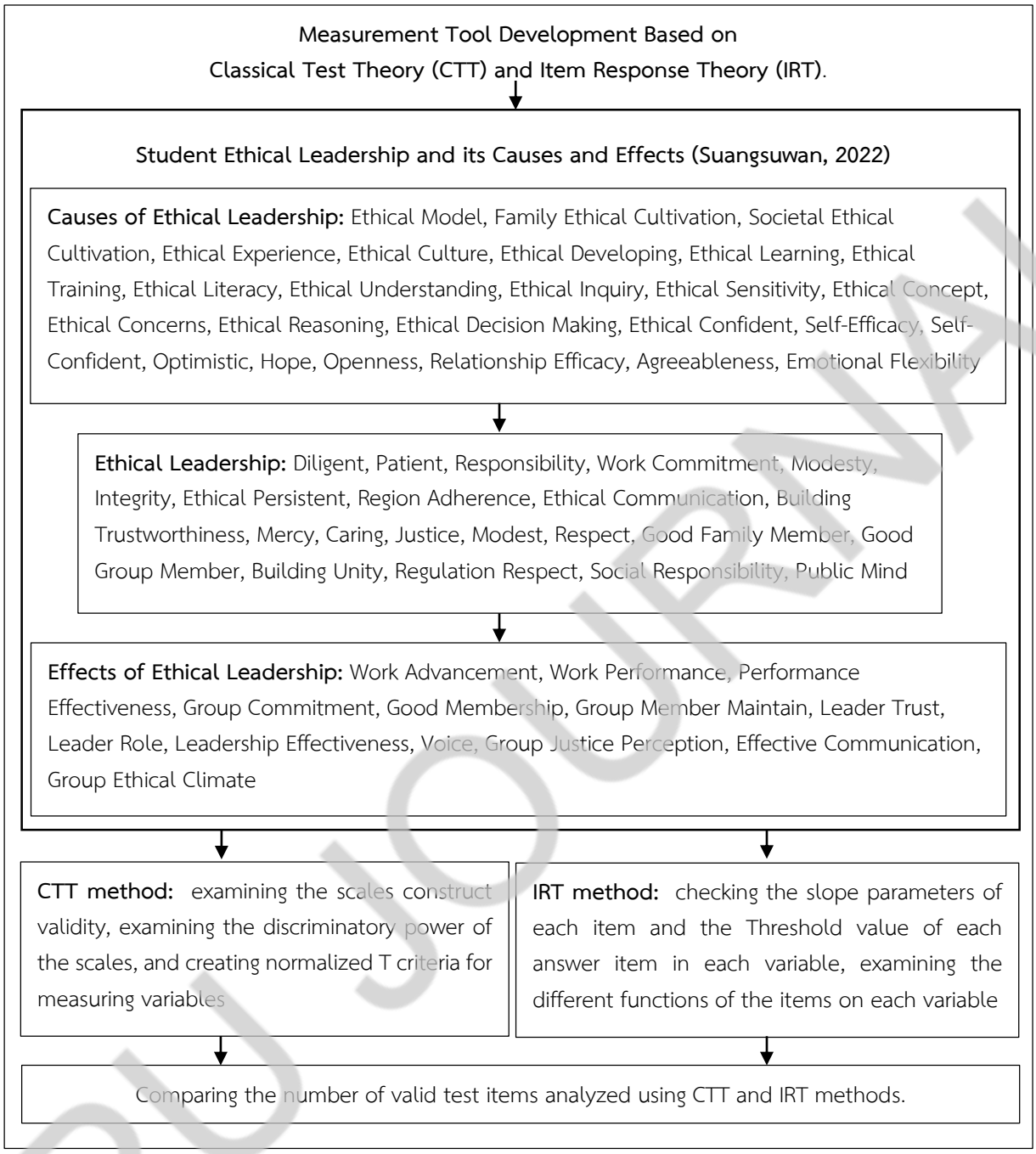



**Figure 1** Research Framework

An example of variables explanation in this study based on primary research (Suangsuwan, 2022) were presented in Table 1.

