EXPLORING ITEM RESPONSE THEORY IN FORCED CHOICE PSYCHOMETRICS FOR CONSTRUCT AND TRAIT INTERPRETATION IN A CROSS-CULTURAL CONTEXT

BY

TENG-WEI HUANG

Submitted in accordance with the requirements for

the degree of

MASTER OF ARTS

in the subject

PSYCHOLOGY

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: DR HC JANEKE

MARCH 2011

DECLARATION

Student no. 35943726

I declare that ‘Exploring item response theory in forced choice psychometrics for construct and trait interpretation in a cross-cultural context’ is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

SIGNATURE DATE

(MR T HUANG)

ACKNOWLEDGEMENTS

Firstly, special thanks to my supervisor Dr Chris Janeke – thank you for your support and guidance.

Secondly, I would like to thank the Thomas International South Africa for all the research resource and support.

Thirdly, I would like to thank my family for support and understanding - this thesis would not have been possible without all your support. Thank you.

LIST OF ABBREVIATIONS

3PL-IRT	Three-parameter Logistic Item Response Theory model
Amend A	Amendment of 16 items after Part I research
C	Compliance construct (or overall score in Self mask)
CI	Compliance construct in Work mask
CII	Compliance construct in Pressure mask
CPE	Correlational Parameter Estimation Method
CTT	Classical Test Theory
D	Dominance construct (or overall score in Self mask)
DI	Dominance construct in Work mask
DII	Dominance construct in Pressure mask
DISC	Marston’s DISC theory (Dominance, Interactive, Submission, and Conformity)
ERB	Extreme Response Bias
FC	Forced choice
FCMCQ	Forced choice to Multiple Choice Question (modified model)
GRM	Samejima’s General (graded) Response Model
Hi	High scoring. PPA scoring method - items only scored when marked ‘high’
I	Interactive construct (or overall score in Self mask)
ICC	Item Characteristic Curve
II	Interactive construct in Work mask
IIC	Item Information Curve graph
III	Interactive construct in Pressure mask
Index (D)	Discrimination index
Index (P)	Difficulty index
Index (PI)	Preference index
IRCCC	Item Response Categories Characteristic Curve

IRCCC-s	Item Response Categories Characteristic Curve summary graph
IRT	Item Response Theory
IRTCI	IRT Construct Interpretation Method
ITCC	Item Total Correlation Co-efficient (construct interpretation method)
KTB	Kendal’s Tau B Ordinal Correlation analysis (modified)
Lo	Low scoring. PPA scoring method - items are only scored when marked ‘least’
Part I	Research Part I (Beijing sample, n=650) are collected via old Chinese PPA form
Part II	Research Part II (Beijing sample, n=307) are collected via New Chinese PPA form with Amend A
PPA	Personal Profile Analysis
RICC	Raw Item Characteristic Curve
S	Submission construct (or overall score in Self mask)
SDR	Social Desirable Response
SI	Submission construct in Work mask
SII	Submission construct in Pressure mask
TIF	Test Information Function graph

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This thesis explores item response theory (IRT) in the Personal Profile Analysis (PPA) from Thomas International. The study contains two parts (Part 1 and Part II) for which two sample groups were collected. For Part I of the research 650 participants were collected via the old form (CPPA25/C7) in the Beijing office of Thomas International in China (male=323, Female=267, missing=60). Part II of the research used the amended form in the same area and collected a sample of 307 (male=185, female=119, missing=3).

The study postulates that IRT methods are applicable to forced-choice psychometrics. The results of Part I showed that the current CPPA form functions, to some extent, according to PPA's original constructs. Part I of the research identified 16 items that need to be amended (called Amend A in this research). The amended form was returned to China for the collection of samples for Part II, and the results are deemed acceptable.

The study concludes with a research protocol for PPA-IRT research generated from the current research. The research protocol suggests four levels of analysis for forced choice (FC) psychometrics, namely: 1. Textual analysis, 2. Functional analysis, 3. Dynamic analysis, and 4. Construct analysis.

Key terms

Item Response Theory, Classical Test theory, forced choice, psychometric theory; General (graded) Response Model, Personal Profile Analysis (PPA), cross-culture research, Chinese, translation, psychometric adaptation, research protocol.

CHAPTER 1. INTRODUCTION

1.1. Motivation

1.1.1 Puzzle of response bias in psychometric practice

Research on the use of personality instruments for employee selection has expanded since the early 1990s when meta-analytic reviews demonstrated their use for predicting work behaviours (Barrick & Mount, 1991). Industrial psychometric tests are commonly employed as an aid for decision making in the South African market (Van De Vijver & Rothmann, 2004; Van der Merwe, 2002). Psychometric testing is used in occupational decisions, including selection and classification of human resources.

