Validating Teamology in Domestic and International Settings

Yang Hua

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Mechanical Engineering

Jan Helge Bøhn, Chair
Clinton L. Dancey
Christopher B. Williams

September 8, 2015
Blacksburg, VA

Keywords: Teamology, Test-retest Reliability, Systematic Change over Time for Psychological Traits

Validating Teamology in Domestic and International Settings

Yang Hua

ABSTRACT

In recent years, collaboration between different companies especially global collaboration on oversea product development becomes more and more popular. Forming efficient product design team becomes an important concern for these companies. Team formation strategies not only consider team member's skills and availability, but also gender, race and cultural background. Personality traits are also increasingly considered when composing a team, based on the hypothesis that diversity in personality traits within a team will improve the team's ability to innovate (Park, 2014, Figure 6-3). Wilde released his 20-item psychological preference test together with his Teamology teaming strategy in 2008, with the assumption that its resulting reliability would be approximately 80% over time due to their similarity to the Myers-Briggs Type Indicator (MBTI) questions (Kirby et al, 2007). In this thesis, the overall test-retest reliability of Teamology instrument is proved good since consistency over time for all four Dimensions are higher than 80%. For each of the 20 items, some are considered not reliable with low consistency over time. Systematic change for consistency data over time is discussed as well, a tendency is figured out that for Dimension EI and SN, graduate participants tend to change their preference on dimension EI and SN over time, while no obvious change is shown for Dimension JP and TF. When the culture and language difference is concerned, all four dimensions have good consistency over time, which means language and culture difference will not affect the consistency of Teamology test score. Finally for Park Creativity Index and MBTI Creativity Index, the reliability over time is tested and judged acceptable with Pearson's correlation data of 0.528 and 0.516.

Acknowledgement

I would like to convey my sincerest thanks to my advisor, Dr. Jan Helge Bøhn, for his support, guidance, and patience during my graduate studies.

I also would like to thank my committee members, Dr. Clinton Dancey and Dr. Christopher Williams, for their generous support and feedback during my research.

This research has also received assistance from LISA (Laboratory for Interdisciplinary Statistical Analysis) at Virginia Tech and the Virginia Tech Institutional Review Board, I would like to give them my most sincere thanks and appreciation either.

Finally, I would also like to convey my sincerest thanks to my parents, Jian Hua and Liwei Yang, for their 24-year upbringing, and their fully support during my master's period of time.

Table of Contents

Chapter 1

Preface

1.1 Introduction

Engineering teams have existed in industry for decades. Increasingly, globally collaborating teams have become the norm for product development for global market. A highly efficient team requires not only capable individuals but also good cooperation between the team members. Thus in recent years, more organizations, companies, and universities take part in the research of team forming strategy. Forming a team by individual’s psychological preference is one of the methods developed. Numerous psychological teaming strategies have been established since then. Typical examples for psychological teaming strategy in engineering design education field are Teamology (Wilde, 2008) and Six Thinking Hat problem-solving strategy (DeBono, 1985). Most teaming strategies are based on the principle of Jungian cognition theory (Jung, 1923) about defining different ways of perceiving and judging (Wilde, 2008, p. 1).

Before applying the psychological teaming strategy, individual’s psychological preference should be tested. One famous and reputable instrument is Myers-Briggs Type Indicator (MBTI) created by Myers and Briggs in 1962, which describe a person’s psychological type by a combination of four letters (Myers, 1962). An abbreviated psychological preference test based on MBTI was created by Wilde in 2008, which provide individual’s numerical psychological preference result for team forming. After testing the psychological preference of individuals, a typical step for teaming forming is to seed each team with the individuals with highest possible creativity score first, then fill the rest participants into teams by maximizing personality type diversity based on the hypothesis that diversity in personality traits within a team will improve the team’s ability of innovation (Park, 2014, Figure 6-3).

To achieve a high validity, psychological teaming strategies require the reliability of the psychological measurement instrument. Only when the psychological preference of individuals is measured accurately, excellent engineering design team can be formed by psychological teaming strategies. Repeatability of the test result is judged as part of the reliability of psychological measurement (Schaubhut, et al., 2009, p. 4), if the repeatability, or called test-retest reliability is high, the result of measurement instrument is determined believable and reputable. For the most updated MBTI form M, the test-retest reliability has been determined acceptable with Pearson's correlation data from 0.67 to 0.73 (Schaubhut, et al., 2009, p. 7), while for Teamology instrument, there is no publication verifying its test-retest reliability so far.

As a prerequisite for testing the repeatability, psychological type of individual is believed remain stable over life span (Quenk, 2009, p. 32). But researches indicate that it is not always the case, the psychological preference for individuals can have a systematic change over time because of his or her specific experience such as team-based learning during this time interval (Wilde, 2008, p. 54). Current research on MBTI does not show any evidence that systematic change on personal type over time do exists. Since Wilde's Teamology instrument provides a more numerical psychological data, psychological preference change over time can be determined if it does exists. If proved that no systematic change over time exists, individuals' psychological preference results can be used repeatedly during the life span and it is not necessary to do the same test any more. On the other hand, if systematic change does exist, the change rate can be calculated, and period of validity of psychological tests can be expected.

For the creativity score used for seeding the team, psychological preference instrument cannot measure creativity directly. Creativity index is one approach measuring one's creativity based on the result of psychological test result. MBTI Creativity Index is one example indicating the creativity of general male (Gough, 1981). For engineering student, Park Creativity Index is developed by Yong-Seok Park in 2014 (Park, 2014, p. 104).

1.2 Problem Statement

Team formation strategies typically consider team member skills and availability as appropriate for the team's task at hand. These skills may include technical skills, proficiencies, experiences, and insights; and gender, race, and cultural backgrounds may be sought as proxies for experiences and insights that might impact the capabilities of the team as a whole. Personality traits are also increasingly considered when composing a team, based on the hypothesis that diversity in personality traits within a team will improve the team's ability to innovate (Park, 2014, Figure 6-3).

Myers-Briggs Type Indicator (MBTI) is the long-established benchmark for identifying personality traits. Research has shown that the current MBTI version, the 93-question Form M, has a test-retest reliability in the range of 0.67 to 0.73 (Schaubhut, et al., 2009, p. 7). Wilde developed a more condensed 20-question instrument, referred to as Teamology, based on the

MBTI questions, with the assumption that its resulting reliability would be approximately 0.8 due to their similarity to the original MBTI questions (Kirby et al, 2007). However, the reliability of the Teamology instrument has not yet been measured. The objective of this thesis is therefore to measure the test-retest reliability of the Teamology instrument.

1.2.1 Hypothesis

The test-retest reliability of the Teamology instrument is 0.8 over time.

1.2.2 Research Questions

RQ1. Is the test-retest reliability consistent across all 20 Teamology questions? This question is answered in Section 3.2.

RQ2. Is there a systematic change in reliability over time? This question is answered in Section 3.3.

RQ3. Do native languages and/or cultural backgrounds impact the reliability? This question is answered in Section 3.4.

RQ4. Does the reliability substantially affect the MBTI Creativity Index and/or the Park Creativity Index? This question is answered in Section 3.5.

1.3 Solution Outline

The purpose of this thesis is to determine the test-retest reliability of Teamology instrument. The objective is to compare the participants' Teamology test score with their previous test score recorded several months ago, measure the test-retest correlation and predict the systematic change of personal type over time. The impact of language and culture on stability and the reliability of creativity measurement tool (Park Creativity Index) are tested as well.

Teamology instrument created by Wilde (Wilde, 2008) has a close relationship with another famous psychological type measurement tool: The Myers-Briggs Type Indicator (MBTI). Test-retest reliability is analyzed and compared by calculate the absolute average difference and consistency percentage between old and new scores. The difference in time period between two tests and nationality difference lead to the analysis on systematic change over time and impact caused by language and culture difference. Finally, Pearson's correlation coefficient is applied on measuring the test-retest reliability of Park Creativity Index.

Two main groups of participants are used in this thesis research:

ME 2024: 56 US sophomore mechanical engineering students from 5 previous semesters who took Teamology test before. The time interval between their two tests ranges from 10 months to 34 months.

PACE Global Course: 75 senior class or graduate level engineering students from 3 previous semesters who took Teamology test before. The time interval between their two tests ranges from 10 months to 34 months. The nationality of these participants includes China, German, Mexico and USA.

1.4 Thesis Outline

The outline of the thesis paper is as follows:

Chapter 1 provides a brief introduction to the project, the problem statement, research questions, hypothesis, solution overview and the thesis paper outline.

Chapter 2 includes the related literature review on Myers-Briggs Type Indicator (MBTI), Teamology instrument as well as the test-retest reliability of these psychological preference type measurement tools.

Chapter 3 is the main part of the thesis paper. The test-retest reliability of Teamology instrument is measured. The analysis includes not only 4 dimensions but also all 20 items used in Teamology test to prove if the test-retest reliability is higher than 0.80. Five different time intervals are applied to figure out if systematic change on personality type over time do exists or not. Test-retest reliability for different age groups and different nationality are discussed in this chapter as well. The repeatability of Park Creativity Index over time is measured as well. Data set used: ME 2024 and Global Course

Chapter 4 presents the result, discussion and future plan. All 4 Dimension and 20 items are discussed and compared to see if any item has significant low test-retest reliability.

Chapter 2

Literature Review

In this chapter, previous work related to The Myers-Briggs Type Indicator (MBTI), Teamology score, and Test-retest Reliability for psychological type instruments are reviewed. For the Myers-Briggs Type Indicator (MBTI), the literature review focuses on its internal consistency, test-retest reliability and impact caused by culture, language and age of test-taker. For Teamology, the focus is the correlation between it and Myers-Briggs Type Indicator (MBTI). At the same time, Test-retest reliability measurement method is reviewed for the data analysis in this thesis project. By reviewing these previous works, a brief idea of psychological preference test can be generated. In this chapter, the MBTI will be discussed in Section 2.1, Teamology will be discussed in Section 2.2 and the Test-retest reliability is in Section 3.3.

2.1 The Myers-Briggs Type Indicator (MBTI)

The Myers-Briggs Type Indicator (MBTI) assessment is one of the most widely used personal psychological preference measurement tools (Schaubhut et al., 2009, p. 4). It was initially designed by Katharine Briggs and Isabel Myers and firstly published in 1962 (Myers, 1962), based on Carl Jung's research on psychological types (Jung, 1923). The MBTI Manual has since been updated and optimized by Mary McCaulley (1985), and by Naomi Quenk and Hammer Allen (1998), to increase its reliability and reduce testing time. The most recent MBTI standard form for Step I assessment is Form M created by Quenk and Allen in 1998 (Myers, et al, 1998), which contains 93 items and usually takes more than 25 minutes for participant to complete (Quenk, 2009, p. 92).

The Myers-Briggs Type Indicator (MBTI) describes personality type by a combination of four alphabetic characters, such as ESFJ, INTP, etc. Each character represents a dimension of personal type using two opposite words: Extraversion (E) vs. Introversion (I); Sensing (S) vs. iNtuition (N); Thinking (T) vs. Feeling (F); and Judging (J) vs. Perceiving (P) (Myers, 1962). The detail description of four dimensions can be seen in Table 2.1 and in the official MBTI Manual (Myers, et al, 1998).

Table 2.1 Four Dimensions of the MBTI

Extraversion (E) vs. Introversion (I): Attitudes or Orientations of Energy	
Extraversion (E): Directing more energy to outer world (other people and objects)	Introversion (I): Directing more energy to inner world (own experience and ideas)
Sensing (S) vs. iNtuition (N): Function and Processes of Perception	
Sensing (S): Tend to focus more on his/her own five senses	Intuition (N): Tend to focus more on patterns and interrelationships of objects
Thinking (T) vs. Feeling (F): Function and Processes of Judgment	
Thinking (T): Making conclusion based on logical analysis	Feeling (F): Making conclusion based on social and personal value
Judging (J) vs. Perceiving (P): Attitude or Orientation to outer world	
Judging (J): Prefer decisiveness and closure, dealing outer world using T or F dimension	Perceiving (P): Prefer flexibility and spontaneity, dealing outer world using S or N dimension

The four dichotomies created by Myers and Briggs have slightly difference from Jung's psychological type theory (Jung, 1923). The most notable difference is the additional Judging (J) vs. Perceiving (P) dimension, which represent person's preferred extraverted function. The four dimensions lead to altogether 16 personality types, which can be seen in Table 2.2 below, together with the estimate percentage of 16 types in U.S. population. (Center for Applications of Psychological Type (CAPT), 2010)

Table 2.2 16 Personality Types and its estimate percentage in US population

ISTJ 11–14%	ISFJ 9–14%	INFJ 1–3%	INTJ 2–4%
ISTP 4–6%	ISFP 5–9%	INFP 4–5%	INTP 3–5%
ESTP 4–5%	ESFP 4–9%	ENFP 6–8%	ENTP 2–5%
ESTJ 8–12%	ESFJ 9–13%	ENFJ 2–5%	ENTJ 2–5%

Although Myers-Briggs Type Indicator (MBTI) is popular and widely used in the business and education fields, it faces several severe criticisms from academics, including its statistic validity (Pittenger, 1993, p. 3), correlation with other psychological instruments, and its reliability. For the statistic validity, Pittenger claims that a bimodal distribution should be expected from the result of MBTI since it is a typology (Pittenger, 1993, p. 3). However, like many other psychological preference measurement tools, the result of MBTI shows a normal distribution (Stricker & Ross. 1964), so it is possible that two individuals with similar scores will be divided into two totally opposite types.

MBTI is also being criticized for its low reliability. Research shows that psychological type will not shift measurably during an individual's life time (Tieger, Barron-Tieger, 1993, p. 2), which means the 4-letter MBTI type will not change in a short period of time. However, Pittenger and some other researcher show that only half of the individuals keep their type after a period of time, ranging from 5 weeks to 9 months (Pittenger, 1993, p. 4) (Harvey, 1996, p. 5- 29), which indicate a low test-retest reliability.

Other than the criticisms mentioned in the previous paragraph, there are some problems which cannot be neglected when MBTI is applied. These include culture difference, native language, age and English reading level of the test taker. These problems are highly related to the terminology of MBTI. The terminology of MBTI is criticized as vague and general, while the accuracy of MBTI requires honest self-reporting by the test taker (Myers & McCaulley, 1985). The reliability of the test result will decrease if the test takers cannot understand and answer the questionnaire correctly and clearly.