NPU JOURNAL

**Table 1** Example of Variables Explanation

| No. | Ethical Leadership Variables (Variables Explanation) |
|---|---|
| 1. | Diligent (Persistent effort; consistent and focused work; overcoming obstacles mindfully) |
| 2. | Patient (Tolerance of delays and discomfort; emotional maturity; calm under pressure) |
| 3. | Responsibility (Commitment to duties; accountable performance; perseverance despite challenges) |
| 4. | Work Commitment (Commitment to duties; accountable performance; perseverance challenges. |
| 5. | Modesty (Mindful spending; self-discipline in consumption; modest lifestyle) |
| 6. | Integrity (Integrity in action; keeping promises; consistency between words and deeds) |
| 7. | Ethical Persistent (Adherence to ethics; standing firm in righteousness; moral self-identity) |
| 8. | Region Adherence (Faith-based conduct; practicing moral teachings; spiritual discipline and control) |
| 9. | Ethical Communication (Constructive messaging with mutual understanding; transparent and ethical exchanges) |
| 10. | Building Trustworthiness (Reliable behavior; emotional consistency; inspiring confidence in others) |
| 11. | Mercy (Loving kindness; supportive attitude; seeing the good in others) |
| 12. | Caring (Empathy and support; joyful cooperation; nurturing others' growth) |
| 13. | Justice (Impartiality in actions; equal treatment; consistent ethical judgment) |
| 14. | Modest (Respectful demeanor; polite language and conduct; dignified interactions) |
| 15. | Respect (Valuing others' opinions; polite acceptance; treating everyone with dignity) |
| 16. | Good Family Member (Supportive family role; avoiding harm; fundamental ethical practice) |
| 17. | Good Group Member (Shared responsibility; collective goals; contribution to team success) |
| 18. | Building Unity (Cooperation and cohesion; shared success; harmony over division) |
| 19. | Regulation Respect (Abiding by laws; ethical compliance; principled civic behavior) |
| 20. | Social Responsibility (Community-mindedness; responsible citizenship; acting for the common good) |
| 21. | Public mind (Selfless service; helping without reward; joyful acts for public benefit) |

## Research Methodology

### Population and Sample

This secondary research used data from research on a development of indicators. The cause-effect model of ethical leadership of students at TNSU (Suangsuwan, 2022). The target population for this research is approximately 13,000 students who enrolled as TNSU student in Academic Year 2022 from 17 campuses. A sample was 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 17 campuses. The, random sampling method which is suitable for data with larger populations, and able to reduce sampling bias was utilized for that data collection.

Sample were 1,048 students from all TNSU campuses had replied entire the online form. The response rate was 8.06%. The proportion of student number from all campuses were 0.1% to 25.0%. As this research was not aimed to study the scale development for each campus. Hence, the differentiate sample size from the campuses were not related to this research analysis result.

NPU JOURNAL

**Research Instrument**

This secondary research focuses on 62 variables with totaling 211 items, that were used as a foundation of this scale development. In terms of content, the variables of interest in this research are in accordance with Prior study framework (Suangsuwan, 2022). The variables consist of three main components: ethical leadership with 21 variables; ethical leadership causes factor with 28 variables; and ethical leadership effects factor with 13 variables. The research tools were five Likert scales ranging from at least to the most. The validity testing showed that IOC values were greater than 0.5, reliability analysis with alpha Cronbach were 0.70-0.96. Only two variables showed high IOC with mediocre reliabilities (0.60 and 0.65).

**Data Collection**

Data utilized in this research deriving from ethical leadership and its causes and effects databased (Suangsuwan, 2022) for effective usage, access large data set, and broader understanding of a scale development topic. The timeframe for data collection was May to July, 2022. In that data collection, the five Likert scales had been transformed into google form and applied it for data collection. Its link had been sent to the Assistant to the Vice President for Student and Special Affairs who had been assigned by the vice president as a coordinator in that research. Then, the coordinators passed the questionnaire links and related documents to all students on their campus who were then able to respond. Researcher had been considering and tried to avoid low response rate by sending reminder or follow up messages, making participation easy such as form using via mobile or laptop format, contacting trusted people to deliver research tool, assuring confidentiality, explaining how the data will be used and how the research benefits to students at TNSU. Moreover, controlling response bias of prior research by analysis sample data indicated that the response sample varied in age, faculty, gender, region, and year class. Although this research was based on secondary data, it was still processed to ensure that it was conducted under ethical concerns. The research ethics code number deriving from research ethics committee at TNSU was TNSU-EDU 001/2566.

**Data Analysis**

This research was interested in studying the development of instruments using CTT and IRT methods to provide information for further scale development. Therefore, this research involved more steps than the instruments development in each method. Seven main research steps were summarized in Table 2.

In Table 2, research analysis procedures were as follows.

1. Basic data analysis. Analyzed 1,048 samples to describe the sample characteristics using frequency and percentage. The data were creating a data log file, recording, preparing a file using statistics including frequency percentages, and all 62 variables were analyzed with means and standard deviation.

**Table 2** Analysis Process, Number of Samples, Statistics, and the Obtained Result

| Analysis Process | Sample Number | Statistics | Obtained Result |
|---|---|---|---|
| 1. Preliminary analysis | Sample with 1,048 students | Frequency, Percentage, Means, Standard deviation | -Characteristics of sample |
| 2. Checking construct validity | random group of 107 students | Developing measurements model with CFA | - Results of the 62 variables scales validit |

NPU JOURNAL

**Table 2** (Continued)

| Analysis Process | Sample Number | Statistics | Obtained Result |
|---|---|---|---|
| 3. Verifying the discriminant power of scales | 15 students with the lowest score rank and 15 students with the highest score rank | 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 |
| 4. Developing of normal criteria for each variable measurement | All sample with 1,048 students | - Developing normal criteria using T-score normalization. | -Criteria for interpreting scores from 62 variable measurement scales |
| 5. Checking the common slope parameters of each item | All sample with 1,048 students | - Analyzing the common slope and Threshold value of each item using Graded-Response Model analyzed with MULTILOG | - Items with reliability and discriminant 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 |