From the blue-collar level, such as assembly line operators and drivers, to white-collar levels such as sales representatives, administrative personnel, and top management, there is scarcely a job for which some kind of psychometric test has not proved helpful in labour related matters. In the South African labour market, psychometrics play an important role in selection, job assignment, transfer, promotion, and retrenchment (Foxcroft & Roodt, 2007). Therefore, it is not surprising that respondents are highly motivated to achieve 'excellence' in psychometric performance. Past research indicated that such eagerness to 'do well' in a psychometric test could lead to various types of response biases (Brown & Harvey, 2003; Crowne, 1960; Dicken, 1963; Griffith, Chemielowski, & Yoshita, 2007).

Studying response biases is important for the validation and interpretation of personality and attitude measures by self-report. The most commonly observed response styles, or response biases, are 'acquiescence' and 'extreme response' bias (ERB) (Meisenberg & Williams, 2008). Paulhus (2003) has suggested a comprehensive model for motivation of response biases, under the umbrella term 'socially desirable responding' (SDR).

Researchers have investigated SDR for more than half a century (Bernreuter, 1933; Paulhus, 1991, 2003; Paulhus, Fridhandler, & Hayes, 1997). More than 50 years ago, the negative impact of SDR on psychometric validity had already been reported (Bernreuter, 1933; Vernon, 1934). The SDR effect has been a particular focus in areas such as personality measurements (Gough, 1947; Lönnqvist, 2008; McKinley, Hathaway, & Meehl, 1948), industrial research (Levashina & Campion, 2007), academic performance (Hirsh & Peterson, 2008), social attitudes (Rachlin, 2002), and sensitive (socially-unacceptable) behaviour measurements (Meisenberg & Williams, 2008; Rouse, Kozel, & Richards, 1985).

1.1.2 The complexity of SDR

It has been established by researchers that respondents of self-report measures tend to manipulate the results of such instruments by providing what they deem to be socially acceptable responses (Griffith et al., 2007; Komar, Brown, Komar & Robie, 2008; Morgeson, Campion, Dipboye & Hollenbeck, 2007; Paulhus, 1981; Paulhus et al., 1997; Rees & Metcalfe, 2003; Vernon, 1934). According to Cheung and Rensvold (2000), common expressions of such response styles are ‘extreme’ and ‘acquiescence’.

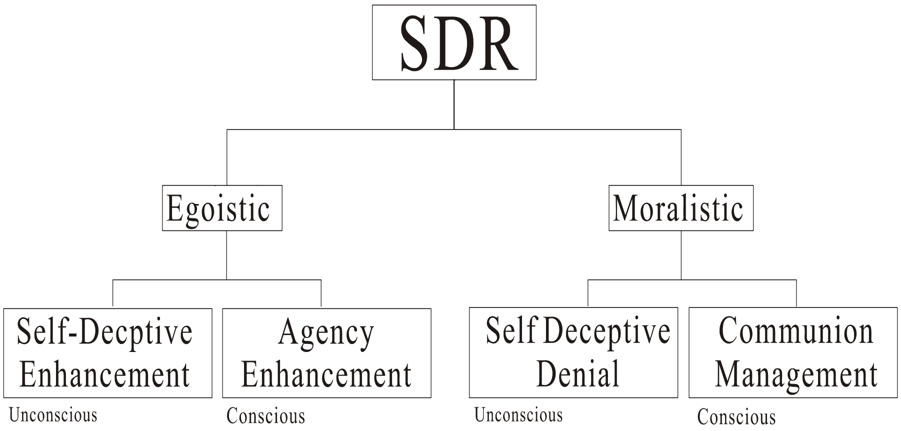


Figure 1.1 Socially Desired Response constructs (Paulhus, 2002, 2003)

Extreme response style (ERS) is a preferential selection of the end points of a scale (Meisenberg & Williams, 2008). **Acquiescence response style (ARS)** refers to agreement with a statement (Cheung & Rensvold, 2000). Early theorists used a two-component model to classify the motivation behind response biases, namely self deception at an unconscious level, and conscious deception in which a respondent knowingly tries to deceive others (Sakeim & Cur, 1978).

This model was further developed into a four-factor model by Paulhus (2002). *Egoistic bias* is a self-deceptive tendency to exaggerate one's social and intellectual status. *Moralistic bias* is a self-deceptive tendency to deny socially deviant impulses and claim sanctimonious status (Paulhus, 2002). Egoistic and moralistic bias can be further subdivided into conscious and unconscious dimensions. Egoistic bias can be separated into self-deceptive enhancement and agency enhancement. Moralistic bias can be separated into self-deceptive denial and communal management.

In terms of its theoretical aspect, a valid psychometric instrument should have the ability to counter, or reduce, all four types of bias. However, in practice, these biases are still commonly observed in industrial assessment, and are not directly confronted. It seems that most psychometric instruments, especially the Likert scale, do not fulfil their validity criterion (Brown & Harvey, 2003; Day & Carroll, 2008; Griffith et al., 2007; Komar et al., 2008; Morgeson et al., 2007; Rees & Metcalfe, 2003). Moreover, differences in response styles have been documented in cross cultural applications of assessment instruments (Billiet & Davidov, 2008; Cheung & Rensvold, 2000). This makes the already difficult measurement issue even more complex when it occurs in a culturally diverse setting such as in South Africa.