The culture and language might affect the response as well. Although MBTI has been officially translated to 20 languages currently, American English version is still the most reliable version because it has been examined and validated repeatedly. Church (2001) believes that personality traits varies between cultures, thus personality test can only be applied in the area where it is constructed. However, more recent research indicates that differences are mostly caused by translation difficulties rather than culture difference (Beuke, et al, 2006, p. 8). Participants that have English as their second language, often have a hard time fully understanding the questionnaire since the MBTI tools expect an eight-grade reading level (Quenk, 2009, p.35). Translations can also be affected by subtle differences in the meaning of translated words and phrases (Hofstede & Hofstede, 2005).

2.2 Teamology

Teamology was published by Douglass Wilde in 2008. The Teamology questions were derived from the MBTI questions based on his 16 years study on engineering student design project teams (Wilde, 2008, p. vii). The questionnaire consists of 20 A-B-Both-Neither questions that can be easily finished in 5 minutes. The benefits expected from the smaller questionnaire size and simpler expression than 93-item MBTI test is that Teamology is more likely to have a high attendance rate and honest response because of its simplicity.

Wilde's questionnaire is highly related to Jung's cognition theory and MBTI (Wilde, 2008, p. 18). He utilized Myers' four dichotomies used in MBTI: EI, JP, SN, and TF (Myers, 1962). The selection of 20 items falls into 4 dimensions evenly with 5 items per dichotomy since it can represent 80% consistency over time (Kirby et al, 2007). One of the most important improvements Wilde did is to transform qualitative psychological preference type into quantitative scores, which makes it easier to apply quantitative analysis by psychological preference like product design team forming. The 20 items questionnaire is shown below in Table 2.3 (Wilde, 2008, p. 10).

The Teamology questions are derived from the 93 MBTI items. The selection and refining of 20 items are based on statistical studies of correlations between all 93 MBTI questions. According to Quenk and Hammer (Myers, et al, 1998), statistic study on MBTI shows the questions in each of the four dimensions can be concluded into five groups called "facets". The results of all other items in the same facets can be predicted if the response of one item has been determined. Thus 20 Teamology items are generated from 93 MBTI items to represent all 20 MBTI facets in four dimensions (Wilde, 2008, p. 18).

In order to avoid the vague and general phrasing in MBTI, Wilde tried to use simple and clear expression for his 20 items (Wilde, 2008, p. 18). A brief example based on question EI1 gives a brief view of this procedure. The original MBTI questions are shown below:

> "Initiating: People at this pole get pleasure from mingling with others in large or small gatherings. . . . "
>
> "Receiving: People at this pole are much more comfortable letting conversations come to them than initiating contact. . . . "

These MBTI questions represent the “Initiating-Receiving” facet in the EI dimension. The corresponding Teamology question is rephrased as “You are more sociable, or more reserved?” All 20 Teamology questions are established under similar way to represent 5 facets for each of the 4 MBTI dimensions.

Table 2.3 Teamology Questionnaire

Energy Direction: Outward or Inward		
EI1 You are more:	(e) sociable	(i) reserved
EI2 You are more:	(e) expressive	(i) contained
EI3 You prefer:	(e) groups	(i) individuals
EI4 You learn better by:	(e) listening	(i) reading
EI5 You are more:	(e) talkative	(i) quiet
EI difference: Σ e - Σ i = EI____		
Orientation: Structured or Flexible		
JP1 You are more:	(j) systematic	(p) casual
JP2 You prefer activities:	(j) planned	(p) open-ended
JP3 You work better:	(j) with pressure	(p) without pressure
JP4 You prefer:	(j) routine	(p) variety
JP5 You are more:	(j) methodical	(p) improvisational
JP difference: Σ j - Σ p = JP____		
Information Collection process: Fact or Possibilities		
SN1 You prefer the:	(s) concrete	(n) abstract
SN2 You prefer:	(s) fact-finding	(n) speculating
SN3 You are more:	(s) practical	(n) conceptual
SN4 You are more:	(s) hands-on	(n) theoretical
SN5 You prefer the:	(s) traditional	(n) novel
SN difference: Σ s - Σ n = SN____		
Decision-Making process: Objects or People		
TF1 You prefer:	(t) logic	(f) empathy
TF2 You are more:	(t) truthful	(f) tactful
TF3 You see yourself as more:	(t) questioning	(f) accommodating
TF4 You are more:	(t) skeptical	(f) tolerant
TF5 You think judges should be:	(t) impartial	(f) merciful
TF difference: Σ t - Σ f = TF____		

Based on the structure of Teamology questionnaire, the score range for each of the four dimension is [-5, 5]. As far as the score for all four dimensions (EI, JP, SN and TF) are determined, it can be simply transferred into eight cognitive modes (ES, EN, IS, IN for collecting information and ET, EF, IT, IF for making decision) (Wilde, 2008, p. 11). The numerical transform method can be found in Table 2.4.

Table 2.4 Cognitive modes Transform Table

Collection mode [C-mode]	**Decision making mode [D-mode]**
ES = EI – JP + 2SN	ET = EI + JP + 2TF
EN = EI – JP - 2SN	EF = EI + JP - 2TF
IS = - EN	IT = - EF
IN = - ES	IF = - ET

2.3 Test-retest Reliability and Long Term Drift

Test-retest reliability, also called repeatability, is defined as the variation of measurement when the same instrument is applied on the same item in a short period of time. If the variation is smaller than an acceptance criteria determined in advance, the measurement can be determined as “repeatable.” Only when the result is repeatable can a measurement tool be judged valid and reliable. Test-retest reliability is not only an important requirement for physical measurement tool such as ruler, it can also be applied on psychological measurement tool like Myers-Briggs Type Indicator (MBTI) mentioned in Section 2.1.

Generally speaking, in order to minimize the error caused by other factors, a list of conditions should be fulfilled before the repeatability is tested. These conditions include same observer, same measurement tool under same condition, same location and same object within a short time period of the previous measurement (Taylor, 2009, p. 14). However, when the repeatability of psychological measurement tool is concerned, some of these conditions can hardly be met. The retest result of a psychological test might be significantly different from the first one due to three possible reasons (Davidshofer & Murphy, 2013, p.123):

1. Object’s attribute changes between two tests;
2. Effect of taking the first test;
3. Carryover Effect if the interval between two tests is short.

The object's attribute changes between two tests means participants might change their preference because of the experience between two tests, for example a language test is given to a kid when he is 4 years old and retest when he is 5. His language skill definitely increases during this period of time, thus a change of test result can be expected. Effect of taking the first test means that the experience of taking the first will affect the answer when retest, for example anxiety inventory test could increase a person's level of anxiety. Carryover Effect is simply that participants might still remember their reply when they firstly took the test. For Myers-Briggs Type Indicator (MBTI) and Teamology test discussed in this thesis paper, the first and the third condition could occur. Thus to test the repeatability of psychological measurement tool, time period between the first and the second test should not be too short to avoid the Carryover Effect, while at the same time drift of result can be expected due to test taker's attribute change during this period of time.

Pearson's Correlation Coefficient is a measure of the linear correlation between two variables (Karl Pearson, 1880). It is widely used for testing the degree of linear dependence between two groups of data. The range of correlation coefficient is [-1, +1], with +1 means total positive correlation and -1 represent total negative correlation. When social science like psychology is concerned, it is highly likely that the correlation coefficient falls into the range from .00 to .60, with most of them lower than 0.30 (Guilford, 1965, p. 146). For test-retest reliability of psychological test, the result is higher with acceptable range from .50 to .60. A correlation coefficient for repeatability of psychological test is judged high with correlation coefficient higher than .60 (Cohen, 2013, p. 78). In the following paragraph, the correlation coefficients mentioned are all Pearson's coefficient.

As mentioned in Section 2.1, test-retest reliability of MBTI has been questioned for several times. As the most widely used psychological type measurement tool around the world, the test-retest reliability for MBTI has been verified by several academic publications. One of the earliest tests was done by Stricker and Ross in 1964, a 14-months test-retest interval was applied and lead to a test-retest correlation coefficient of 0.69 to 0.73 for all dimensions, despite the Thinking-Feeling of 0.48 (Stricker & Ross, 1964). A later test managed by Levy, Murphy and Carlson in 1972 used a 8-week interval and present a test-retest correlation coefficient of 0.69 to 0.80 for males and 0.78 to 0.83 for females (Levy, et al, 1972). Carskadon held a test in 1977 using a 7-week interval and lead to a test-retest correlation coefficient result of 0.73 to 0.87, however there is an exception of 0.56 for male's Thinking & Feeling Dimension (Carskadon, 1977).

The previous three researches are all related to the first edition of MBTI published in 1962. For the most updated Form M is concerned, Quenk provides a group of test-retest reliability correlation data in the MBTI Manual as well as its supplement document. Quenk claims that for a 4-week interval test, 66% of the participants reported all four letters the same and 91% reported 3 or more (Quenk, 2008, p. 81). A more systematic test was held by Schaubhut and Herk. About 400 pairs of data were collected, which include half male and half female. The time interval varies from less than one week to more than 4 years. The overall correlation data ranges from 0.57 to 0.81. If gender is concerned, male's data range from 0.53 to 0.93 and female range from 0.56 to 0.92. The overall test-retest correlation coefficient is judged ranging from 0.67 to 0.73 (Schaubhut, et al., 2009, p. 7). Although correlation data are different under each time intervals, there is no evidence shows that longer time lead to lower correlation.

For Teamology instrument, currently there is no academic publication available measuring the test-retest reliability of it. As mentioned by Wilde, the Teamology Test is highly related to the MBTI measurement by covering all its 20 facets (Wilde, 2008, p. 18), thus by design Teamology instrument and MBTI are intended to share the same validity including the test-retest correlation, which means the type should not shift during lifetime, while the Teamology score have a 80% reliability over time (Kirby et al, 2007). However, Wilde also suggested that specific experience such as team-based learning might develop individual participants (Wilde, 2008, p. 54). This means it is possible that one's psychological preference score will drift after a period of time.

2.4 Brief Conclusion for Literature Review

This chapter has reviewed the literature in the following areas:

The Myers-Briggs Type Indicator (MBTI)(Section 2.1). It includes its origin, improvement, 4 dimensions and totally 16 psychological types, as well as the criticism MBTI faces.

The Wilde's Teamology instrument (Section 2.2). The origin of Teamology survey, method of question generation and its correlation with MBTI psychological types is covered.

The test-retest reliability (repeatability) and long term drift of psychological survey (Section 2.3). Include the previous work on test-retest repeatability of MBTI and Wilde's Teamology Instrument, the statistic tool for testing the repeatability as well as some common knowledge for psychological survey repeatability test.

Based on the literature review, some observation and brief conclusions are listed below:

Due to the Carryover Effect, test-retest reliability of psychological measurement tool examined in a short period of time might not represent the actual reliability of it because participants may still remember his or her previous response and it will cause error to the result generated. However, if the time interval is long enough to eliminate the Carryover Effect, the retest result might face a drift compared with the first response because of the test taker's experience during this period of time and attribute change.

The previous works on test-retest reliability of Myers-Briggs Type Indicator (MBTI) provide a clear idea how MBTI results change over time. As a combination of different test source, the test-retest reliability for the most updated MBTI manual is 0.67 to 0.73, which is good enough for a psychological preference test.

For Wilde's Teamology instrument, currently there is no related article discussing its test-retest reliability. However, Wilde assumed that it should share the same validity with MBTI because the Teamology items are well generated from all 93 MBTI items. Since there are only 5 items in each dimension to represent 5 MBTI facets, the assumed test-retest reliability is judged as 80%, which means on average only 1 out of 5 items might face difference between two tests. Wilde also claims that specific experience can develop individual participants so as to change the psychological preference.

Chapter 3

Test-retest Reliability and Systematic change over time for Teamology Instrument

The Teamology teaming strategy and the 20-items psychological test were first published in 2008 (Wilde, 2008). Wilde believes that his Teamology instrument is, by design, highly related to the Myers-Briggs Type Indicator (MBTI), thus they should share the same validity (Wilde, 2008, p. 20). Currently there is no publication testing the validity of the Teamology instrument, thus in order to certify its validity, one recognized way is to measure its test-retest reliability to determine if the result from participants are consistent over time.

The aim of this chapter is to measure the test-retest reliability of the Teamology instrument, investigate any systematic change in participant responses over time, and determine the impact by culture or language on participant responses to the Teamology instrument. The test-retest reliability helps represent the validity of the Teamology instrument, while the systematic change over time leads to the measurement of the valid time period for psychological preference tests. The impact on consistency by culture and language difference is tested as well to ensure that the Teamology instrument can be used by participants whose first language is not English because English version is the first and most reliable edition for Teamology instrument currently.

In this chapter of the thesis, Section 3.1 covers the experiment design and target research objects used in the succeeding analysis. Section 3.2 measures the test-retest reliability of Teamology test, which includes the performance of all 20 items as well as 4 dimensions. Systematic change on reliability and psychological preference shift is discussed in Section 3.4, including the performance of each dimension as well as discussion on shift tendency. In Section 3.5, participants' data from different countries (China, Germany, Mexico and U.S.) are compared to verify if the culture and language have an impact on reliability. Finally, in Section 3.6, the creativity index, Park CI (Park, 2014, p. 104) is calculated and compared for both data entries to verify its validity and definition. The suggestions to improvement Teamology questionnaire and Park CI is available at the end of this chapter as well.

3.1 Experiment Design

In order to verify the test-retest reliability of the Teamology instrument, an experiment is needed to collect and analyze the retest data from previous participants. As mentioned by Wilde (2008, p. vii), the establishment of Teamology instrument is based on his sixteen year study on engineering student design teams. Thus in order to decrease the predictable error as much as possible, the test object is set as engineering students who took part in design project teams. Teamology test was given to the selected participants twice after a period of time to collect the test-retest data. The collected data were compared to measure the test-retest reliability, while the time interval between two tests were recorded as well to determine if consistency have a systematic change over time.

The test object has been restricted to engineering students who took Teamology test before when they took part in design project teams several months ago. In total 303 participants were chosen from 8 different course sections. Among these 8 sections, 5 are from ME2024 Introduction to Engineering Design and Economics at Virginia Tech Mechanical Engineering Department, in which participants are mostly sophomore mechanical engineering students. The remaining 3 sections are from PACE Global Course sponsored by Partner for the Advancement of Collaborative Engineering Education Program (PACE) and lead by Virginia Tech. The participants from PACE Global Course are mainly senior or graduate engineering students from four different countries: China, Germany, Mexico, and the U.S. The detail information for all 8 sections can be seen in Table 3.1 below.