2. Construct validity of the 62 variable scales were validated based on CTT methods. 107 students from 1,048 were randomly selected through SPSS. According to Hair et al. (1998), there is no exactly rule for identifying sample size for SEM analysis. The most common SEM estimation procedure is maximum likelihood estimation (MLE). Hair et al. (2019) suggested the minimum sample size for a measurement model be at least 100 observations, and an oversize of sample is impractical and can cause the measurement model to be overfitted. Hence, this research randomly selected about 100 cases to validate measurement model. This analysis step included checks instrument quality regarding construct validity through Confirmatory Factor Analysis (CFA) with SEM, and uses $\chi^2$ and the index to measure the harmony of the model, adjust model assumptions to be consistent with empirical data, and reports model validated results.

3. Verification of the scales discriminatory power. For this stage, the researchers used the Jung Teh Fan's principle of selecting a sample of 27% to analyze the discriminatory power. 284 random sample 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. A value of Low group and high group was assigned to 30 respondents on each variable (15 people for high/low score groups), then each variable discriminatory power was analyzed by means difference testing through independent t-test.

4. Creation of normal criteria for each variable scales. Scores for each variable were analyzed using the Z test, creating a T-score norm by converting raw scores into T-score, setting normal criteria, and determining how to interpret normal scores. 1,048 sample were utilized as subject analysis through SPSS. The criteria for interpreting scores from 62 variable measurement scales were presented, and used for categorizing students into three groups: students with high, mediocre and low level of each 62 variables.

NPU JOURNAL

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

6. Examining the different item functions. 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 development methods of the two theories. The number of items passed the test quality criteria of each theory were compared using t-test dependent, and analyzed the effect size (d) for paired sample t-test using Cohen'D formula (d=t/$\sqrt{\square}$).

## Results and Discussion

The analysis result showed that most sample were male (68.45%). Regarding academic year, 281 students (26.81%), 194 students (18.51%), 251 students (23.95%), and 322 students (30.73%) were in the fourth year to first year, respectively. In terms of faculty affiliation, 360 students (34.35%) were enrolled in the Faculty of Sports Science, 109 students (10.40%) in the Faculty of Liberal Arts, and 579 students (55.25%) in the Faculty of Education. An analysis of 62 variables found that the mean scores for each variable were at a 'moderate' to 'high' level. All variables were used in developing the measurement tool ($\overline{X}$ = 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. The six objectives analysis results and their discuss were as follows.

1. The results of examining the construct validity of SEL and its causes and effects scales at TNSU indicated that all 62 variables were validated, the concordance index values of all models are not statistically significant. The consistency index values of all models were not statistically significant. The value of $\chi^2$ = 0.01 to 7.98, df = 1 to 4, P-value = 0.12 to 0.99, GFI = 0.98 to 1.00, AGFI = 0.89-1.00, and RMR = 0.00 to 0.08. This is due to the development of measurement tools in the primary research followed the criteria and methods for developing tools by defining variables, verification by experts, and checking for reliability with the Alpha coefficient. This also means that developed tools based on the basic CTT method are valid (e.g., checking for validity, and reliability). 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. The results of examining the discriminatory power of the developed scales showed that the 62 variables scales had discriminatory power. It was found that the mean scores of the group with high scores and the group with low scores of each 62 variables are different. The t values for each item are <1-31.00, df = 14 to 28. Comparing the average values for each variable reveals that the t value is <1 to 424.281, df = 14-28, indicating that all 62 developed measuring instruments have discriminant power. In terms of the discriminatory power, the verification results are at a level that passes the specified criteria, indicating that the scale has a construct validity. These are in lines with Singhasiri et al. (2021) who presented the discriminatory power of the scales deriving from dividing the respondents into high groups and low groups, then finding the value of the t -test. If the means scores are different, it shows that the scales have the power to classify groups. That is, the measurement results can tell that those with high total scores indicate high levels of the trait, while those with low total scores indicate low levels of the trait.

3. The results of creating normalized T criteria for measuring the variables revealed that the normalized T-scores for all variables were able to be classified into differentiate groups. It was found that the scores are in the raw score = 2 to 20 points. The normal T score = 2.25 to 65.24 in measuring the characteristics that are intended to be measured. The low measurement score group, which is a group with a T score < 35, had a T value = 2.25 to 34.98. The moderate measurement score group, which is a group

NPU JOURNAL

with a T score ≥ 35 to 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 ≥ 50, had a T value = 50.03 to 67.23. 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 scores interpretation clear and able to compare individual characteristics. These findings are consistent with Pradujprom et al. (2021) that the normal criterion (Norm) is an important of standardized tests used for interpreting scores, 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. They are in lines with Phatthiyani (2019) who indicated that raw scores cannot provide any information to determine what the measurement results reflect. Therefore, creating T score that will help interpret the results obtained from the measurement will yield information that is consistent with the measurement purpose, and is in a condition that can be useful information.