1.1.3 Forced-choice focus of research

Many methods of coping with SDR have been suggested by past researchers (Nederhof, 1985; Paulhus, 1981). Coping methods include rational techniques, factor analytic techniques, covariate techniques, and demand reduction techniques (Paulhus, 1981, 2003). In the family of rational techniques, forced choice (FC) is one of the common methods used.

The FC psychometric method first appeared around the 1930s and the concept was initially developed by clinical measurement researchers such as the Humm-Wadesworth (Humm, 1939a) and Horst-Wherry teams (as cited by Travers, 1951; Zavala, 1965). It received recognition after the 54th APA annual meeting in 1946 (Staff, 1946). FC reached the height of its popularity between 1950 and 1960 with a cooling off during 1970s (Hicks, 1970). FC items are among the earliest developed methods of coping with a social desirability bias by providing more than one socially equal preferable item that prevents candidates from 'spotting' the 'right answer' among several 'right answers'.

Example 1: PPA FC item (English item set 4, Irvine, 2003)

□	Open Mind	□	Obliging	□	Will power	□	Cheerful

Instruction: In each line select the word that **most** describes you in the work situation and place an M in the box to the right of that word. Choose a word from each line that least describes you in the work situation and place an L in the box to the right of that word.

In example 01 above, the four items can be considered as socially preferable by most of respondents although the item set does not give any hints of a social situation. It only indicates that the respondent should answer this question as relating to a work situation, and as such, all four items could be 'right answers'. This format would prevent candidates from answering in accordance with socially preferred patterns.

Example 2: Self-report Likert format. (Trait Emotional Intelligence, Petrides, 2009)

No TEIQue items

2. Generally, I do not take any notice of other people's emotions.

25. I believe I have many personal weaknesses.

36. I normally find it difficult to keep myself motivated.

61. I would describe myself as a calm person.

85. I can handle most difficulties in my life in a cool and composed manner.

88. I believe I have many personal strengths.

Instruction: Please answer each statement by circling the number that best reflects your degree of agreement or disagreement with that statement. There is no right or wrong answer. You have seven possible responses, ranging from 1 = Completely Disagree, to 7 = Completely Agree.

In example 02, it is easy to ‘guess’ what the ‘intended’ responses are from the self-report Likert scales. It is also easy to observe/guess that items 2, 25, and 36 are negative items and that they are likely to elicit negative responses. In contrast, 61, 85, and 88 are positive items likely to elicit positive responses. Furthermore, various culture groups that tend to give extremes (i.e. select 1 or 7) and acquiesce responses (i.e. selecting 3~5) could easily select extreme or moderate responses in this format.

In contrast to the FC scale (see example 01), alignment of equally social preferable options into an item set would make the extreme response, acquiesce, and socially desirable response impossible because no such options are provided in the format (see Example 01) (Edwards, 1957, 1970; Ford, 1964; Goldman, 1964; Humm & Wadsworth, 1939).

It is within the framework of this research that the present study was conducted. Its main aim was to explore new approaches for FC via contemporary item response theory algorithms and to establish a system of research protocols that further increase the practical applicability of FC measurements and indirectly reduce all forms of SDR (Brown & Harvey, 2003; McCloy, Heggestad & Reeve, 2005).

1.1.4 Using IRT in FC

Although previous theorists have suggested that the forced choice approach has the ability to decrease SDR (Edwards, 1957, 1970; Ford, 1964; Goldman, 1964; Humm &

Wadsworth, 1939; Nederhof, 1985; Zavala, 1965), FC methods are not applied, due mostly to the added complexity they cause for test construction, calculation, and interpretation (Brown & Harvey, 2003; Martinussen, Richardsen & Varum, 2001; Nederhof, 1985).

Theorists tend to be critical of FC methods because they often preclude the use of classical statistical methods such as Cronbach's Alpha for internal consistency and factor analysis for construct validity (Bartram, 2007; Cornwell & Dunlap, 1994; Martinussen et al., 2001; Tenopyr, 1988).

Inspired by past research studies, attention has shifted recently towards Item Response Theory (IRT) (Bartram, 2007; Brown & Harvey, 2003). IRT emerged during the 1940s, at which time true score Classical Test Theory (CTT) was at its height of popularity. IRT originated from debates and discussions regarding the various shortcomings of CTT and from the postulation of a new model that could fix most of the CTT mistakes. Since the 1940s the IRT model has evolved from dealing with simple true-false cognitive tests to multiple–choice tests, the Likert scale, and timed tests (speed and accuracy). It has also been adapted for use in various computerised item-banks and testing systems. The IRT theory is characterised by the construction of item characteristic curves (ICC), higher requirements on statistical training, and a large homogenous sample size (n>250~500) (Jooste, 2003).