Table 3.1 Detail Information about 8 Course Sections Selected as Participant Pool

Section #	1	2	3	4	5	6	7	8
First Response	Aug. 2012	Jan. 2013	Aug. 2013	Jan. 2014	Aug. 2014	Aug. 2012	Aug. 2013	Aug. 2014
Second Response	Jun. 2015	Jun. 2015	Jun. 2015	Jun. 2015	Jun. 2015	Jun. 2015	Jun. 2015	Jun. 2015
Time Interval	33 months	28 months	21 months	16 months	9 months	33 months	21 months	9 months
Participants	29 U.S. Sophomore	31 U.S. Sophomore	30 U.S. Sophomore	31 U.S. Sophomore	33 U.S. Sophomore	21 Chinese 8 Germany 10 Mexican 18 U.S.	9 Chinese 11 Germany 8 Mexican 13 U.S.	16 Chinese 11 Germany 7 Mexican 14 U.S.

An invitation letter for the second test was distributed to all 303 selected participants via E-mail. 131 valid replies were received in 3 weeks, yielding an overall participant rate of 43.2%. Among these replies, 56 of them came from the ME2024 Introduction to Engineering Design and Economics sections, and 75 were from PACE Global Course sections.

3.2 Test-retest Reliability of the Teamology Instrument

In this section, the test-retest reliability of Teamology instrument is verified by calculating the Pearson's correlation coefficient, the average absolute difference, and the consistency percentage between the two tests. Take dimension EI for example, the formula is shown below in Eq. 3.1 and Eq. 3.2. In Eq. 3.1, n is the sample number, in Eq. 3.2, Scale is the score range, which is 2 for each item and 10 for each dimension.

$$\text{Average Absolute Difference} = \overline{\left(\sum_1^n |New\ EI\ score - Old\ EI\ score|\right)/n} \qquad \text{Eq. 3.1}$$

$$\text{Consistency Percentage} = \left(1 - \frac{Average\ Absolute\ Difference}{Scale}\right) \times 100\% \qquad \text{Eq. 3.2}$$

If the test-retest reliability is high by having consistency percentage higher than 80%, the Teamology instrument can be judged as an acceptable psychological preference measurement tool with high validity. As designed by Wilde, each item in Teamology instrument have two choices while the participants have the option of selecting one, the other, both words or neither words (Wilde, 2008, p. 11). By selecting the first option, one point is added to the total score, while minus one point will be added to the total score if the second option is chosen. This gives each item a score range of [-1, 1], and the maximum amount can change for single item is 2. For the same reason, the score range for each psychological dimension is [-5, 5] with a maximum change amount of 10 since there are 5 items in one dimension. As mentioned in Section 2.3, the reason of having five questions for single dimension is to reflect approximately 80% consistency of personal traits over time (Kirby et al, 2007). Thus the average absolute difference between first and second score for each item should be around 0.4, and for each dimension, the average absolute difference should be around 2. In the following section, the average absolute difference and percentage of consistency for each question and each dimension between first and second test will be calculated to determine the test-retest reliability of Teamology survey. The hypothesis of 80% consistency over time will be validated as well.

3.2.1 Overall Performance

Since two tests are administrated to same group of people to determine the test-retest reliability, the most straightforward way to figure out the performance is to calculate the absolute difference between the scores from first test and second test. The average absolute differences for each question and each dimension from all 131 participants are shown below in Table 3.2 through 3.5. In these four tables, the first and second row represents the average absolute difference and standard deviation of the difference for each question or dimension, while the third row shows the percentage of consistency over time. The data of consistency over time is calculated by the unchanged percentage of each item in its score range. Based on the hypothesis of 80% consistency over time made in Section 1.2.1, an average of 80% consistency could be expected for each dimension, while lower score for specific items could be expected as well since change for each individual is not evenly distributed into all questions. For each item, score range is [-1, 1], so scale for average difference is 2. For overall score of a single dimension, the score range is [-5, 5], which lead to a scale of 10, as shown in Table 3.2 through Table 3.5.

Table 3.2 Average Absolute Difference for Dimension EI

Items	**EI1** You are more: (e) sociable (i) reserved	**EI2** You are more: (e) expressive (i) contained	**EI3** You prefer: (e) groups (i) individuals	**EI4** You learn better by: (e) listening (i) reading	**EI5** You are more: (e) talkative (i) quiet	**EI Overall**
Average Difference	0.45 [2]	0.50 [2]	0.65 [2]	0.59 [2]	0.50 [2]	1.78 [10]
Standard Deviation	0.81	0.83	0.88	0.87	0.84	1.65
Percentage of Consistency	77.27%	75.21%	67.36%	70.66%	75.21%	**82.21%**

Table 3.3 Average Absolute Difference for Dimension JP

Items	**JP1** You are more: (j) systematic (p) casual	**JP2** You prefer activities: (j) planned (p)open-ended	**JP3** You work better: (j) with pressure (p)without pressure	**JP4** You prefer: (j) routine (p) variety	**JP5** You are more: (j) methodical (p)improvisational	**JP Overall**
Average Difference	0.72 [2]	0.48 [2]	0.38 [2]	0.59 [2]	0.58 [2]	1.86 [10]
Standard Deviation	0.91	0.84	0.76	0.89	0.84	1.64
Percentage of Consistency	64.05%	76.03%	80.99%	70.66%	71.07%	**81.37%**

Table 3.4 Average Absolute Difference for Dimension SN

Items	**SN1** You prefer the: (s) concrete (n) abstract	**SN2** You prefer: (s)fact-finding (n)speculating	**SN3** You are more: (s) practical (n)conceptual	**SN4** You are more: (s) hands-on (n)theoretical	**SN5** You prefer the: (s) traditional (n) novel	**SN Overall**
Average Difference	0.50 [2]	0.68 [2]	0.67 [2]	0.32 [2]	0.62 [2]	1.95 [10]
Standard Deviation	0.81	0.90	0.91	0.71	0.88	1.84
Percentage of Consistency	75.21%	66.12%	66.53%	83.88%	69.01%	**80.46%**

Table 3.5 Average Absolute Difference for Dimension TF

Items	**TF1** You prefer: (t) logic (f) empathy	**TF2** You are more: (t) truthful (f) tactful	**TF3** You see yourself as more: (t) questioning (f) accommodating	**TF4** You are more: (t) skeptical (f) tolerant	**TF5** You think judges should be: (t) impartial (f) merciful	**TF Overall**
Average Difference	0.26 [2]	0.54 [2]	0.57 [2]	0.65 [2]	0.54 [2]	1.66 [10]
Standard Deviation	0.64	0.84	0.83	0.88	0.81	1.40
Percentage of Consistency	86.78%	73.14%	71.49%	67.36%	73.14%	**83.44%**

The complete raw data of this test is available in Appendix B. As mentioned in Table 3.2 through 3.5, all four overall absolute average differences for each dimension (EI, JP, SN, TF) is higher than 80%, with $\mathrm{EI} = 82.21\%, \mathrm{JP} = 81.37\%, \mathrm{SN} = 80.46\%$ and $\mathrm{TF} = 83.44\%$. This means the hypothesis that 80% consistency of personal traits over time is true. On the other hand, the Pearson's correlation coefficient (Karl Pearson, 1880) between old data and new data for all four dimensions are calculated by Eq. 3.3, where cov is the covariance, and ρ is the standard deviation. The result of Pearson's correlation coefficients is listed below in Table 3.6.

$$\rho_{\text{old data,new data}} = \frac{\text{cov(old data,new data)}}{\rho_{\text{old data}} \times \rho_{\text{new data}}} \qquad \text{Eq. 3.3}$$

Table 3.6 Pearson’s correlation coefficient and Percentage Consistency for each Dimension

Dimension	EI	JP	SN	TF
Percentage of Consistency	82.21%	81.37%	80.46%	83.44%
Pearson’s Correlation	0.70	0.55	0.46	0.43

Table 3.6 shows that the Pearson’s correlation coefficient for each dimension ranges from 0.43 to 0.70. As mentioned in Section 2.3, a Pearson’s correlation coefficient for test-retest reliability of psychological test is judged high if it is higher than 0.6 (Cohen, 2013, p. 78), while a data between 0.3 to 0.6 is still acceptable. Compared with the MBTI test-retest correlation data of 0.67 to 0.73 mentioned in Section 2.3 (Stricker & Ross, 1964), Teamology has a lower result, while it is still in the acceptable range. Stricker and Ross also mentioned a result of 0.48 for TF dimension (Stricker & Ross, 1964), while here also shows that TF dimension has the lowest correlation data (TF = 0.43) between old result and new result, which means the test-retest reliability for dimension TF could be worse than other 3 dimensions.

3.2.2 Performance for Each Item

As far as the test-retest reliability performance for all four dimensions are proved satisfactory, the performance for each item is considered to figure out if test-retest reliability is consistent across all 20 Teamology questions. As mentioned in Table 3.2 through Table 3.5, the consistency percentage for each item varies from 64.05% (Dimension JP Item 1) to 86.78% (Dimension TF Item 1), which indicate that some items have a good consistency over time but some are not satisfactory. Box plot is used to figure out the item with relatively poor performance on test-retest reliability. The box plot in Figure 3.1 and statistical data in Table 3.7 shows the distribution of average absolute difference and consistency percentage of all 20 Teamology items. Related definitions are listed below:

$$\text{First quartile (Q1): splits off the lowest 25\% of data from the highest 75\%}$$

$$\text{Third quartile (Q3): splits off the highest 25\% of data from the lowest 75\%}$$

$$\text{Interquartile range (IQR)} = \text{Q3} - \text{Q1} \qquad \text{Eq. 3.4}$$

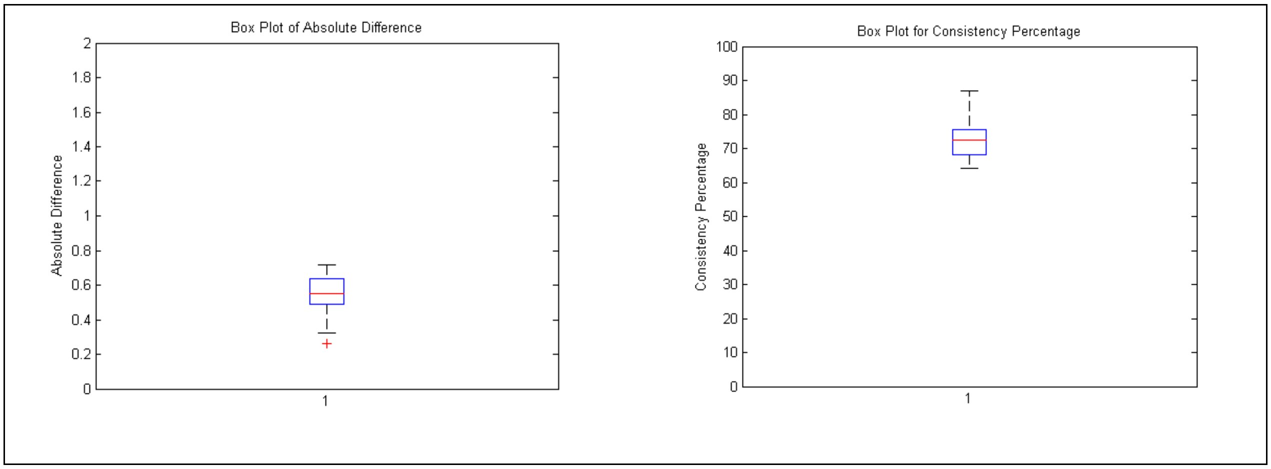


Figure 3.1 Box Plot for 20 Items

Table 3.7 Statistic Data for 20 Items

	Minimum	Q1	Median	Q3	Maximum	Average	IQR
Absolute Difference	0.26	0.49	0.55	0.63	0.72	0.54	0.14
Consistency Percentage	64.05%	68.60%	72.31%	75.41%	86.78%	73.06%	6.82%

Based on the data shown in Table 3.7, the outlier zone of data points can be determined:

$$\text{MAX Bound for Consistency Percentage} = \text{Q3} + 1.5\text{IQR} = 85.64\% \qquad \text{Eq. 3.5}$$

$$\text{MIN Bound for Consistency Percentage} = \text{Q1} - 1.5\text{IQR} = 58.37\% \qquad \text{Eq. 3.6}$$

Which means if the overall consistency over time for specific item is lower than 58.37%, it is judged to be out of the reasonable range, and it is highly likely to be determined having a poor test-retest reliability. If the box plot in Figure 3.1 is concerned, the data point will be lower than the dash line period and judged invalid. When all 20 Teamology items are determined, the lowest data is 64.05% (Dimension JP Item 1), which means all items are acceptable when overall result is concerned. The same calculation is applied to all dimensions under all time period as well in order to figure out the test-retest reliability of each item under the condition of different time period. If there is too much outlier under different time period for specific item, it will still be judged as not acceptable. The outlier minimum range for consistency percentage over time under different time period is listed in Table 3.8.

Table 3.8 Outlier Range for Each Time Period

Section	Q1	Q3	IQR	MIN Range
2 (Spring 2013)	50.00%	83.33%	33.33%	0.00%
3 (Fall 2013)	63.89%	78.47%	14.58%	42.01%
4 (Spring 2014)	61.54%	81.73%	20.19%	31.25%
5 (Fall 2014)	61.11%	88.89%	27.78%	19.44%
6 (Fall 2012)	62.50%	81.25%	18.75%	34.38%
7 (Fall 2013)	68.75%	84.62%	15.87%	44.95%
8 (Fall 2014)	71.97%	77.65%	5.68%	63.45%
Fall 2013 Overall	67.90%	78.41%	10.51%	52.13%
Fall 2014 Overall	68.75%	76.79%	8.04%	56.70%

Table 3.8 is generated based on the distribution of consistency percentage among all 20 Teamology questions. If any of the consistency percentage data is lower than the minimum range in its own time section, it is judged as poor test-retest reliability. The complete data set for each time period can be found in Appendix C. The outlier and lowest score for each time period are counted to determine the stability of each item over time.

As a result, the performance of Dimension JP Item 1 is not satisfactory because it has totally four outliers and lowest data among all time period, which is shown below. The performance for all time period, fall 2013 and 2014 graduate student are all lowest among all items, while for fall 2014 overall performance, it is in the outlier area which means the consistency data is significantly lower than other items. The bold in the item JP1 description table highlight these lowest data or outlier.