4. The results of checking the slope parameters of each item, and the Threshold value of each answer item in each variable pointed out that the common slope parameters and the Threshold had $\beta$ values distributed over a parameters range, the Threshold values of each item are $\beta_1<\beta_2<\beta_3<\beta_4$. It showed that the $\beta$ values spread across the range of $\theta$, The common slope parameter of the questions or individual assessment items = 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 $\beta_1$ = -6.17 to - 2.89, $\beta_2$ = -2.99 to -2.13, $\beta_3$ = -1.17 to -0.48, and $\beta_4$ = 0.22 to 0.99. It was also found that every question had the same answer selection curve, that is, people with high $\theta$ values had a probability value of selecting response Items 4 and 5 higher than response Items 1, 2, and 3. It indicates that the scales have the power to discriminate through the criteria set according to the concept of IRT. These results are in lines with Samart et al. (2022) who indicated that the scales can appropriate measure the trait if it was found that the $\beta$ value spreads over the range of the joint slope parameters of the questions, and the Threshold values of each answer item are $\beta_1<\beta_2<\beta_3<\beta_4$.

5. The results of examining the different functions of the questions or individual assessment items on each variable notified that there were items biased toward subgroup. It is found that some items are bias in favor of the reference group and focal group when analyzing the questions different functions for respondents of different genders, faculties, years, and regions. The bias items are presented in Table 3.

**Table 3** Number of Items in Which Test Function Was Detected Using IRT Method

| Classification Variable (Total bias) | List of bias items (Item number of bias items) | | |
|---|---|---|---|
| | Ethical leadership | Cause of ethical leadership Factor | Effect of ethical leadership Factor |
| Gender (11 items) | (3 items) | (7 items) | (1 item) |
| Female | 14 | 80, 81, 82, 85, 89, 98, 143, 151 | - |
| Male | 34, 39 | 157 | 177 |
| Faculty (18 items) | (7 items) | (10 items) | (1 item) |

**Table 3** (Continued)

| Classification Variable (Total bias) | List of bias items (Item number of bias items) | | |
|---|---|---|---|
| | Ethical leadership | Cause of ethical leadership Factor | Effect of ethical leadership Factor |
| 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 (16 items) | (5 items) | (11 items) | (1 item) |
| Higher Year (3st & 4st) | 34, 37 | 69 , 90, 104, 114, 130, 143, 145 | 205 |
| Lower Year (1st & 2nd) | 9,18, 20 | 75, 122, 152, 153 | - |
| Region (7 items) | (4 items) | (2 items) | (1 item) |
| Central | 37, 42, 54 | 155 | 177 |
| Not Central | 23 | 87 | - |

In Table 3, the results of examining the 211 different functions questions showed that the different functions of the exams were classified according to gender, faculty, academic year level, and region. There are 11, 18, 16, and 7 items with gender, faculty, academic year level, and region bias, respectively. The result showed that some scales should be concerned for applying to measure ETS and its causes and effects at TNSU. These results were consistent with Mostert et al. (2024) who suggested that the bias items should be improved or excluded from use in developing tools because it may cause unfairness. 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. The analysis results as an example aligned to objectives 1-5 were shown in Figure 2.

NPU JOURNAL

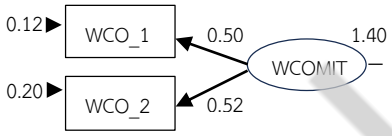
**Variable**: Work commitment or WCOMIT (Commitment to Duties; Accountable Performance; Perseverance

Item 1. You consistently dedicate your body, mind, and time to the assigned tasks.

2. You work with great effort to ensure the successful completion

**Objective Results Based on CTT (Objectives 1 to 3) for Work Commitment Scale Development**

1. Work commitment (WCOMIT) measurement model was valid, Chi-square=1.096, df=1, P=0.295, GFI=0.990, AGFI=0.971, RMR=0.050.




Chi-square=1.10, df=1, P-value=0.29511, RMSEA=0.030

Work Commitment Measurement Model

2. T difference between 15 students with high (H.) commitment and low (L.) commitment were difference ($p < .01$, df=14).

| Items | H.( $\bar{X}$) | H.(S.D.) | L.( $\bar{X}$) | L.(S.D.) | t | df |
|---|---|---|---|---|---|---|
| WCO1 | 5.00 | 0 | 2.93 | 0.52 | 31.00 | 14 |
| WCO2 | 5.00 | 0 | 2.88 | 0.62 | 16.00 | 14 |

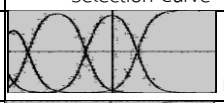
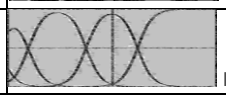
3. T Criteria for work commitment evaluation were able to categorize student in to three groups such as the low, medium and high work commitment group.

| Score | T-score | Percentile | Group |
|---|---|---|---|
| ≤ 5 | ≤ T28.24 | 0.19-1.43 | Low |
| 6-8 | T35.46-T42.68 | 18.32-67.56 | Medium |
| 9-10 | ≥ T57.12 | 81.01-100.00 | High |

The result of WCO scale development showed that two items were able to measure WCO, the measurement model was valid, and WCO T score could be categorized student into three groups.