It is necessary to understand the fundamental applicability of IRT in FC psychometrics. Consequently, this study focuses mostly on exploring such a possibility with the intention of creating a system, or protocol, of research for future researchers.

1.2. Research in Chinese language group in mainland China

The aim of this research is to run IRT with FC. However, the entry requirement for running IRT (GRM model) would require a large and homogenous sample. According to past research, such samples (n>500) are necessary to conduct an IRT analysis (Hambleton & Swaminathan, 1985; Jooste, 2003; McKee, Klein & Teller, 1985).

The homogeneous samples need to be relatively similar in terms of language, education, culture, literacy level, and ethnicity.

However, it is very difficult to obtain such compositions in the current South African sampling frame, due to the fact that the South African sample contains too many different languages (11 languages) and cultures. Furthermore, the literacy/education level of the majority of the population would not be suitable for conducting psychometric research. If limited to individuals with a higher education, the sampling would be heterogeneous, and the South African sample would therefore not be suitable for conducting this study.

Fortunately Thomas International's branch in China had approached the researcher for assistance in evaluation of the Chinese PPA form. This is a large-size sample comprising one main language group. It appeared suitable for the intended research, and the main sample of this study is therefore based on a set of Chinese respondents.

1.3. Aim of current study

This research is mainly an explorative study. The research question would be. ***'Can item response theory (IRT) be used in forced-choice psychometric (FC) adaptation?'*** The final product of this research should be a system of IRT methods or protocols for FC that can be used in future research. The following section explains briefly the areas to be explored, and the reasons for exploring them.

For this protocol to be truly applicable, the research should first explore the IRT parameter estimation methods. The IRT parameters would be the three basic numbers that would help to define the ICC. This IRT curve would be used to abstract the probability of an item that had been used against differing levels of ability (from low to high). In IRT, it is assumed that if an item demonstrates a positive relationship with a target construct, the item 'belongs' to the construct. By creating an adequate parameter estimation method, one is able to adapt and interpret item constructs via response patterns. Therefore, one of the major aims of this protocol is to establish the most applicable *IRT parameter estimation method for FC*.

The other area of interest is the *item construct estimation*. The latter would normally be conducted through correlational analysis in a common Likert scale study. Although, due to the complexity of FC internal dynamics, Pearson product-movement correlational statistics would not be suitable. Hence it is important to investigate alternative methods for replacing Pearson-correlational-base statistics. The item construct estimation method should be able to indicate the relationship between target items and construct. This would be beneficial for interpreting an item's construct nature from the responses of the participants. This interpretation would improve an item's construction and modification through different cultural and time periods. Based on these considerations, this study explores the *possible construct estimation methods* suitable for FC study.

The third area of focus is *cross-cultural applicability.* The practice of using translational psychological or educational testing instruments has become common practice in the contemporary international market (Sireci, Yang, Harter & Ehrlich, 2006; Yu, Lee & Woo, 2004). This study aims to create a research protocol that embeds this consideration. This method should help test-creators and researchers to modify their tools for use in different cultural settings.

This protocol should therefore cover three areas: *parameter estimation method*, *item construct estimation*, and *cross-cultural applicability*. The protocol would need to cover these three areas to reach the two levels of equivalence of adaptation; i.e. textual and constructual. The three areas are explained in more detail in the following sections. This study aims to contribute to the incremental improvement of psychometric accuracy and quality in industrial assessment. The areas and reasons for exploring them are introduced briefly in the following sections. The various strengths of IRT are also introduced across three areas, these being the main reasons for selecting the IRT method as the focus method for this research.

1.3.1 Parameter estimation method in forced choice questions

The term '*parameter*' here is used to describe the IRT parameters. Three parameters are *difficulty parameter* (b), *discrimination parameter* (a), and *guessing parameter* (c) (see Figure 1.2).

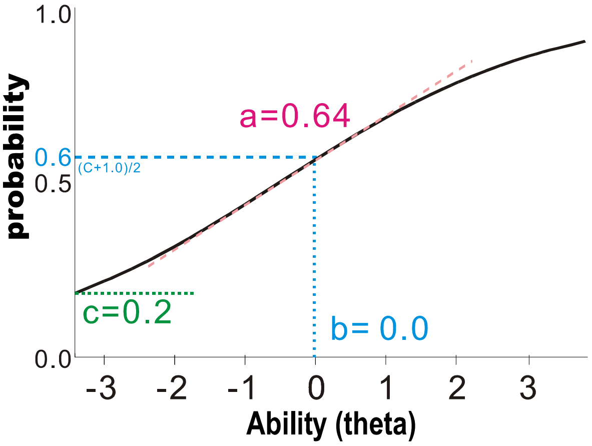


Figure 1.2 The illustrations of three parameters of item response theory

Note: a=0.64 represents low discrimination. b=0.00 represents current item difficulty is relatively neutral; c=0.2 represents at least a 20% of chance of guessing this item correctly.