Item JP1 You are more: (j) systematic (p) casual										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	**64.05%**	75.00%	83.33%	**57.69%**	77.78%	65.91%	76.92%	**46.97%**	66.67%	**51.19%**

For the same reason, the consistency over time for dimension EI Item 3, dimension SN Item 2, and dimension TF Item 4 are also doubtful, which is shown below. These items have a worse performance on consistency over time than other items since it has two or three outliers and lowest score. The score comparison Table can be found in Appendix D. All lowest data and outliers for all time period among all time period are listed.

EI3 You prefer: (e) groups (i) individuals									
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	67.36%	**53.13%**	66.67%	84.62%	63.89%	76.14%	65.38%	**65.15%**	61.11%	**64.29%**

SN2 You prefer: (s) fact-finding (n) speculating										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	66.12%	62.50%	66.67%	84.62%	**50.00%**	70.45%	**46.15%**	72.73%	55.56%	69.05%

TF4 You are more: (t) skeptical (f) tolerant										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	67.36%	68.75%	100.00%	59.62%	75.00%	65.91%	61.54%	**60.61%**	83.33%	**65.48%**

3.2.3 Brief Conclusion

Based on statistical analysis in this section, the overall test-retest performance for each of the four dimensions (EI, JP, SN, TF) are satisfactory. The hypothesis that 5 questions in each dimension can reflect 80% consistency over time is confirmed, which means the overall test-retest reliability of Teamology survey is good. At the same time, for the 20 items in the Teamology survey, the test-retest reliability are generally acceptable, though some questions should be used with concern, including Dimension JP Item 1, Dimension EI Item 3, Dimension SN Item 2, and Dimension TF Item 4 because of its poor performance on some time period.

3.3 Systematic Change and Long Term Drift of Teamology Survey

The prerequisite of verifying test-retest reliability of a measurement tool is that the object measured will not change during the test-retest time period. When the psychological preference measurement tool is concerned, this lead to the hypothesis that psychological type of individual people remains stable over life span (Quenk, 2009, p. 32). However, it is not always the case; younger participants like college students are more likely to develop their psychological type during life span (Quenk, 2009, p. 32). As mentioned before in Section 2.3, the psychological test result might yield systematically over time due to the following reasons: Attribute changes between two tests; Effect of taking the first test and Carryover Effect if the interval between two tests is short (Davidshofer & Murphy, 2005, p. 123). Some examples suggested that special experience like team-based learning can develop individual participant and guide healthy psychological development (Wilde, 2008, p. 54). Thus in this section, the performance of test-retest reliability of Teamology survey under different time periods are compared to figure out if there is systematic change over time.

In this section, the hypothesis is that individual's psychological type remains stable during the life span, which means no systematic change should occur over time. If the absolute difference between two tests becomes obviously larger over time, or on the other word the test-retest reliability reduce when time period between two tests increase, systematic change for psychological type over time is determined to exist. The detail comparison and analysis will be shown in the following section.

As mentioned in Section 3.1, the participants used in this thesis mainly comes from two different resources: 5 sections 56 students from undergraduate design course and 3 sections 75 students from graduate design course. In order to eliminate the influence caused by different participant group, the data of overall performance, undergraduate student performance and graduate student performance are presented simultaneously. The consistency result for dimension EI is listed below in Figure 3.2 and Table 3.9.

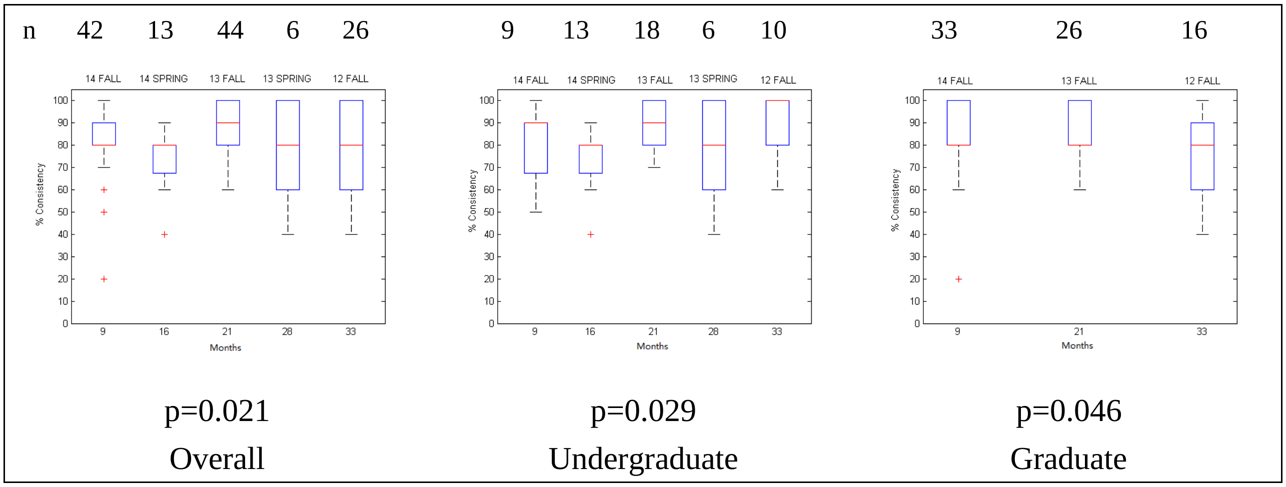


Figure 3.2 Performances for dimension EI over time

Table 3.9 Performances for dimension EI over time

	14FA	14SP	13FA	13SP	12FA	14FA(U)	13FA(U)	12FA(U)	14FA(G)	13FA(G)	12FA(G)
Participants	42	13	44	6	26	9	18	10	33	26	16
Mean (%)	80.95	73.85	88.41	76.67	79.23	80.00	90.56	88.00	81.21	86.92	73.75
Q1 (%)	80.00	70.00	80.00	65.00	60.00	70.00	82.50	80.00	80.00	80.00	60.00
Q3 (%)	90.00	80.00	100.00	95.00	100.00	90.00	100.00	100.00	100.00	100.00	85.00

In table 3.2, the average consistency for dimension EI is 82.21%. As shown in Figure 3.2 and Table 3.9, there is no obvious evidence prove that the overall performance for dimension EI has a systematic change over time. However, when the box plot for overall EI performance is concerned, it can be easily figure out that the distribution becomes bigger when the time period between two tests increase, which means there is a tendency that some participants change their preference in Dimension EI over time. The separate undergraduate student and graduate student data indicate it in detail: undergraduate students have a relatively stable consistency data over time for Dimension EI, while graduate student shows a tendency that the consistency is decreasing when the time period between two tests increase. Although accurate quantative analysis could not apply due to the small amount of data points, a preliminary speculation is that systematic change on dimension EI might occur for graduate students after more than 30 months.

In order to achieve a quantative conclusion, analysis of variance (ANOVA) is applied to determine if the systematic change is significant. If the p-value generated by ANOVA is equal or smaller than the significance level of 0.05, it means this group of data has a significant difference. Pairwise t-test is applied afterward between each group of data to figure out the exact group with significantly different. For Dimension EI shown in Figure 3.2 and Table 3.9, the p-value for overall performance is 0.021, for graduate student it is 0.046, and 0.029 for undergraduate student. These results lead to a primary conclusion that the consistency for Dimension EI has a significant difference over time. Pairwise test is applied as well to figure out which group of data shows a significant difference. Result of pairwise test shows that all p-value related to data for fall 2013 semester are significantly small, which means the consistency of result for 2013 fall semester is significantly different from others. When track back to box plot and average data, it shows that the consistency for 2013 fall semester is significantly higher than other semesters. For p-value between 2012 fall graduate student and 2014 fall graduate student, a result of 0.13 is shown, which means difference is much likely to be found but it is not significant enough currently. However, Fisher indicates that the significant level can be set according to specific circumstances (Quinn & Keough, 2002, p. 46–69). Analysis of variance indicates that a systematic change on consistency over time for dimension

EI could be expected, but it is not significant enough currently. The full result for p-value is available in Appendix E.

The result for dimension JP is listed below in Figure 3.3 and Table 3.10.

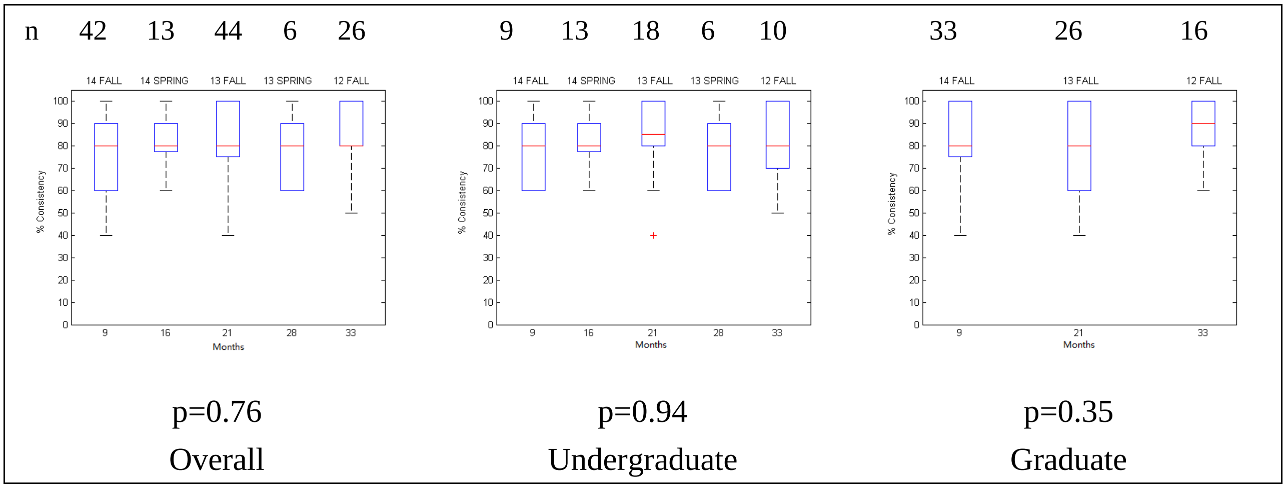


Figure 3.3 Performances for Dimension JP over time

Table 3.10 Performances for Dimension JP over time

	14FA	14SP	13FA	13SP	12FA	14FA(U)	13FA(U)	12FA(U)	14FA(G)	13FA(G)	12FA(G)
Participants	42	13	44	6	26	9	18	10	33	26	16
Mean (%)	79.76	81.54	81.14	78.33	85	78.89	83.33	82	80	79.62	86.88
Q1 (%)	65	80	77.5	65	80	60	80	72.5	100	62.5	80
Q3 (%)	90	90	100	87.5	100	90	97.5	100	80	100	100

The performance for Dimension JP over time is generally consistent. As shown in Table 3.10, the difference between maximum data and minimum data is less than 10%. When the box plot in Figure 3.3 is concerned, there is no evidence shown that systematic change exist in Dimension JP as well,

ANOVA test shows a similar result for dimension JP. The p-value for dimension JP is 0.76 for all participants, 0.35 for graduate students and 0.94 for graduate student. This indicate that no significant change can be expected among dimension JP, Thus a primary conclusion for Dimension JP is that systematic change over time is not obvious and could not be expected currently.

The result for dimension SN is listed below in Figure 3.4 and Table 3.11.

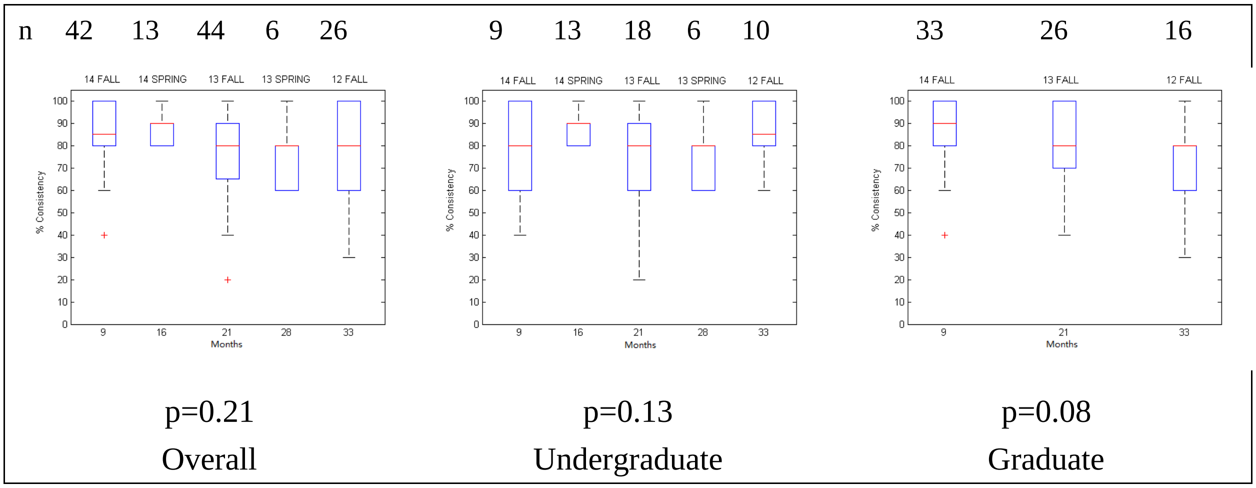


Figure 3.4 Performances for Dimension SN over time

Table 3.11 Performances for Dimension SN over time

	14FA	14SP	13FA	13SP	12FA	14FA(U)	13FA(U)	12FA(U)	14FA(G)	13FA(G)	12FA(G)
Participants	42	13	44	6	26	9	18	10	33	26	16
Mean (%)	84.29	86.92	76.82	76.67	78.08	78.89	70	85	85.76	81.54	73.75
Q1 (%)	80	80	67.5	65	60	60	60	80	80	70	60
Q3 (%)	100	90	90	80	97.5	100	90	100	100	100	80

The performance of Dimension SN is similar to the performance of Dimension EI. The overall performance presented in Figure 3.4 and data in Table 3.11 shows that there is a slightly decrease over time for the consistency of Dimension SN with the difference between maximum and minimum near 13%. When undergraduate student is concerned, the existing data cannot confirm the tendency but a slightly decrease over time could be expected. On the other hand, the data from graduate student indicate a systematic change might occur over time.