**Objective Results Based on IRT (Objective 4 to 5) for Work Commitment (WCO) Scale Development**

4. The developed items were valid, the $\beta$ spread across the range of $\theta$, and the Threshold value of each answer item had a value of $\beta_1 < \beta_2 < \beta_3 < \beta_4$

| Items | $\alpha$ (S.E.) | $\beta_1$ (S.E.) | $\beta_2$ (S.E.) | $\beta_3$(S.E.) | $\beta_4$ (S.E.) | Items Response Selection Curve |
|---|---|---|---|---|---|---|
| 1 | 4.65 (0.30) | -3.27 (0.37) | -2.42 (0.13) | -0.76 (0.04) | 0.73 (0.04) |  Item1 |
| 2 | 4.03 (0.25) | -3.17 (0.28) | -2.42 (0.14) | -0.76 (0.04) | 0.65 (0.04) |  Item2 |

5. Item 1 biased towards students from different faculty ($\hat{\beta}^*$=0.09, p<.05). Item 2 biased towards students from different year class level ($\hat{\beta}^*$=0.08, p<.05).

| items | Gender bias | | | Faculty bias | | | Year Class bias | | | Region bias | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\beta}^*$ | S.E. | P | $\hat{\beta}^*$ | S.E. | P | $\hat{\beta}^*$ | S.E. | P | $\hat{\beta}^*$ | S.E. | P |
| 1 | 0.04 | 0.04 | 0.31 | 0.07 | 0.04 | 0.06 | -0.08 | 0.04 | 0.04* | -0.03 | 0.04 | 0.53 |
| 2 | -0.06 | 0.04 | 0.15 | 0.09 | 0.04 | 0.02* | -0.01 | 0.04 | 0.77 | -0.09 | 0.48 | 0.72 |

The result showed that items were valid according to their reliabilities, but each item using should be concerned according to the bias toward students with difference faculty and year class level.

**Figure 2** Example of Scale Analysis Result: Work Commitment Scale Analysis

NPU JOURNAL

6. The results of comparing the number of valid test items analyzed using CTT and IRT methods verified that 27 variables had the same number of items, there were 33 variables with no statistically-significantly of item number difference. The numbers of each item validated by CTT and ITR and their comparative were presented in Table 4.

**Table 4** Number of Items Validated by CTT and ITR, and Items Comparisons

| Components/Variables | | Measurement Theory | | | | Number of Items Comparisons | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CTT | | IRT | | | | | |
| | | No. item | $\bar{X}$ (S.D.) | No. item | $\bar{X}$ (S.D.) | t | df | p | d |
| | Ethical Leadership | | | | | | | | |
| 1. | Diligent | 2 | 1.00 (0) | 2 | 1.00 (0) | Unable to compare | | | |
| 2. | Patient | 3 | 1.00 (0) | 3 | 1.00 (0) | Unable to compare | | | |
| 3. | Responsibility | 3 | 1.00 (0) | 3 | 1.00 (0) | Unable to compare | | | |
| 4. | Work Commitment | 2 | 1.00 (0) | 0 | 1.00 (0) | Unable to compare** | | | |
| 5. | Modesty | 3 | 1.00 (0) | 3 | 1.00 (0) | Unable to compare | | | |
| 6. | Integrity | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 7. | Ethical Persistent | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 8. | Region Adherence | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 9. | Ethical Communication | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 10. | Building Trustworthiness | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 11. | Mercy | 2 | 1.00 (0) | 2 | 1.00 (0) | Unable to compare | | | |
| 12. | Caring | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 13. | Justice | 2 | 1.00 (0) | 1 | 0.67 (.58) | 1.00 | 1 | 0.50 | 1.00 |
| 14. | Modest | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 15. | Respect | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 16. | Good Family Member | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 17. | Good Group Member | 2 | 1.00 (0) | 2 | 1.00 (0) | Unable to compare | | | |
| 18. | Building Unity | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 19. | Regulation Respect | 4 | 1.00 (0) | 2 | 0.50 (.58) | 1.73 | 3 | 0.18 | 0.58 |
| 20. | Social Responsibility | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 21. | Public mind | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| | Cause of Ethical Leadership | | | | | | | | |
| 1. | Ethical Model | 4 | 1.00 (0) | 3 | 0.75 (.50) | 1.00 | 3 | 0.39 | 0.33 |
| 2. | Family Ethical Cultivation | 4 | 1.00 (0) | 3 | 0.75 (.50) | 1.00 | 3 | 0.39 | 0.33 |
| 3. | Societal Ethical Cultivation | 4 | 1.00 (0) | 2 | 0.50 (.58) | 1.73 | 3 | 0.18 | 0.33 |
| 4. | Ethical Experience | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.58 |