The IRT uses these parameters to define the shape of the item characteristic curve (ICC). An effective psychometric instrument, the ICC is used to summarise the probability of an item being used across different levels of the target construct (theta). A successful ICC would look like Figure 1.2, with a higher right side and a lower left side. This implies that the higher the ability, the more probability an item has of being marked as correct.

Figure 1.2 illustrates the difficulty (b) of this item as 0.00 (the ability across (C+1.0)/2 probabilities), which means the difficulty of this item is fairly neutral. The discrimination (a) as 0.64 (slope), this indicates that the item's discrimination of the candidate is relatively low.

This implies that one cannot discriminate between different candidates' ability easily by using the current item. The guessing probability (c) is 0.2, this indicates the item has a 20% chance of being marked correctly, even with very low ability (Hambleton & Swaminathan, 1985).

The concepts of three parameters were not developed concurrently. It started from the one parameter logistic (1PL), two parameter logistic (2PL), to three parameter logistic (3PL) and further (Hambleton & Swaminathan, 1985). The one-parameter logistic (1PL) model would use the difficulty parameter (b) only. (see Figure 1.3)

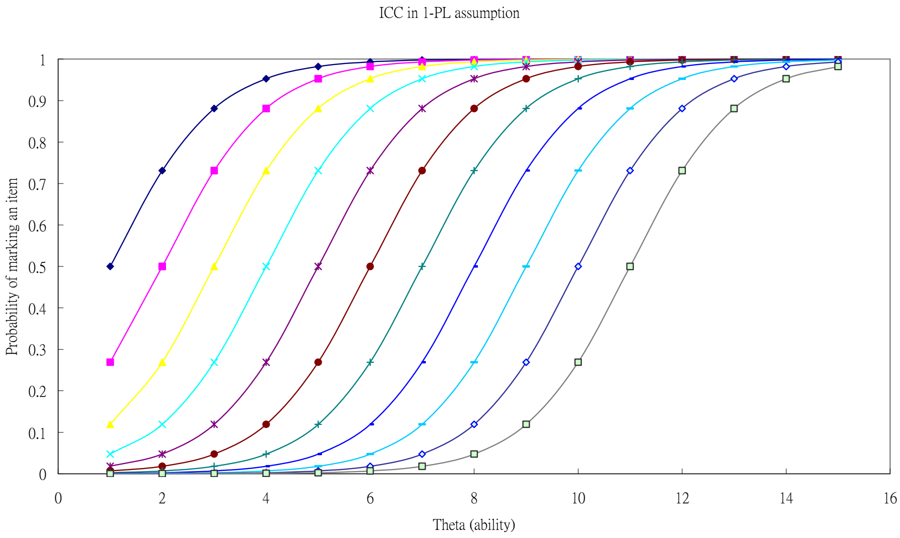


Figure 1.3 1PL: One parameter logistic model, b (difficulty parameter)

Note: b (difficulty) parameter is the only parameter that is used in 1PL model. The above illustration shows items with different difficulty parameters (b). From the left to right are the easiest to the most difficult items.

The two parameter logistic (2PL) model uses both difficulty (b) and introduces discrimination parameter (a) (see Figures 1.2 and 1.4).

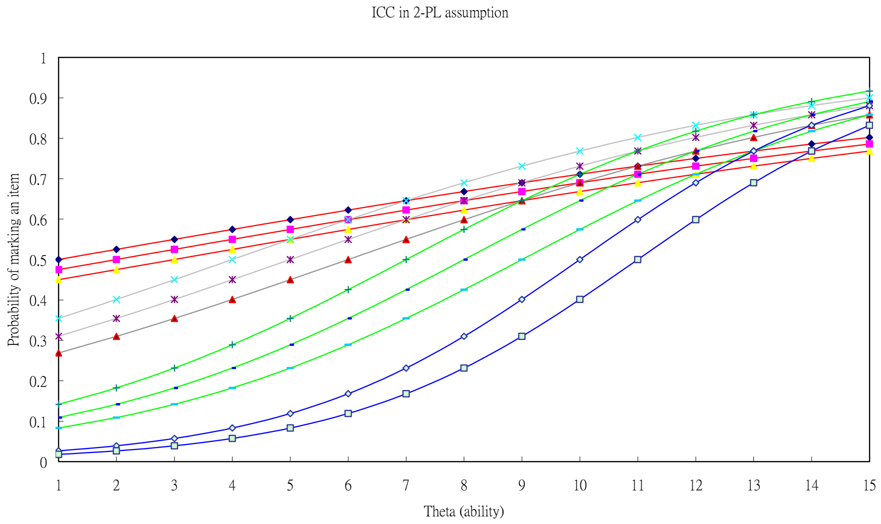


Figure 1.4 2PL. Two parameter logistic model, b (difficulty parameter) and a (discrimination parameter)

Note: This figure indicates four (the red, grey, green, and blue) sets of item with same discrimination parameter (a), but different difficulty parameter (b). Within the red set, items have the same discrimination parameter (same slope) but a different difficulty would formulate parallel curves.