For ANOVA result on dimension SN, overall data shows a p-value of 0.21, while graduate student get a 0.08 and undergraduate get a 0.13, which means it is likely that systematic change for consistency occurs over time. When pairwise test is applied, for graduate students, p-value between 0.30 and 0.48 is observed, which means no significant difference between time periods happened. However, the p-value of 0.07 and decreasing mean indicate the fact that it might decrease smoothly over time. For undergraduate student, pairwise f-test shows that spring 2014 semester has a significantly high consistency, if ANOVA is applied to other 4 semesters separately, it shows a p-value near 0.35, which means there is no systematic change over timefor undergraduate students. Thus, a speculation for Dimension SN is that a systematic change over time could be expected for graduate participant, while the change is smooth.

Finally, the result for dimension TF is listed below in Figure 3.5 and Table 3.12.

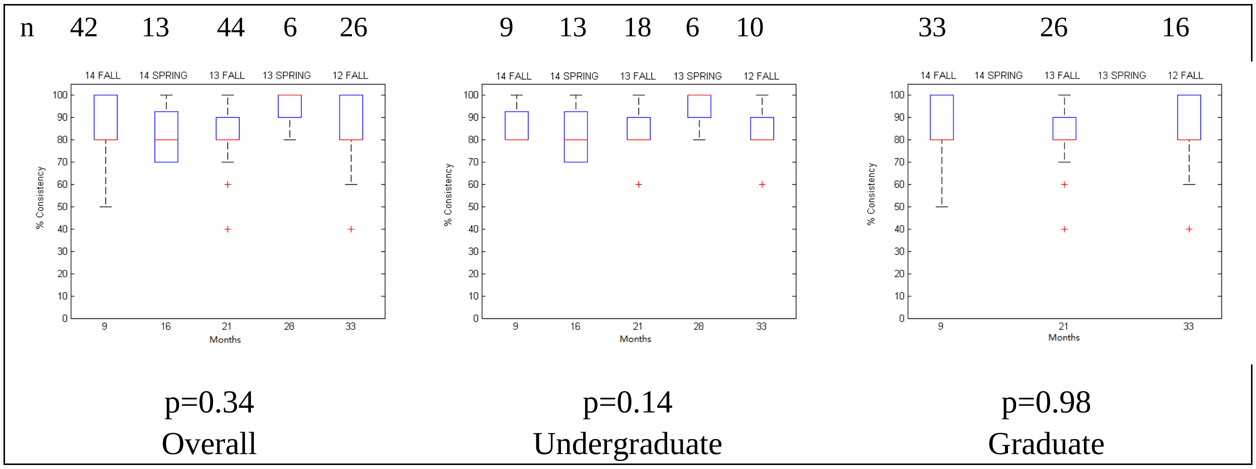


Figure 3.5 Performances for Dimension TF over time

Table 3.12 Performances for Dimension TF over time

	14FA	14SP	13FA	13SP	12FA	14FA(U)	13FA(U)	12FA(U)	14FA(G)	13FA(G)	12FA(G)
Participants	42	13	44	6	26	9	18	10	33	26	16
Mean (%)	83.57	81.54	82.95	95	82.31	86.67	83.33	83	82.73	82.69	81.88
Q1 (%)	80	70	80	92.5	80	80	80	80	80	80	80
Q3 (%)	100	90	90	100	97.5	90	90	87.5	100	90	100

For Dimension TF presented in Figure 3.5 and Table 3.12, box plot and statistical data shows that the result have a pretty high consistency over time, no matter whether it is from undergraduate student or graduate student. ANOVA raise an overall p-value of 0.34 and 0.98 for graduate student. For undergraduate the p-value is 0.14 while it is caused by data from 2013 spring semester since it is significantly high and small sample size. If 2013 spring semester is eliminated, a p-value of 0.77 is shown which represent a conclusion that no systematic change on consistency can be expected. Thus for Dimension TF, no systematic change over time can be found, no matter the participants are.

3.3.1 Brief Result

In Section 3.3, four dimensions in Teamology survey has been tested separately to determine if there is systematic change for the reliability of psychological preference test over time. The box plot for each dimension under different time period is plotted; the statistical data including average consistency over time are presented as well to indicate if systematic change do exists. ANOVA and pairwise test is applied to verify if the brief speculation from box plot is true or not.

A conclusion based on the current data has been presented: Systematic change for reliability can be observed for dimension EI and SN, especially for participants from graduate design course. For dimension JP and TF, it is more likely to keep its reliability over time, or say no obvious systematic change is observed for these two dimensions. However, based on quantitative analysis based on ANOVA and pairwise test, the systematic change over time on dimension EI and SN for graduate student are not significant enough, further research is needed to figure out if the change will becomes significant when time period becomes even longer.

Since the undergraduate participants in this research experiment are mostly sophomore mechanical engineering student when they took the first test, the undergraduate student data collected mainly covers the time period from sophomore through graduate, thus it can be simply judged as testing if systematic change of psychological preference will occur during undergraduate time period. For the same reason, the graduate student data collected can be simplified as post-graduate a participant. The result shows that it is less likely that psychological preference will have a systematic change over undergraduate time period, while psychological preference on Dimension EI and SN might have some change over time after graduate.

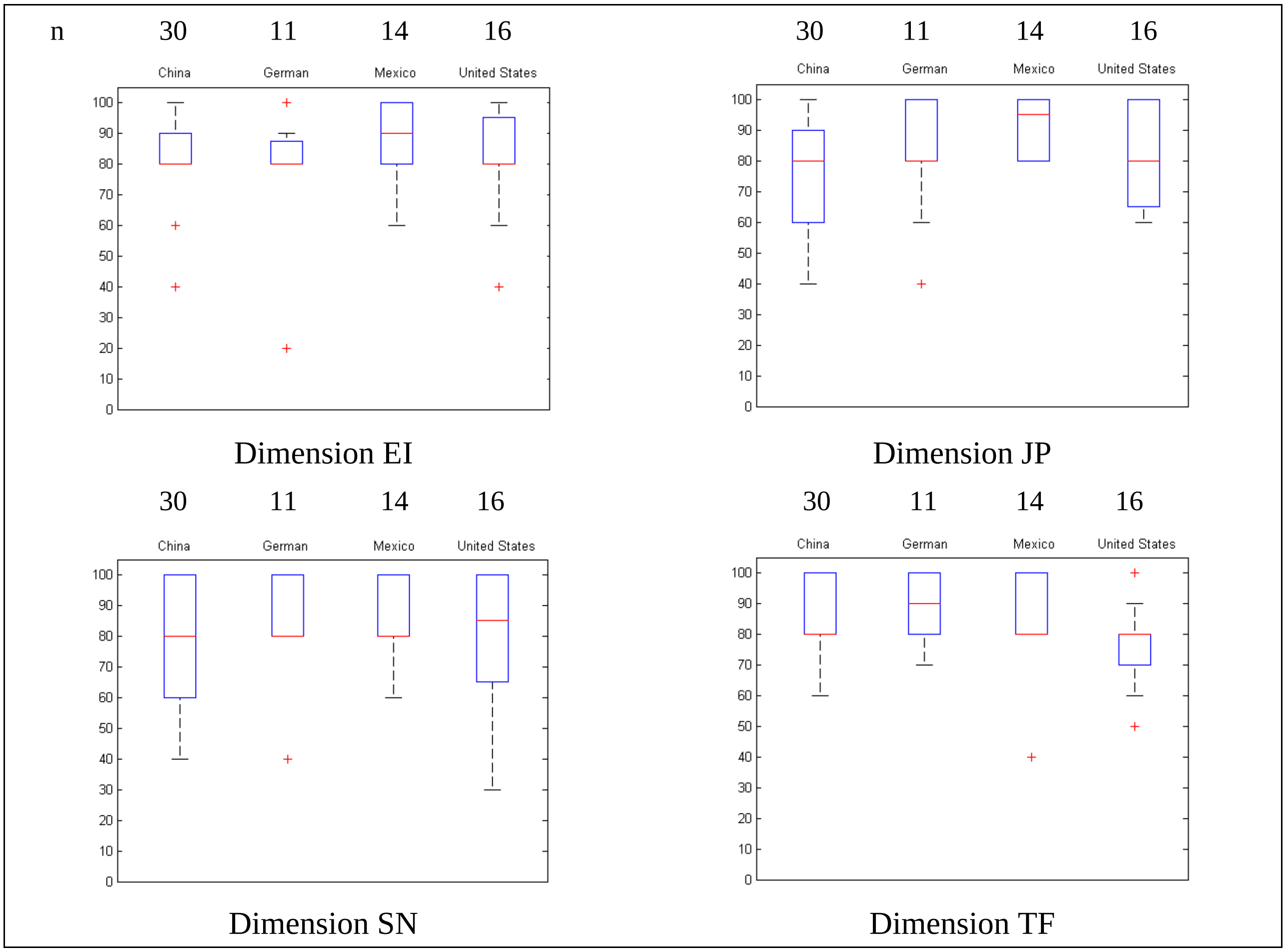
The conclusion presents above is still primary based on limit data points. If condition permitted, a third test can be administrated to figure out if systematic change does exist over time for each of the participants.

3.4 Impact Caused by Language and Culture Background

MBTI and Teamology are all firstly designed in English. Benefit from its long history and high reputation, MBTI has been officially translated into 20 languages. At the same time, English is still the most reliable language version for Teamology survey. However, even for MBTI, participants whom English is his or her second language will have some difficulty answering the survey questions due to the unfamiliar expression used (Quenk, 2009, p. 35). Thus it is necessary to determine if the language and culture difference will affect the reliability of Teamology survey because of obscure words used. In the following section, participants are sorted into four groups based on their nationality. Their reliability results for Teamology test are compared to figure out the impact caused by language and culture.

As mentioned in Section 3.1, there are altogether 131 participants in the experiment, from which most of the undergraduate participants are American. Graduate participants mainly come from 4 different countries: China, German, Mexico and United States. Since the impact caused by language, or say mother tongue is the most important fact considered in this section, overseas student is sorted by his or her homeland. In order to unify the participant group and decrease the error during analysis, only data from graduate participants are used in this section. Finally there are 30 participants from China, 11 from German, 14 from Mexico and 16 from United States. The raw test-retest result data used in this section can be found in Appendix F.

The participants' performance from different countries can be found in Figure 3.6 and Table 3.13 below. In Figure 3.6, the upper left chart represent Dimension EI, upper right is Dimension JP. Dimension SN is in the lower left side while Dimension TF is at lower right side. All four countries are listed for comparison.



Dimension EI

Dimension JP

Dimension SN

Dimension TF

Figure 3.6 Reliability Test for Different Language and Culture Background

Table 3.13 Average Consistency over Time for Different Countries

	China	German	Mexico	United States	ANOVA p
EI	79.67%	79.09%	88.57%	81.25%	0.36
JP	75.00%	83.64%	90.71%	82.50%	0.04
SN	78.33%	85.45%	85.00%	81.88%	0.58
TF	83.67%	89.09%	78.57%	76.88%	0.18

As shown in Figure 3.6 and Table 3.13, participants from Mexico have a relatively higher performance in Dimension EI and JP, while Chinese has a low consistency on Dimension JP. When Dimension SN is concerned, participants from all four countries have pretty close results, while Chinese and Americans tend to have a wider consistency distribution. For Dimension TF, American has a comparatively low reliability than people from other countries. ANOVA for this section lead to quantative results of consistency over time. For dimension EI and SN, ANOVA indicates that there is no significant difference. For dimension JP, an ANOVA p-value of 0.04 indicate that there is significant difference between countries, while pairwise test shows that the result from Mexican is significantly high, and for Chinese, the difference is acceptable. A similar result shows on dimension TF, Mexican shows a significantly high score, while Americans are acceptable. In general, reliability results from different countries are satisfactory. Mexican has higher score in some dimension, but it could be caused by the limitation of sample size. Thus as a brief conclusion is that currently there is no clear evidence shows culture and language difference will lead to a systematic difference on the test-retest reliability for Teamology Test, while some obscure expression used in specific items might still be a burden for participants who use English as second language, for example Dimension JP for Chinese.

3.5 Sensitivity of the Park Creativity Index and MBTI Creativity Index

Psychological preference test such as MBTI and Teamology test cannot measure one’s creativity directly. Alternatively some researches develop experimental calculation of creativity based on psychological test scores. One famous approach is MBTI Creativity Index developed by Gough (1981). MBTI CI is developed for all people, it is a simple linear transformation of MBTI scores which judges I, N, F and P as positive values and E, S, T and J as negative values. The score for each individual represent his or her creativity by the quantity of ideas generated in a specific period of time. The higher your score is, the more likely you can generate more ideas in the same period time. The formula of MBTI CI is shown below:

$$\mathrm{MBTI\ CI} = 250 + 3\mathrm{N} + \mathrm{P} - \mathrm{I} - 0.5\mathrm{F} \qquad \text{Eq. 3.7}$$

MBTI CI is designed for measuring the creativity of general human. For specific participant group such as college mechanical engineering students, modification is needed for experimental formula calculating creativity. Park developed the Creativity Index for undergraduate mechanical engineering student (Park, 2014, p. 104). For Park CI, it measures one's innovation in engineering teams, higher score represent better performance in engineering teams. The formula of Park CI which is shown below:

$$\mathrm{Park\ CI} = 3.28 - 0.01\mathrm{N} - 0.38\mathrm{P} - 0.33\mathrm{I} + 0.05\mathrm{F} \qquad \text{Eq. 3.8}$$

Since the consistency over time of Teamology survey has been confirmed in Section 3.2, the test-retest reliability of MBTI CI and Park CI should be tested as well to prove the validity. In this section, the correlation between two groups of data is calculated to figure out the reliability of MBTI CI and Park CI. Simultaneously the average absolute difference and creativity level difference are presented as well to indicate the performance of this evaluation system. The raw data used in this section can be found in Appendix B. For MBTI CI, the data used in the formula is MBTI score rather than Teamology score. However, Wilde suggested that Teamology score can be transferred to MBTI score with proper scaling (Wilde, 2008, p. 20). Teamology score can be easily transferred to MBTI score by multiplying 6 times.

The box plot for two groups of Park CI data and MBTI CI data are shown in Figure 3.7. As shown in the box plot, there is no big difference in distribution between two pairs of data. Pearson's Correlation coefficient (Karl Pearson, 1880) is 0.528 for MBTI CI and 0.516 for Park CI. As mentioned in Section 2.3, 0.5 to 0.6 is an acceptable range for Pearson's correlation coefficient when test-retest reliability is measured (Cohen, 2013, p. 78). Thus the test-retest reliability for MBTI CI and Park CI are both okay. However, when the average difference and level difference is concerned, the consistency of Park CI over time is not as good as performance of Teamology survey result. The detail data is shown below in Table 3.14.