NPU JOURNAL

**Table 4** (Continued)

| Components/Variables | | Measurement Theory | | | | Number of Items Comparisons | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CTT | | IRT | | | | | |
| | | No. item | $\bar{X}$ (S.D.) | No. item | $\bar{X}$ (S.D.) | t | df | p | d |
| 5. | Ethical Culture | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 6. | Ethical Developing | 3 | 1.00 (0) | 0 | 0 (0) | Unable to compare* | | | |
| 7. | Ethical Learning | 4 | 1.00 (0) | 2 | 0.50 (.58) | 1.73 | 3 | 0.18 | 0.58 |
| 8. | Ethical Training | 3 | 1.00 (0) | 1 | 0.33 (.58) | 2.00 | 2 | 0.18 | 1.00 |
| 9. | Ethical Literacy | 4 | 1.00 (0) | 3 | 0.75 (.50) | 1.00 | 3 | 0.39 | 0.33 |
| 10. | Ethical Understanding | 4 | 1.00 (0) | 2 | 0.50 (.58) | 1.73 | 3 | 0.18 | 0.58 |
| 11. | Ethical Inquiry | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 12. | Ethical Sensitivity | 4 | 1.00 (0) | 2 | 0.50 (.58) | 1.73 | 3 | 0.18 | 0.58 |
| 13. | Ethical Concept | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 14. | Ethical Concerns | 4 | 1.00 (0) | 3 | 0.75 (.50) | 1.00 | 3 | 0.39 | 0.33 |
| 15. | Ethical Reasoning | 4 | 1.00 (0) | 3 | 0.75 (.50) | 1.00 | 3 | 0.39 | 0.33 |
| 16. | Ethical Decision Making | 4 | 1.00 (0) | 2 | 0.50 (.58) | 1.73 | 3 | 0.18 | 0.58 |
| 17. | Ethical Confident | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 18. | Self-Efficacy | 4 | 1.00 (0) | 3 | 0.75 (.50) | 1.00 | 3 | 0.39 | 0.33 |
| 19. | Self-Confident | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 20. | Optimistic | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 21. | Hope | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 22. | Openness | 3 | 1.00 (0) | 1 | 0.33 (.58) | 2.00 | 2 | 0.18 | 1.00 |
| 23. | Relationship Efficacy | 3 | 1.00 (0) | 3 | 1.00 (0) | Unable to compare | | | |
| 24. | Agreeableness | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 25. | Emotional Flexibility | 4 | 1.00 (0) | 1 | 0.25 (.50) | 3.00 | 3 | 0.06 | 1.00 |
| 26. | Performance Commitment | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 27. | Self-Regulation | 3 | 1.00 (0) | 3 | 1.00 (0) | Unable to compare | | | |
| 28. | Working Standard Setting | 3 | 1.00 (0) | 3 | 1.00 (0) | Unable to compare | | | |
| | Effect of Ethical Leadership | | | | | | | | |
| 1. | Work Advancement | 3 | 1.00 (0) | 3 | 1.00 (0) | Unable to compare | | | |
| 2. | Work Performance | 3 | 1.00 (0) | 2 | 0.67 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 3. | Performance Effectiveness | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |

NPU JOURNAL

**Table 4** (Continued)

| Components/Variables | | Measurement Theory | | | | Number of Items Comparisons | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CTT | | IRT | | | | | |
| | | No. item | $\bar{x}$ (S.D.) | No. item | $\bar{x}$ (S.D.) | t | df | p | d |
| 4. | Group Commitment | 4 | 1.00 (0) | 3 | 0.75 (.50) | 1.00 | 3 | 0.39 | 0.33 |
| 5. | Good Membership | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 6. | Group Member Maintain | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 7. | Leader Trust | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 8. | Leader Role | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 9. | Leadership Effectiveness | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 10. | Voice | 4 | 1.00 (0) | 4 | 1.00 (0) | Unable to compare | | | |
| 11. | Group Justice Perception | 3 | 1.00 (0) | 2 | 0.68 (.58) | 1.00 | 2 | 0.42 | 0.50 |
| 12. | Effective Communication | 3 | 1.00 (0) | 3 | 1.00 (0) | Unable to compare | | | |
| 13. | Group Ethical Climate | 3 | 1.00 (0) | 3 | 1.00 (0) | Unable to compare | | | |

Note: 1. “Unable to compare” means that t value was not able to be analyzed due to the same number of items or the same two means with their zero-standard deviation.

2. * The number of items or the two means with zero-standard deviation is obviously different without the needs of t-test analysis.