The three-parameter logistic (3PL) model uses both difficulty (b) and discrimination (a), and introduces the new concept (c), the guessing parameter (see Figures 1.2 and 1.5).

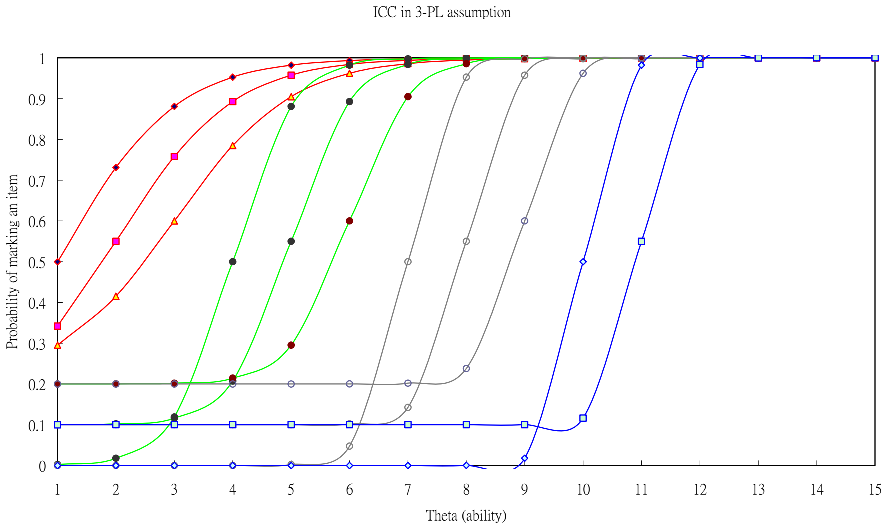


Figure 1.5 2PL: Two parameter logistic model, b (difficulty parameter), a (discrimination parameter), and c (guessing parameter)

Note: This figure indicates four (the red, green, grey, and blue) sets of item with the same discrimination parameter (a), different difficulty parameter (b), and different guessing parameter (c).

The *'parameter estimation method'* implies methods to calculate three parameters from raw FC data. This is illustrated by Figure 1.6. The real probability of an item being marked positive is indicated by the blue dots. The blue dots in Figure 1.6 indicate that as the construct increases in strength, the probability of the item being marked positively is increasing. The pink curve indicates a theoretical abstraction of the actual data. The curve is normally plotted via application of three parameters. The method of calculating parameters from the raw data is called a 'parameter estimation method'.

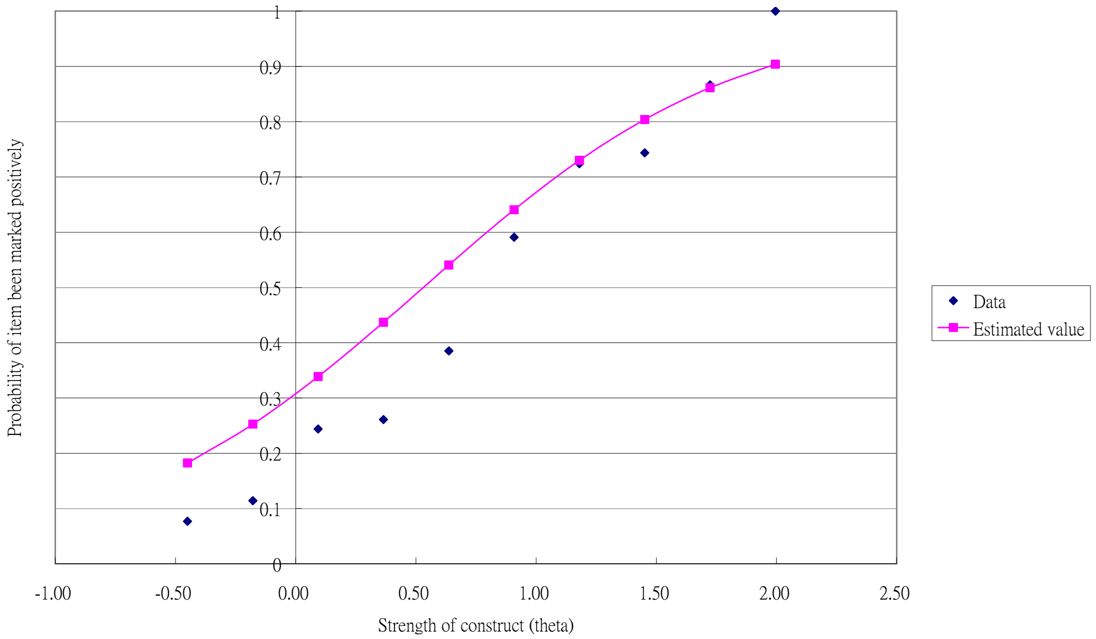


Figure 1.6 Sample of parameter estimation, three-parameter logistic model.