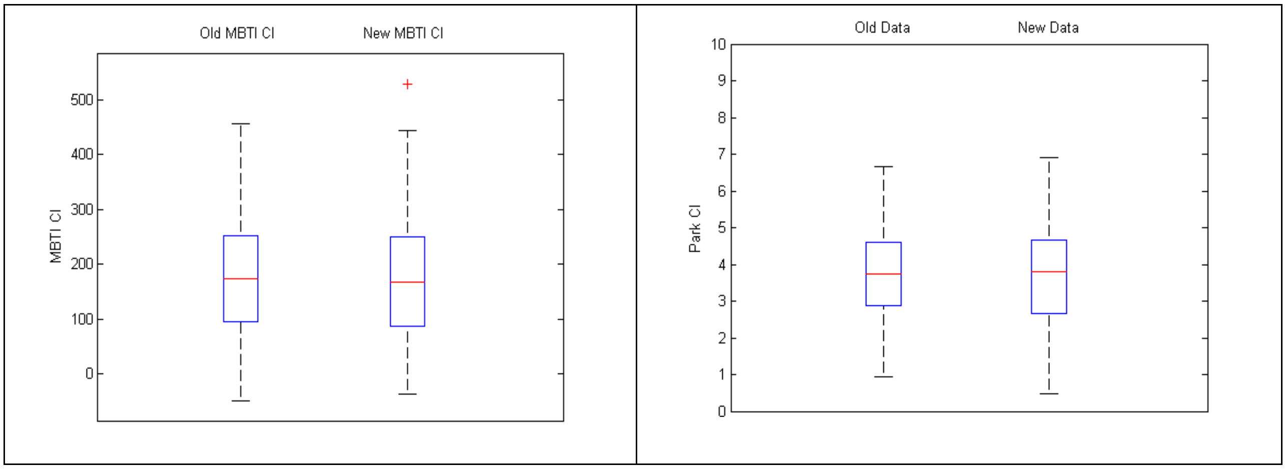


Figure 3.7 Box Plot for MBTI CI (left) and Park CI (right)

Table 3.14 Statistical Data for Park CI Comparison

Creativity Index	Difference Average	Average Absolute Difference	Consistency %	Pearson's Correlation Coefficient	p-value
Park	0.011	1.00	84.76%	0.516	0.11
MBTI	5.91	85.88	82.82%	0.528	0.16

The data in Table 3.14 indicate the performance of Park CI and MBTI CI over time. The average absolute difference shows that the average change for Park CI over time is about 1.00. An average difference near zero means that the chance that Park CI increase or decrease is almost same. For a scale with theoretical maximum value of 7.13 and minimum value 0.57, an absolute difference of 1.00 lead to a consistency of 84.76% for Park CI, while MBTI CI is 82.82%. These data indicate that the test-retest reliability of Park CI and MBTI CI are generally acceptable, but not good enough. The weighted average sensitivity is calculated in Eq. 3.9 and Eq. 3.10:

$$\text{MBTI CI average sensitivity} = \frac{3SN + JP + EI + 0.5TF}{3 + 1 + 1 + 0.5}$$

$$= \frac{3 \times 0.8046 + 0.8137 + 0.8221 + 0.5 \times 0.8344}{5.5} = 0.8121 \qquad \text{Eq.}\,3.9$$

$$\begin{aligned}\text{Park CI average sensitivity} &= \frac{0.01SN + 0.38JP + 0.33EI + 0.05TF}{0.01 + 0.38 + 0.33 + 0.05} \\ &= \frac{0.01 \times 0.8046 + 0.38 \times 0.8137 + 0.33 \times 0.8221 + 0.05 \times 0.8344}{0.77} \\ &= 0.8185 \qquad \text{Eq. 3.10}\end{aligned}$$

It can be easily determined that the sensitivity level for both MBTI CI and Park CI derived from weighted average method are around 80%, which means the MBTI CI and Park CI have a similar consistency level compared with MBTI score or Teamology score. From Eq. 3.9, MBTI CI is more sensitive on dimension SN, and from E1. 3.10, Park CI is more sensitive on dimension JP and EI. The weight difference for each dimension is determined by the difference focus on participants and projects. In section 3.3, a brief conclusion is that dimension SN and EI have higher probability that they will change over time, when track back to MBTI CI and Park CI, MBTI CI has a higher change that it will change over time because the overall weight for dimension EI and SN is higher.

In this section the test-retest reliability of MBTI CI and Park CI is discussed. Its effectiveness of measuring creativity is not discussed. However, the consistency of it has been tested with Pearson’s Correlation Coefficient and pairwise test. The reliability of Park CI and MBTI CI are acceptable with Pearson’s Correlation Coefficient 0.516 and 0.528. The sensitivity of both Creativity Index are calculated, MBTI CI is more likely to change over time because it is more sensitivity to dimension SN.

Chapter 4

Result and Conclusion

In this thesis, four research questions are addressed. Research Question 1 asked if test-retest reliability are consistent across all 20 Teamology items and 4 Dimensions, as well as determining if hypothesis that 80% consistency is true or not. This research question is answered in Section 3.2. For research Question 2, the systematic change in reliability for Teamology test over time is concerned and it is covered in Section 3.3. In Section 3.4, Research Question 3 is answered for impact on reliability by language and culture difference. Finally the Research Question 4 is answered in Section 3.5 for the reliability of Park CI.

4.1 Research Questions

Research Question 1: Is the test-retest reliability consistent across all 20 Teamology questions?

For Research Question 1, a brief conclusion is that overall test-retest performance for each of the four dimensions (EI, JP, SN, TF) are satisfactory with consistency percentage higher than 80%. The hypothesis that 5 questions in each dimension reflect 80% consistency over time is confirmed, and the Pearson's correlation coefficient for each dimension ranges from 0.43 to 0.70, thus overall test-retest reliability of Teamology survey is good. For the 20 items in the Teamology survey, the reliability is acceptable with specific items not satisfactory enough due to their poor performance in several time periods, including Dimension JP Item 1, Dimension EI Item 3, Dimension SN Item 2, and Dimension TF Item 4 because its consistency over time is not as good as other items.

Research Question 2: Is there a systematic change in reliability over time?

In Section 3.3, Research Question 2 about systematic change over time is covered. Four dimensions are tested separately to figure out if systematic change for the reliability over time does exist. The box plot and statistical data are used to indicate if systematic change do exists. A preliminary conclusion based on the current data is listed below: Systematic change in reliability is observed on Dimension EI and SN, but it is not significant enough currently. For Dimension JP and TF, no obvious systematic change is observed over time.

Research Question 3: Do native languages and/or cultural backgrounds impact the reliability?

For the impact caused by culture and language raised in Research Question 3 and answered in Section 3.4, as a brief conclusion, currently no clear evidence prove that culture and language will lead to a systematic difference on the test-retest reliability of Teamology Test. Some obscure expression used in specific items might still be a burden for participants who use English as second language, for example Dimension EI for Chinese, but the impact is limited and can be eliminated when psychological preference is tested.

Research Question 4: Does the reliability substantially affect the MBTI Creativity Index and the Park Creativity Index?

The Research Question 4 mentioned Park CI and MBTI CI. It is discussed in Section 3.5. The test-retest reliability are calculated and discussed. The Creativity Index's effectiveness of measuring creativity is not discussed in this section, while a Pearson's Correlation Coefficient between old data and new data of 0.528 and 0.516 lead to the conclusion that consistency of it is acceptable. Sensitivity analysis on MBTI CI and Park CI as well, result indicate that MBTI CI is more likely to change over time based on the conclusion that participants' psychological preference on dimension EI and SN are more likely to change over time, while the equation of MBTI CI indicates that it is more sensitivity on dimension SN.

The result presented in this thesis is a primary one. Due to the limitation of data collected, error of data is inevitable. As a future plan, a third administration of test is preferred if condition allowed so that there are three data entries for each participants. Under this condition, the test-retest reliability and systematic change over time can be tested more accurately since the error is eliminated. Furthermore, in systematic change section, more effort can be applied to figure out if age does matter with psychological change or not. Current result shows that graduate student tend to change more than undergraduate student in Dimension EI and SN, if more data can be collected, the impact caused by age, or if it is caused by special event during this period of time such as job hunting can be indicated then. For Park CI, a new level sorting regulation should be establish based on the combined analysis between creativity score and project team performance.

Overall, Teamology test is proved to have a good reliability and validity. Qualitative psychological preference test will definitely perform an important role in accurate psychological analysis and team forming. As a 20-item test with good validity, Teamology test will be popular in this field in the future.

4.2 Research Contributions

The thesis paper has made the following research contributions:

1. **Validating the reliability of 20-questions Teamology test:** The Teamology instrument is designed to share the same validity with MBTI, while the reliability of it, especially test-retest reliability has not been tested yet. In this thesis paper, the overall test-retest reliability of Teamology is validated and proved to be good, which indicate that the reliability of Teamology test is acceptable.

2. **Figure out 4 Teamology items which are not well derived:** By verifying the test-retest reliability of all 20 Teamology items, it is found that reliability result is not consistent among all 20 questions. By using box plot and statistical analysis, 4 Teamology items have been judged not satisfactory enough on test-retest reliability, revised is needed for these four items.

3. **Prove that personality traits measured by Teamology might have a systematic change over time for specific group of participants on some dimensions:** By comparing the test-retest result of different dimension under different time period, it is discovered that graduate level participants tend to change their answer on Dimension EI and SN over time, while there is little change on other participants or other dimensions.

4. **Validate that language and culture difference does not have obvious impact on psychological preference change over time:** The data grouped by nationality of participants represent that there is no obvious change on psychological test-retest reliability when culture and mother language is different, even if all participants are using English version questionnaire.

5. **Prove the validity of Park Creativity Index:** Park Creativity is based on Teamology score. Test proved that reliability of Park Creativity Index over time is acceptable with Pearson’s correlation 0.52.

Reference

Beuke, C. J., Freeman, D. G., & Wang, S. (2006). Reliability and validity of the Myers-Briggs Type Indicator® Form M when translated into Traditional and Simplified Chinese characters.

Carskadon, T. G. (1977). Test-retest reliabilities of continuous scores on the Myers-Briggs Type Indicator. Psychological Reports, 41(3), 1011-1012.

Cohen, J. (2013). Statistical power analysis for the behavioral sciences. Academic press.

Church, A. T. (2001). Personality Measurement in Cross‐Cultural Perspective. Journal of Personality, 69(6), 979-1006.

Davidshofer, Kevin R. Murphy, Charles O. (2005). Psychological testing: principles and applications (6th ed.).

Gough, H. G. (1981). Studies of the Myers-Briggs Type Indicator in a personality assessment research institute. *Isabel Briggs Memorial Library*.

Guilford, J. P. (1965). Fundamental statistics in psychology and education.

Harvey, R J (1996). Reliability and Validity, in MBTI Applications A.L. Hammer, Editor. *Consulting Psychologists Press*: Palo Alto, CA. p. 5- 29.

Jung, C. G. (1923). Psychological types: or the psychology of individuation.

Kirby, L. K., Kendall, E., & Barger, N. J. (2007). Type and culture: Using the MBTI instrument in international applications: CPP.

Myers, I. B., McCaulley, M. H., & Most, R. (1985). Manual: A guide to the development and use of the Myers-Briggs Type Indicator

Park, Y. (2014). Theory and Methodology for Forming Creative Design Teams in a Globally Distributed and Culturally Diverse Environment.

Pittenger, D. J. (1993). Measuring the MBTI… and coming up short. Journal of Career Planning and Employment, 54(1), 48-52.

Quinn, Geoffrey R.; Keough, Michael J. (2002). Experimental Design and Data Analysis for Biologists (1st ed.). Cambridge, UK: Cambridge University Press.

Quenk, N. L. (2009). Essentials of Myers-Briggs type indicator assessment (Vol. 66).

Schaubhut, N. A., Herk, N. A., & Thompson, R. C. (2009). MBTI® Form M manual supplement. Estados Unidos: CPP.

Stricker, L. J. & J. Ross. (1964). An Assessment of Some Structural Properties of the Jungian personality Typology.” *Journal of Abnormal and Social Psychology*, Vol. 68, pp. 62-71.

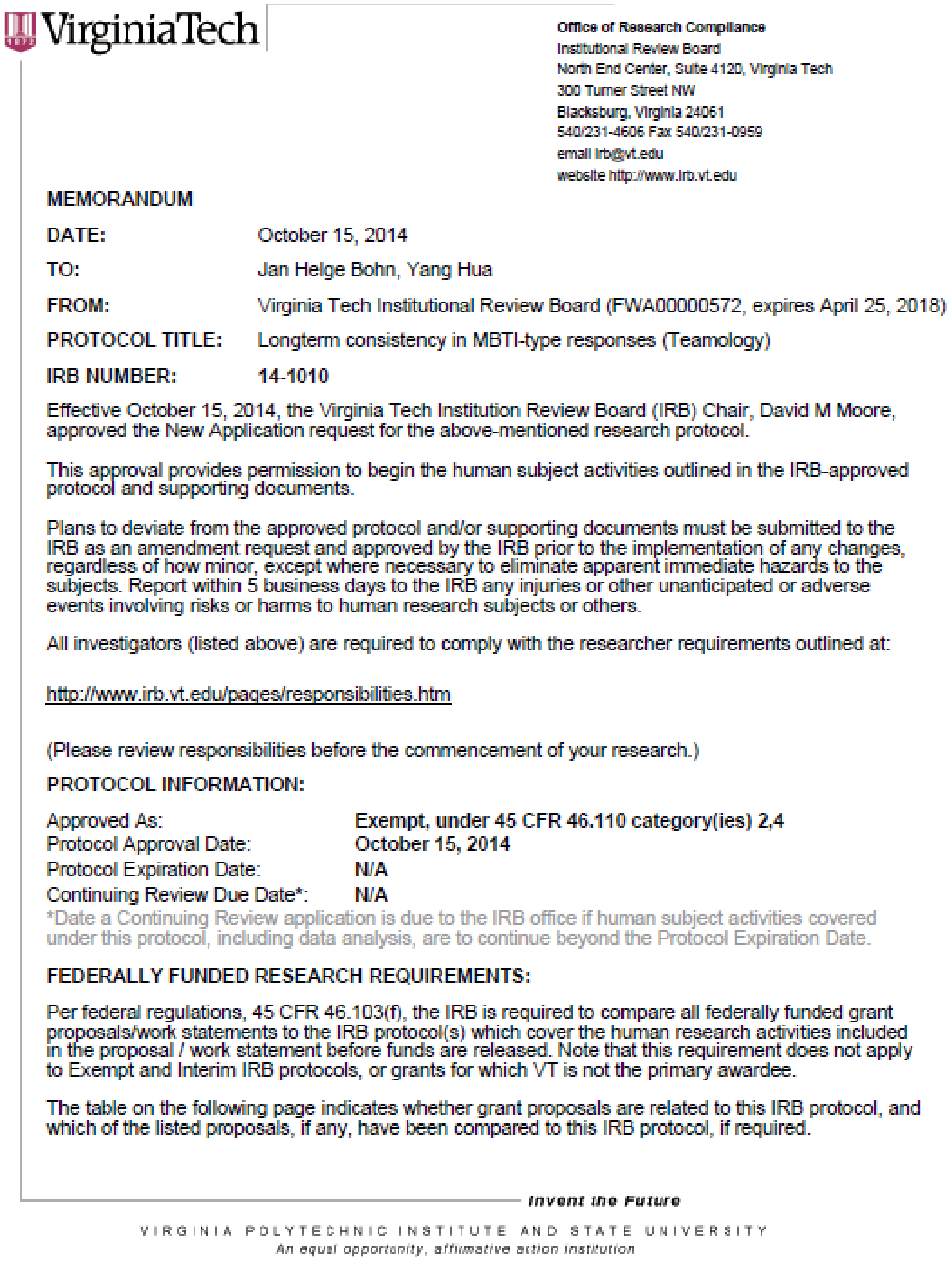
Taylor, B. N. (2009). Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results (rev. DIANE Publishing.