In Table 4, The items of 62 variables are ranging from 2-3. The numbers of item passing the CTT criteria are 2-3 whereas the numbers of item passing the IRT criteria are 0-4. The results of comparing the number of items passing the criteria of CTT and IRT methods indicated that some variables are the same number of questions. Even some variables with different number of questions, p value of t indicate no statistical difference ($\bar{x}$ = .25 to .75, S.D. = .50 to .58, t = 1.00 to 2.00, $p > .05$). In detail, there are seven variables in which the mean number of questions are not significantly different. In other words, most of the questions in the variables obtained from the development of the instrument using CTT and IRT has a number of items that are not statistically-significantly different. There are only two variables that are clearly different according to t-test. However, the effect sizes of t paired differences are moderate to high (d were ranging from 0.33 to 1.00), only four variables had high effect sized ($d > .80$). This indicated that these four variables, justice, ethical training, openness, and emotional flexibility should be of more concerned for scales development and application. The scales development with differences results can happen because of many reasons which are in lines with Kanchanawasi (2007) who indicated that 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. 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. However, these two theories are significant for tools development like Coulacoglou & Saklofske (2018) presented that the main aim of studying testing theory is to use it as a source of knowledge for understanding the measurement model, developing tools and items It is based on

NPU JOURNAL

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 decision-making.

## Conclusion

The research aimed at examining the validation and comparison of variables using both CTT and IRT produced several key findings. The results were the following.

1. The results of examining the construct validity of SEL and its causes and effects scales at TNSU indicated that all 62 variables were successfully validated.

2. The results of examining the discriminatory power of the developed scales showed that the 62 variables scales had a discriminatory power.

3. The results of creating normalized T criteria for variables measuring revealed that the normalized T-scores allowed for the classification of sample characteristics into three distinct group scores.

4. The results of checking the slope parameters of each item, and the Threshold value of each answer item pointed out that the joint slope parameters and threshold values of the 211 items revealed a wide range of values, and the threshold values indicating variability in the response patterns across items.

5. The results of examining the different functions of the questions or individual assessment items on each variable notified that those certain items showed bias toward specific subgroups based on the sample background.

6. The results of comparing the number of valid test items analyzed using CTT and IRT methods verified that 27 out of 62 variables showed no significant difference in the number of items, while 33 variables showed no statistically significant difference in the average number of scales, with only two variables displaying a distinct difference. According to the effect size value, only four variables, as justice, ethical training, openness, and emotional flexibility had high mean difference (d=1.00). These results suggest that both CTT and IRT provide valuable similar results for scales development; IRT offers more detailed analysis of item properties, while CTT provides straightforward assessment of overall scale performance.

### Contribution

This research contributes to the understanding of how both Classical Test Theory (CTT) and Item Response Theory (IRT) can be applied to validate and compare measurement scales. The study provides evidence of the reliability and discriminatory power of scales across 62 validated variables, offering insights into the effective classification of sample characteristics using normalized T-scores. By analyzing the joint slope parameters and threshold values of 211 items, the research uncovers the variability in response patterns and identifies potential biases in item responses based on the sample background. The comparison of CTT and IRT methods highlights the strengths of each approach, revealing that while both theories yield comparable results in terms of the number of items and scales, IRT offers a more in-depth analysis of individual item characteristics. This contribution enhances the understanding of scale development and validation, emphasizing the utility of both CTT and IRT in different contexts and offering a foundation for future research on improving measurement accuracy in educational assessments.

### Suggestions

1. Suggestions for improving educational policy. A purpose of this research was to develop scales for evaluating the student characteristics in order to improve personnel development. The developed scales from this study were able to be applied to measuring SEL and its causes and effects. TNSU administrators

NPU JOURNAL

and staff might use the measurement results as an evaluation database, and set the program for developing SEL based on analysis results. They also can be used to develop students in terms of the success of student development. In addition, the results in the scales development 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 tools application and their 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 using secondary data. It does not focus on the information obtained from the measurement. According to prior research, there were low response rates and different percentages for each campus though online response. These issues should be of concern for further data collection. 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. The tools can be developed as 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. It is possible to study the developed model 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, year class, faculty, and region of the campus using multiple group strategy. The further research is able to use the concept of CTT or/and IRT for scale development, and examined the scale development result to approve them, as well as to provide suggestions for developing tools to gain accurate research results. According to IRT, some items with bias such as justice, ethical training, and ethical development causes different outcomes between CTT and IRT. The causes of these bias items should be studied. Moreover, the scale development based on IRT should be recommended for the item development for more precise student measurement.

**Limitations**

First, the research focused on a secondary specific sample, which may limit the generalizability of the findings to other populations. Hence, ensuring the data is up-to-date, it is appropriate to collect new research data using a newly developed instrument, which has been revised based on the results of item-level analysis using IRT. Second, it is based on secondary data which utilized random sampling via online collection which is hard to control the response rate. This may cause result bias, the further study might minimize them and suggests proper ways for online, on site, mail, or other data collection methods for conducting research. Third, the use of variables and items may not fully represent the complexity and diversity of all possible measurement scales. Research with a broader range of variables and items could offer a more comprehensive understanding for SEL and its causes and effects. Fourth, the study only examined one aspect of measurement theory, and future research could explore other models or incorporate longitudinal data to assess the stability and validity of the scales over time. Fifth, the analysis may have been influenced by specific software (e.g., Multilog and SIBTEST), which could introduce limitations related to the software's algorithmic assumptions or the result interpretation based on the chosen models. Lastly, this research is not aimed to explain the cause of item bias. Further research might aim to identify

the causes of item bias, and replace, remove, or revise the bias item, and check it through interview of experts, lecturers or students to provide the information for item bias elimination.