Note: The blue dots represent the actual data, the pink line represents the modelled IRTICC curve. This is an example of a positive ICC.

This topic is important because studying parameter estimation methods would increase the application to forced choice (FC), and indirectly benefit the psychometric society. The current research postulates that the nearest parameter estimation model for PPA FC system would be Samejima's (1999) general (grade) response model (GRM) as the GRM is an IRT model designed for Likert scale-type responses (see Figure 1.7). If a study treats every single PPA item within an item set as single response, and all sub categories as ordinal responses, the PPA items can be interpreted as Likert scale items. This study examines the correctness of this postulate (see Example 3 below).

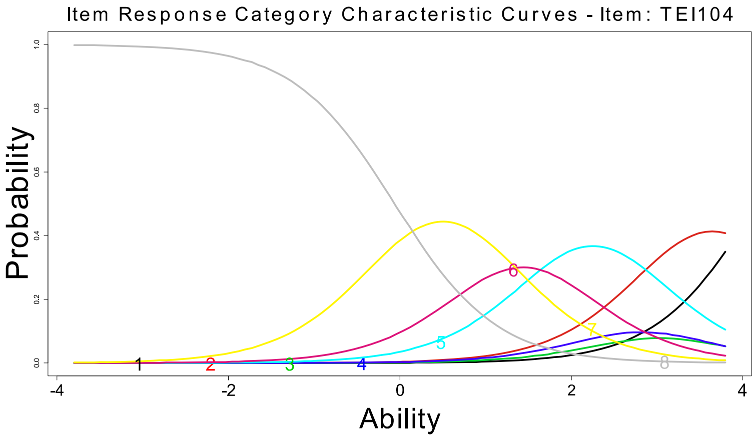


Figure 1.7 The General Grade Response model (GRM) illustration

Note: The general/grade response model (GRM) enables the researcher to explore the relationship under each sub-category of an ordinal item. The above figure indicates the category characteristic curve of the eight categories (1~8) of a Likert scale. The GRM method would help the researcher to define the true distance between each sub category of a Likert scale.

Example 3: PPA FC item (English item set 4, Irvine, 2003)

M	Open Mind	□	Obliging	□	Will power	L	Cheerful
			||				
1	Open Mind	2	Obliging	2	Will power	3	Cheerful

Note: the forced choice item can be converted into four separate Likert scale items.

This study explores and compares various models and their benefits. The parameter estimation methods, efficiency, and accuracy will be compared. Methods for parameter estimation will include *Correlational parameter estimation method* (CPE) and GRM models. Methods for rechecking CPE and GRM's validity are *Forced choice to multiple choice Questions* (FCMCQ), Raw Item Characteristic Curve (RICC), and Kendall's Tau-B (τ) (KTB). Details of above methods are discussed in the method chapter.

1.3.2 Item construct estimation (through response pattern)

The term ‘construct’ can be differentiated from the term ‘trait’, according to Leovinger (1957/1967). *‘Traits exist in people; Constructs exist in minds and magazines of psychologists’* (as cited in Braun, Jackson & Wiley, 2002.4). However, at the application level, psychometric practitioners often forget there is still a fine line between ‘trait’ and ‘construct’. Therefore a method of research needs to be established to constantly reconfirm and re-examine the relationship between actual traits and psychometric constructs. This aspect is explored in this study.

In the classical test history, a construct is normally created via Pearson product-movement correlation statistics, such as factor analysis, correlation coefficient, and reliability. A good example of this approach is Spearman’s general intelligence ‘g’. The idea of different types of intelligence is constructed through correlations of different types of items. Using a data reduction technique such as factor analysis, a construct of ‘general intelligence’ is then inferred.

However, ‘construct’ creation in psychometrics is not limited to the correlation family only. The latent trait model, which is another name for IRT, proposes an alternative method for ‘construct’ creation. The latent trait model can also be used to extract the dominant ‘component’ or ‘factor’ within a group of items (IRT’s *‘uni-dimensionality’* assumption). Furthermore, the latent trait model also has the ability to explore other multiple constructs, which in IRT’s terms is called ‘multi-dimensionality’ (Lord, 1980a).

In terms of IRT, if an item demonstrates a positive relationship with a target construct (or ability) it is more likely that this item will belong to this construct (see Figure 1.6 for an example of a typical positive ICC). By creating an adequate parameter estimation method, one is able to adapt and interpret item constructs via response patterns. Therefore, using IRT in item selection and application would be more accurate due to the fact that the item is generated from the target population’s response pattern.