Tieger, P. D., & Barron-Tieger, B. (1993). Personality typing: A first step to a satisfying career. Journal of Career Planning and Employment, 53(2), 50-56.

Wilde, D. J. (2008). Teamology: The Construction and Organization of Effective Teams: The Construction and Organization of Effective Teams.

Appendix A:

IRB Application



VirginiaTech

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120, Virginia Tech
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-4606 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE:	October 15, 2014
TO:	Jan Helge Bohn, Yang Hua
FROM:	Virginia Tech Institutional Review Board (FWA00000572, expires April 25, 2018)
PROTOCOL TITLE:	Longterm consistency in MBTI-type responses (Teamology)
IRB NUMBER:	**14-1010**

Effective October 15, 2014, the Virginia Tech Institution Review Board (IRB) Chair, David M Moore, approved the New Application request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Exempt, under 45 CFR 46.110 category(ies) 2,4**
Protocol Approval Date:	**October 15, 2014**
Protocol Expiration Date:	**N/A**
Continuing Review Due Date*:	**N/A**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Invent the Future

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
An equal opportunity, affirmative action institution

Appendix B:

Raw Data

The raw data used in this thesis paper is listed in Appendix B. In the column “Study Code”, code start with “a” is fall 2012 undergraduate, “b” for fall 2012 graduate, “c” is spring 2013 undergraduate, “d” is fall 2013 undergraduate, “e” is fall 2013 graduate, “f” is spring 2014 undergraduate, “g” is fall 2014 undergraduate, “h” is fall 2014 graduate. Data start with “Origin” means it is collected when participant attend the team, and for those with “Retested”, they are all collected at May 2015.

#	Study	Original	Original	Original	Original	Original	Original	Retested	Retested	Retested	Retested	Retested	Retested
	Code	EI	JP	SN	TF	Park CI	MBTI CI	EI	JP	SN	TF	Park CI	MBTI CI
1	a1	1	-1	3	1	3.21	168.5	5	4	5	1	6.45	84.5
2	a2	5	3	5	-1	6.17	84.5	5	1	5	1	5.31	120.5
3	a3	5	-4	1	5	3.17	348.5	5	-1	-1	4	4.34	378.5
4	a4	1	5	3	3	5.39	108.5	-1	3	3	5	3.87	120.5
5	a5	1	3	3	1	4.73	120.5	-1	3	-1	-1	4.13	228.5
6	a6	-5	5	1	1	3.49	96.5	-5	5	1	-1	3.59	84.5
7	a7	3	3	5	1	5.41	72.5	3	3	5	1	5.41	72.5
8	a8	-3	-1	3	1	1.89	120.5	-3	3	2	-3	3.6	84.5
9	a9	3	1	3	1	4.63	168.5	3	1	5	3	4.55	108.5
10	a10	5	5	-1	3	6.67	300.5	1	3	3	1	4.73	120.5
11	b1	-1	3	1	5	3.85	192.5	1	3	3	3	4.63	132.5
12	b2	5	3	3	5	5.85	192.5	5	3	1	4	5.88	258.5
13	b3	3	0	1	3	4.13	264.5	5	-3	-3	-1	3.81	444.5
14	b4	3	-1	-1	3	3.73	348.5	1	1	1	-3	4.15	192.5
15	b5	1	5	1	3	5.37	180.5	-5	3	3	5	2.55	72.5
16	b6	-3	3	-2	3	3.26	264.5	-3	3	1	5	3.19	168.5
17	b7	-1	-1	5	1	2.57	72.5	-1	-1	1	3	2.43	228.5
18	b8	1	5	5	1	5.51	24.5	3	1	5	1	4.65	96.5
19	b9	-5	5	3	3	3.41	36.5	-3	5	5	3	4.09	-11.5
20	b10	-3	5	5	1	4.19	-23.5	1	3	1	1	4.71	192.5
21	b11	-1	1	1	1	3.29	192.5	-3	-1	1	1	1.87	192.5
22	b12	-3	3	1	1	3.39	144.5	3	3	3	3	5.29	156.5
23	b13	-1	5	-1	1	4.79	216.5	-5	3	3	3	2.65	60.5

24	b14	-2	-3	3	1	1.46	156.5	-2	-3	3	1	1.46	156.5
25	c1	5	-1	-3	1	4.47	432.5	1	-5	1	-1	1.77	276.5
26	c2	-1	1	5	5	3.13	72.5	5	1	-2	1	5.24	372.5
27	c3	-5	1	3	1	1.99	72.5	-5	-1	5	1	1.25	24.5
28	c4	-3	3	3	-1	3.51	60.5	-5	3	5	-1	2.87	-35.5
29	c5	-3	1	-3	1	2.59	312.5	3	-1	-1	1	3.83	336.5
30	c6	-1	-3	3	-1	1.89	156.5	-1	1	3	-1	3.41	108.5
31	c7	5	-1	3	4	4.38	234.5	1	-2	-1	3	2.69	336.5
32	c8	-3	5	-3	3	4.01	276.5	-5	1	1	1	1.97	144.5
33	d1	-2	2	1	2	3.29	174.5	-3	3	5	3	3.33	12.5
34	d2	0	-2	1	4	2.33	258.5	-2	0	2	3	2.49	168.5
35	d3	1	2	4	1	4.36	96.5	1	1	3	-3	4.17	120.5
36	d4	1	-3	3	-1	2.55	180.5	1	-3	-3	1	2.39	408.5
37	d5	-2	4	0	0	4.14	174.5	1	3	1	1	4.71	192.5
38	d6	1	3	3	3	4.63	132.5	1	3	1	-1	4.81	180.5
39	d7	-5	3	2	1	2.74	84.5	-5	5	1	3	3.39	108.5
40	d8	-5	3	1	1	2.73	120.5	-5	5	3	-1	3.61	12.5
41	d9	1	-5	-3	5	1.43	456.5	-1	-3	3	3	1.69	180.5
42	d10	-2	1	3	3	2.88	120.5	-3	-1	-5	5	1.61	432.5
43	d11	5	1	-3	5	5.03	432.5	3	5	5	5	5.97	72.5
44	d12	1	3	-1	-1	4.79	252.5	1	-3	-1	1	2.41	336.5
45	d13	-3	3	5	3	3.33	12.5	-2	2	4	2	3.32	66.5
46	d14	-4	3	3	5	2.88	84.5	-5	3	4	5	2.56	36.5
47	d15	-2	4	-2	1	4.07	252.5	-5	4	1	3	3.01	120.5
48	d16	2	3	1	-1	5.14	192.5	1	3	3	1	4.73	120.5
49	d17	1	3	2	3	4.62	168.5	1	5	5	3	5.41	36.5
50	d18	-1	-1	1	3	2.43	228.5	-1	3	5	1	4.09	24.5
51	e1	1	4	3	3	5.01	120.5	3	1	1	1	4.61	240.5
52	e2	-3	3	3	1	3.41	72.5	-1	3	1	3	3.95	180.5
53	e3	3	3	-1	3	5.25	300.5	1	-1	-1	5	2.97	336.5
54	e4	-1	5	1	3	4.71	156.5	3	-1	5	3	3.79	132.5
55	e5	-1	2	-5	1	3.61	396.5	-3	1	-1	3	2.51	252.5
56	e6	1	3	0	3	4.6	240.5	-1	-1	3	4	2.4	162.5
57	e7	3	3	-1	3	5.25	300.5	1	3	-1	-1	4.79	252.5
58	e8	3	-2	5	5	3.31	156.5	4	4	-1	5	5.86	312.5
59	e9	-3	5	4	1	4.18	12.5	-3	1	1	1	2.63	168.5
60	e10	1	1	3	-1	4.07	132.5	1	1	1	-2	4.1	198.5
61	e11	5	-1	-1	-2	4.64	342.5	5	-3	3	-3	3.97	216.5
62	e12	3	-1	3	3	3.77	204.5	5	-1	1	5	4.31	312.5

63	e13	1	3	5	1	4.75	48.5	1	3	3	0	4.78	114.5
64	e14	1	3	1	-3	4.91	168.5	3	3	-1	-2	5.5	270.5
65	e15	5	-1	3	3	4.43	228.5	5	-3	3	1	3.77	240.5
66	e16	1	-3	3	1	2.45	192.5	3	-3	3	2	3.06	222.5
67	e17	-1	3	2	3	3.96	144.5	1	5	4	3	5.4	72.5
68	e18	5	-1	3	-1	4.63	204.5	5	-3	3	5	3.57	264.5
69	e19	-1	-1	-1	-3	2.71	264.5	-1	-1	3	3	2.45	156.5
70	e20	-2	-2	5	1	1.86	72.5	-3	2	5	-1	3.15	0.5
71	e21	-1	2	2	0	3.73	138.5	1	3	5	2	4.7	54.5
72	e22	-3	-1	3	1	1.89	120.5	-5	-3	3	1	0.47	120.5
73	e23	-3	-1	5	1	1.91	48.5	-1	3	4	-2	4.23	42.5
74	e24	5	3	1	1	6.03	240.5	5	1	3	1	5.29	192.5
75	e25	-5	3	5	1	2.77	-23.5	-3	5	5	3	4.09	-11.5
76	e26	-5	5	-5	5	3.23	336.5	-5	3	-5	3	2.57	348.5
77	f1	2	3	0	-2	5.18	222.5	1	2	-1	1	4.31	276.5
78	f2	-2	3	1	1	3.72	156.5	-5	5	-1	3	3.37	180.5
79	f3	2	1	2	-1	4.39	180.5	5	5	1	2	6.74	222.5
80	f4	-1	1	-1	1	3.27	264.5	3	-1	-3	1	3.81	408.5
81	f5	-1	5	1	1	4.81	144.5	5	5	3	-1	6.91	132.5
82	f6	1	-1	1	1	3.19	240.5	-1	-3	-1	1	1.75	312.5
83	f7	-3	4	2	3	3.68	108.5	-1	5	3	1	4.83	72.5
84	f8	-1	-1	1	2	2.48	222.5	-3	-1	2	1	1.88	156.5
85	f9	-1	3	4	1	4.08	60.5	-5	5	5	3	3.43	-35.5
86	f10	-1	-4	2	0	1.45	210.5	1	0	3	3	3.49	168.5
87	f11	-1	3	3	1	4.07	96.5	-3	1	3	1	2.65	96.5
88	f12	3	0	5	-2	4.42	90.5	5	3	3	1	6.05	168.5
89	f13	2	4	4	0	5.5	78.5	1	3	3	3	4.63	132.5
90	g1	0	1	1	1	3.62	204.5	-1	3	5	1	4.09	24.5
91	g2	1	-1	1	3	3.09	252.5	3	0	5	5	4.07	132.5
92	g3	-2	-1	5	1	2.24	60.5	-3	3	-1	3	3.27	228.5
93	g4	2	1	3	2	4.25	162.5	1	-3	5	1	2.47	120.5
94	g5	-2	5	5	3	4.42	0.5	-5	5	5	3	3.43	-35.5
95	g6	5	1	3	1	5.29	192.5	1	5	5	3	5.41	36.5
96	g7	5	0	-1	3	4.77	360.5	5	1	-1	1	5.25	336.5
97	g8	4	2	1	-3	5.52	216.5	-1	0	2	-1	3.02	156.5
98	g9	-1	0	2	2	2.87	174.5	-2	1	2	1	2.97	144.5
99	h1	-1	-1	5	-1	2.67	60.5	1	-1	5	1	3.23	96.5
100	h2	4	-1	5	1	4.22	132.5	5	3	3	3	5.95	180.5
101	h3	-1	0	-4	3	2.76	396.5	1	1	-4	3	3.8	408.5