**References**

Andrich, D. & Marais, I. (2019). *A Course in Rasch Measurement Theory: Measuring in the Education, Social and Health Science*. Springer Nature Singapore.

Demars, C. E. (2018). Classical Test Theory and Item Response Theory in Irwing, P., Booth, T. & Hughes, D. J. (Eds.), *The Wiley handbook of psychometric testing* (pp. 48-73). John Wiley & Sons.

Cardinal, B. J. (2023). Sexual Misconduct and Violence in Physical Activity and Sport Settings. *JOPERD: The Journal of Physical Education, Recreation & Dance*, *94*(5), 64. https://doi-org.ejournal.mahidol.ac.th/10.1080/07303084.2023.2185000

Coulacoglou, C. & Saklofske, D. H. (2018). *Psychometrics and Psychological Assessment*. London: Academic Press.

Hair, J. T., Black, W. C., Babin, B. J. & Anderson, R. E. (1998). *Multivariate Data Analysis* (5th ed.). Prentice-Hall.

Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Hampshire, Cengage Learning EMEA.

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. https://www.researchgate.net/publication/379155280_Role_of_Ethical_Leadership_in_Corporate_Governance_A_Systematic_Review

Irwing, P. & Hughes, D. J. (2018). Test development in Irwing, P., Booth, T. & Hughes, D. J. (Eds.), *The Wiley handbook of psychometric testing* (pp. 1-47). John Wiley & Sons. https://doi-org/10.1002/9781118489772

Kanchanawasi, S. (2007). *New Testing Theory* (3rd ed). Faculty of Education, Chulalongkorn University.

Mostert, K., de Beer, L. & de Beer, R. (2024). Invariance and item bias of the Mental Health Continuum Short-Form for South African university first-year students. *African Journal of Psychological Assessment*, *6*(1), 1-9. https://doi.org/10.4102/ajopa.v6i0.143

Opstoel, K., Chapelle, L., Prins, F. J., De Meester, A., Haerens, L., van Tartwijk, J. & De Martelaer, K. (2020). Personal and social development in physical education and sports: A review study. *European Physical Education Review*, *26*(4), 797-813. https://doi-org.ejournal.mahidol.ac.th/10.1177/1356336X19882054

Panphet, K. (2022). Ethical leadership based on Vuddhidhammas for school administrators in the 4.0 era. *Journal of Education Review*, *9*(1), 491-498. https://so02.tci-thaijo.org/index.php/EDMCU/article/view/255546

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

Phramaha Samack Atibhaddho, Kamsuk, K. & Suwanjiraphaisanm W. (2025). Budhist principles and political leadership development in the Thai context. *Journal of Buddhist Innovation and Management, 8*(1), 256-266. https://so06.tci-thaijo.org/index.php/bim/issue/view/18189/6065

Pradujprom, P., Pantong, K. & Kitiyanusan. R. (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. https://so06.tci-thaijo.org/index.php/rdibsru/article/view/254203

Samart, C., Intanam, N. & Rattanaraj, P. (2022). Development of a social skill test for the 9th grade students in the schools under Ubon Ratchatani Primary Educational Service Area by using Polytomous Item Response Theory. *Journal of Graduate School, Pitchayatat, Ubon Ratchathani Rajabhat University*, *17*(1), 77-88. https://so02.tci-thaijo.org/index.php/Pitchayatat/article/view/252009

NPU JOURNAL

Sen, S. & Gocen, A. (2021). A psychometric evaluation of the Ethical Leadership Scale using Rasch analysis and confirmatory factor analysis. *Journal of General Psychology*, *148*(1),84-104. https://doi-org.ejournal.mahidol.ac.th/10.1080/00221309.2020.1834346

Shakeel, F., Kruyen, P. M. & Van Thiel, S. (2024). Ethical leadership in the Netherlands: testing the broader conceptualization and measurement scale. *International Journal of Public Leadership*, *20*(2), 144-167. https://doi.org/10.1108/IJPL-10-2023-0082

Singhasiri, P., Sirisatayawong, P. & Chinchai, P. (2021). Known-group validity and inter-rater reliability of the Dynamic Loewenstein Occupational Therapy Cognitive Assessment (DLOTCA). *ASEAN Journal of Rehabilitation Medicine*, *31*(3), 111-118. https://he01.tci-thaijo.org/index.php/aseanjrm/article/view/249244

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

Toraman, Ç. & Korkmaz, G. (2023). *What is the "Meaning of School" to High School Students? A Scale Development and Implementation Study Based on IRT and CTT.* SAGE Open, *13*(3), 1-15. https://doi.org/10.1177/215824402311990

Transparency International Secretariat. (2024). *Corruption Perception Index 2023*. Retrieved June 2025, from https://images.transparencycdn.org/images/CPI-2023-Report.pdf

Trivedi, C. (2020). *Assessing reliability in research methods.* Retrieved January 2025, from https://conceptshacked.com/assessing-reliability/

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

NPU JOURNAL