102	h4	-5	1	1	5	1.77	168.5	-1	1	1	5	3.09	216.5
103	h5	5	-1	-3	1	4.47	432.5	5	-3	-5	1	3.69	528.5
104	h6	-1	1	3	5	3.11	144.5	-1	-1	3	3	2.45	156.5
105	h7	-3	-1	1	1	1.87	192.5	-1	3	5	1	4.09	24.5
106	h8	-3	-3	-3	1	1.07	360.5	-3	-1	-1	1	1.85	264.5
107	h9	-1	2	-1	3	3.55	264.5	-1	3	-1	-1	4.13	228.5
108	h10	-3	1	5	-1	2.77	12.5	-5	1	5	-1	2.11	-11.5
109	h11	5	3	-1	3	5.91	324.5	3	1	3	1	4.63	168.5
110	h12	-3	1	-3	1	2.59	312.5	-1	-5	1	3	0.91	276.5
111	h13	-3	5	3	-1	4.27	36.5	1	5	3	1	5.49	96.5
112	h14	-5	5	5	1	3.53	-47.5	-3	5	3	-1	4.27	36.5
113	h15	-4	5	5	1	3.86	-35.5	-3	3	5	5	3.23	24.5
114	h16	-5	3	1	1	2.73	120.5	-5	1	3	1	1.99	72.5
115	h17	-3	5	1	-3	4.35	96.5	1	1	-1	1	3.93	288.5
116	h18	5	-1	3	1	4.53	216.5	3	1	3	3	4.53	180.5
117	h19	-1	1	3	-1	3.41	108.5	3	-5	-3	-3	2.49	432.5
118	h20	4	-3	1	1	3.42	300.5	4	-1	1	1	4.18	276.5
119	h21	3	1	3	3	4.53	180.5	1	5	3	5	5.29	120.5
120	h22	-3	5	1	-1	4.25	108.5	5	3	1	-1	6.13	228.5
121	h23	-1	-5	3	3	0.93	204.5	1	-5	3	3	1.59	228.5
122	h24	-1	3	-3	-1	4.11	300.5	1	3	-3	1	4.67	336.5
123	h25	1	1	-1	5	3.73	312.5	3	3	1	2	5.32	222.5
124	h26	-3	5	3	1	4.17	48.5	-3	3	5	1	3.43	0.5
125	h27	-1	1	3	-3	3.51	96.5	-3	3	5	-1	3.53	-11.5
126	h28	5	1	1	1	5.27	264.5	1	1	3	-3	4.17	120.5
127	h29	-3	-1	-5	-1	1.91	396.5	-1	-1	-5	1	2.47	432.5
128	h30	-3	1	3	-1	2.75	84.5	-5	-1	-1	1	1.19	240.5
129	h31	3	3	1	0	5.42	210.5	5	-1	1	5	4.31	312.5
130	h32	5	-3	3	2	3.72	246.5	5	-1	2	1	4.52	252.5
131	h33	1	1	-3	-1	4.01	348.5	1	5	1	3	5.37	180.5

Appendix C:

Consistency over Time for Each Item

In appendix C, the consistency percentage for each item under different time period is listed. For the time period shown in the second row, G12 means fall 2012 graduate student, SP13 means spring 2013 student, G13 means fall 2013 graduate student, FA13 means fall 2013 undergraduate student, FA13 ALL means fall 2013 student, SP14 means spring 2014 student, G14 means fall 2014 graduate student, FA14 means fall 2014 undergraduate student, FA14 ALL means fall 2014 student. The lowest score among all items in the same time period is bold in the following table, while the test procedure of it is available in Appendix D.

EI1 You are more: (e) sociable (i) reserved										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	77.27%	65.63%	83.33%	92.31%	80.56%	87.50%	**46.15%**	74.24%	100%	79.76%

EI2 You are more: (e) expressive (i) contained										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	75.21%	81.25%	83.33%	75.00%	75.00%	75.00%	61.54%	78.79%	66.67%	76.19%

EI3 You prefer: (e) groups (i) individuals										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	67.36%	**53.13%**	66.67%	84.62%	63.89%	76.14%	65.38%	**65.15%**	61.11%	**64.29%**

EI4 You learn better by: (e) listening (i) reading										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	70.66%	62.50%	66.67%	75.00%	72.22%	73.86%	80.77%	69.70%	61.11%	67.86%

EI5 You are more: (e) talkative (i) quiet										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	75.21%	62.50%	8.33%	88.46%	77.78%	84.10%	61.54%	72.73%	77.78%	73.81%

JP1 You are more: (j) systematic (p) casual										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	**64.05%**	75.00%	83.33%	**57.69%**	77.78%	65.91%	76.92%	**46.97%**	66.67%	**51.19%**

JP2 You prefer activities: (j) planned (p) open-ended										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	76.03%	87.50%	41.67%	76.92%	80.56%	78.41%	69.23%	72.73%	88.89%	76.19%

JP3 You work better: (j) with pressure (p) without pressure										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	80.99%	87.50%	83.33%	78.84%	77.78%	78.41%	88.46%	78.79%	77.78%	78.57%

JP4 You prefer: (j) routine (p) variety										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	70.66%	81.25%	33.33%	67.31%	69.44%	68.18%	65.38%	72.73%	88.89%	76.19%

JP5 You are more: (j) methodical (p) improvisational										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	71.07%	78.13%	66.67%	63.46%	61.11%	**62.50%**	84.62%	77.27%	61.11%	73.81%

SN1 You prefer the: (s) concrete (n) abstract										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	75.21%	62.50%	100.00%	75.00%	63.89%	70.45%	92.31%	77.27%	72.22%	76.19%

SN2 You prefer: (s) fact-finding (n) speculating										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	66.12%	62.50%	66.67%	84.62%	**50.00%**	70.45%	**46.15%**	72.73%	55.56%	69.05%

SN3 You are more: (s) practical (n) conceptual										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	66.53%	56.25%	50.00%	73.08%	58.33%	67.05%	80.77%	72.73%	**50.00%**	67.86%

SN4 You are more: (s) hands-on (n) theoretical										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	83.88%	81.25%	83.33%	88.46%	66.67%	79.55%	84.62%	87.88%	94.44%	89.29%

SN5 You prefer the: (s) traditional (n) novel										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	69.01%	56.25%	50.00%	67.31%	77.78%	71.59%	69.23%	75.76%	66.67%	73.81%

TF1 You prefer: (t) logic (f) empathy										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	86.78%	87.50%	100.00%	86.54%	86.11%	86.36%	92.31%	78.79%	100.00%	83.33%

TF2 You are more: (t) truthful (f) tactful										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	73.14%	75.00%	50.00%	71.15%	55.56%	64.77%	80.77%	90.91%	**50%**	82.14%

TF3 You see yourself as more: (t) questioning (f) accommodating										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	71.49%	87.50%	**16.67%**	73.08%	80.56%	76.14%	50.00%	74.24%	77.78%	75.00%

TF4 You are more: (t) skeptical (f) tolerant										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	67.36%	68.75%	100.00%	59.62%	75.00%	65.91%	61.54%	**60.61%**	83.33%	**65.48%**

TF5 You think judges should be: (t) impartial (f) merciful										
Time Period	Overall	G12	SP13	G13	FA13	FA13 All	SP14	G14	FA14	FA14 All
Consistency Percentage	73.14%	65.63%	58.33%	69.23%	80.56%	73.86%	76.92%	69.70%	100.00%	76.19%

Appendix D:

Outliers and Lowest Scores

In appendix D, the procedure of testing the lowest and outliers among items in each time period is presented. The item with lowest consistency percentage and outliers which falls out of the acceptable range is bold and counted. Item with most bold is considered with lowest consistency over time.

Dimension EI and JP

	EI1	EI2	EI3	EI4	EI5	JP1	JP2	JP3	JP4	JP5
Overall	77.27	75.21	67.36	70.66	75.21	**64.05**	76.03	80.99	70.66	71.07
G12	65.63	81.25	**53.13**	62.50	62.5	75.00	87.50	87.50	81.25	78.13
SP13	83.33	83.33	66.67	66.67	83.33	83.33	41.67	83.33	33.33	66.67
G13	92.31	75.00	84.62	75.00	88.46	**57.69**	76.92	78.85	67.31	63.46
FA13	80.55	75.00	63.89	72.22	77.78	77.78	80.56	77.78	69.44	61.11
FA13 All	87.50	75.00	76.14	73.86	84.09	65.91	78.41	78.41	68.18	**62.50**
SP14	**46.15**	61.54	65.38	80.77	61.54	76.92	69.23	88.46	65.38	84.62
G14	74.24	78.79	**65.15**	69.70	72.73	**46.97**	72.73	78.79	72.73	77.27
FA14	100.00	66.67	61.11	61.11	77.78	66.67	88.89	77.78	88.89	61.11
FA14 All	79.76	76.19	**64.29**	67.86	73.81	**51.19**	76.19	78.57	76.19	73.81

Dimension SN and TF

	SN1	SN2	SN3	SN4	SN5	TF1	TF2	TF3	TF4	TF5
Overall	75.21	66.12	66.53	83.88	69.01	86.78	73.14	71.49	67.36	73.14
G12	62.50	62.50	56.25	81.25	56.25	87.50	75.00	87.50	68.75	65.63
SP13	100.00	66.67	50.00	83.33	50.00	100.00	50.00	**16.67**	100.00	58.33
G13	75.00	84.62	73.08	88.46	67.31	86.54	71.15	73.08	59.62	69.23
FA13	63.89	**50.00**	58.33	66.67	77.78	86.11	55.56	80.56	75.00	80.56
FA13 All	70.45	70.45	67.05	79.55	71.59	86.36	64.77	76.14	65.91	73.86
SP14	92.31	**46.15**	80.77	84.62	69.23	92.31	80.77	50.00	61.54	76.92
G14	77.27	72.73	72.73	87.88	75.76	78.79	90.91	74.24	**60.61**	69.70
FA14	72.22	55.56	**50.00**	94.44	66.67	100.00	**50.00**	77.78	83.33	100.00
FA14 All	76.19	69.05	67.86	89.29	73.81	83.33	82.14	75.00	**65.48**	76.19

Appendix E:

ANOVA and pairwise test result

In order to figure out if consistency has a systematic change over time, ANOVA and pairwise test are applied to the consistency data for each time dimension and each group. If the p-value generated by ANOVA is less than 0.05, it can be determined that there is at least one group which is significantly different from other groups. Pairwise test is applied afterwards to determine the specific group with significant difference. For pairwise test, the significant level for p-value is 0.05/number of pairwise test in the same ANOVA group.

ANOVA Result:

	EI	JP	SN	TF
Overall	0.021 (0.65 if delete 2013 fall)	0.76	0.21 (0.29 if delete 2014 spring)	0.34
Graduate Student	0.046	0.35	0.08	0.98
Undergraduate Student	0.029 (0.27 if delete 2013 fall)	0.95	0.13 (0.35 if delete 2014 spring)	0.14 (0.77 if delete 2013 spring)

Pairwise test result:

Dimension EI:

	Graduate 2013 Fall	Graduate 2012 Fall
Graduate 2014 Fall	0.011193	0.132035
Graduate 2013 Fall		0.001246

	Undergraduate 2014 Spring	Undergraduate 2013 Fall	Undergraduate 2013 Spring	Undergraduate 2012 Fall
Undergraduate 2014 Fall	0.28	**0.056**	0.18	0.49
Undergraduate 2014 Spring		0.15	0.064	0.26
Undergraduate 2013 Fall			**0.0058**	**0.046**
Undergraduate				0.19

2013 Spring				

Dimension SN

	Graduate 2013 Fall	Graduate 2012 Fall
Graduate 2014 Fall	0.48	0.30
Graduate 2013 Fall		0.33

	Undergraduate 2014 Spring	Undergraduate 2013 Fall	Undergraduate 2013 Spring	Undergraduate 2012 Fall
Undergraduate 2014 Fall	**0.00014**	0.34	0.23	0.19
Undergraduate 2014 Spring		**0.000011**	**0.0064**	**0.0022**
Undergraduate 2013 Fall			0.13	0.082
Undergraduate 2013 Spring				0.483

Appendix F:

Consistency Data between Different Countries

To determine if culture and native language will affect the consistency or not, graduate student from PACE global course is sorted by their nationality. The absolute difference for each dimension and consistency percentage is listed in the appendix F.

China

EI Difference	JP Difference	SN Difference	TF Difference	EI % Consistency	JP % Consistency	SN % Consistency	TF % Consistency
2	4	0	0	80	60	100	100
4	2	4	0	60	80	60	100
2	2	0	0	80	80	100	100
6	0	2	2	40	100	80	80
4	2	4	2	60	80	60	80
2	3	2	2	80	70	80	80
2	0	2	2	80	100	80	80
2	4	0	2	80	60	100	80
4	6	4	0	60	40	60	100
2	1	4	2	80	90	60	80
2	4	3	1	80	60	70	90
2	0	0	4	80	100	100	60
0	4	3	0	100	60	70	100
0	2	4	1	100	80	60	90
2	4	1	3	80	60	90	70
2	4	4	0	80	60	60	100
0	2	2	0	100	80	80	100
0	1	0	4	100	90	100	60
2	0	0	0	80	100	100	100
2	2	4	2	80	80	60	80
2	6	4	2	80	40	60	80
4	0	0	2	60	100	100	80
2	0	2	2	80	100	80	80
1	2	0	4	90	80	100	60
0	2	2	0	100	80	80	100
4	4	2	4	60	60	80	60

2	2	0	2	80	80	100	80
4	6	6	2	60	40	40	80
0	2	2	0	100	80	80	100
0	4	4	4	100	60	60	60

German

EI Difference	JP Difference	SN Difference	TF Difference	EI % Consistency	JP % Consistency	SN % Consistency	TF % Consistency
2	0	2	0	80	100	80	100
1	6	6	0	90	40	40	100
0	0	2	1	100	100	80	90
2	0	2	2	80	100	80	80
0	2	0	0	100	80	100	100
2	4	0	2	80	60	100	80
8	2	0	0	20	80	100	100
2	0	0	0	80	100	100	100
2	0	0	2	80	100	100	80
2	2	2	3	80	80	80	70
2	2	2	2	80	80	80	80

Mexico

EI Difference	JP Difference	SN Difference	TF Difference	EI % Consistency	JP % Consistency	SN % Consistency	TF % Consistency
2	0	2	2	80	100	80	80
2	2	2	6	80	80	80	40
0	0	3	2	100	100	70	80
0	0	4	2	100	100	60	80
2	0	2	1	80	100	80	90
0	2	0	2	100	80	100	80
2	0	0	1	80	100	100	90
2	2	2	0	80	80	80	100
0	2	0	6	100	80	100	40
0	0	4	6	100	100	60	40
2	1	0	0	80	90	100	100
4	0	0	0	60	100	100	100
0	2	2	0	100	80	80	100
0	2	0	2	100	80	100	80

United States

EI Difference	JP Difference	SN Difference	TF Difference	EI % Consistency	JP % Consistency	SN % Consistency	TF % Consistency
0	0	2	1	100	100	80	90
2	3	4	4	80	70	60	60
0	0	0	0	100	100	100	100
4	4	4	2	60	60	60	80
6	0	7	4	40	100	30	60
1	4	0	2	90	60	100	80
2	1	3	2	80	90	70	80
2	2	0	2	80	80	100	80
0	2	0	2	100	80	100	80
2	0	0	2	80	100	100	80
1	4	2	2	90	60	80	80
4	0	2	4	60	100	80	60
2	0	0	2	80	100	100	80
2	2	4	2	80	80	60	80
2	4	0	5	80	60	100	50
0	2	1	1	100	80	90	90