Construct interpretation with IRT is a new area to be explored. IRT-construct interpretation (IRTCI) is similar to the much more popular ranked correlational technique (Kendall's Tau - B) – and can be used to explore the relationship between a selected item, and a targeted construct. The item-total correlation coefficient (ITCC) measures an item's relation to an assigned construct by analysing the correlation between a single item (i.e. item A) and an assigned construct (item groups excluding item A). ITCC generates a correlation coefficient ranging from -1.0 to +1.0. If an item has a value larger than .3, this would suggest that it has contributed to the assigned construct group (item group) (Allen & Yen, 1979).

Similar to ITCC, IRTCI also compares an item (X) with an assigned construct (item group excluding A). However, the difference between ITCC and IRTCI is that the latter can provide an overview (in probability of positive or successful response across different ability levels) of item A's functionality in different levels of ability (i.e. the assigned construct item group) as shown in Figure 1.8 below.

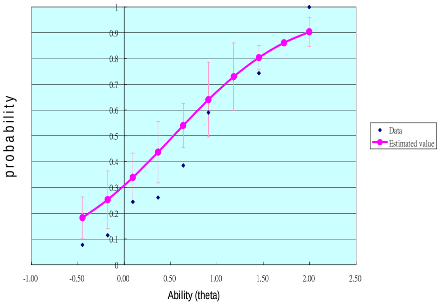
	Item	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
	D_4_01 Original Lo D 做事與眾不同 (old)	0.11	0.78
	D_3_02 Stubborn D 固執 (old)	0.25	0.77
	D_2_03 Bold D 勇敢 (old)	0.46	0.76
	D_3_04 Will power Lo D 有意志力 (old)	0.33	0.77
	D_3_05 Courageous Lo D 有膽量 (old)	0.27	0.77
	D_1_06 Competitive D 喜歡挑戰 (old)	0.43	0.76
	D_3_07 Unconquerable (cannot be beaten) D 好勝 (old)	0.26	0.77
	D_1_08 Brave Hi D 敢與參與 (old)	0.24	0.77
	D_3_09 self-reliant (independent) D 獨立自主 (old)	0.21	0.77
	D_1_10 Adventurous D 喜歡冒險 (old)	0.39	0.76
	D_4_11 Decisive D 遇事果斷 (old)	0.25	0.77
	D_2_12 Daring D 敢做敢爲 (old)	0.44	0.76

Figure 1.8 IRTCI and ITCC comparison

Note: LEFT: IRTCI model, item scoring probability against the target construct. The relationships between an item using a probability of target result (y-axis) versus different strength of the construct (x-axis) are clearly illustrated. RIGHT: ITCC model, the relationship between item and construct are only summarised in terms of correlation strength.

This means, if an item (X) truly functions for this ability/construct/item group, it should have a low probability of being used (or marked as ‘yes’) when an individual has a low ability (construct, or mark rate in this item group).

Vice versa, if a group of candidates marked all other items within this construct, it would be very likely (high probability) that they also marked the item (X) that is associated with a similar construct (see Figure 1.8 for typical Logistic curve).

The probability relationship of an item (X) with different levels of an assigned construct is expressed using a logistic S curve, which indicates the level of agreement between item X and the assigned construct. If there is no relationship, then a flat line of probability will be obtained, which means that the likelihood of the item's utilisation probability has no relationship with the assigned construct (see Figure 1.8, p. 17).

This method opens new possibilities for interpretation. Psychometric researchers interpret an item via a population's response towards that item. In many cases, psychometric developers assume the construct of certain items form the textual interpretation. However, the target population might not interpret an item in accordance with the developer's assumption. In such cases, both ITCC and IRTCI can be applied (see the list of definitions). In ITCC (item-total correlation coefficient), the results would show only whether it 'worked' or 'did not work' in relation to the desired construct. IRTCI (IRT Construct Interpretation) method) would provide more information.

The IRTCI can generate a visual representation (ICC) that explains the item function in much more detail. Other than just 'working' or 'not working', this model explores three phenomena cohesively. An item's 'guessing' (pseudo right) ability, which is the probability of candidate selecting an item without actually 'having' such a construct (parameter c); an item's difficulty (parameter b); and an item's discriminate-ability (parameter a). The CTT also provides all these measurements, such as difficulty index and discrimination index as two unrelated index, not in the cohesive way as does the IRT model. In the IRT model, all parameters are interacting with on another and can be illustrated for better understand (see Figure 1.2). These aspects would help a researcher create a more comprehensive picture of item functioning and could be a very helpful tool in improving the item quality. Exploring this area would help researchers to re-visit the construct and improve the item functionality - all of which contribute to this being a very important area to explore.

1.3.3 International applicability

The practice of using translated psychological or educational testing instruments has become common practice in contemporary international markets. Many psychometrics are either translated or adapted from the dominant culture, while only a small amount of psychometrics are produced locally and validated against all the sub-populations. It is almost guaranteed that in the psychometric practice, practitioners would encounter individuals who cannot fulfil the description of the original research sample, in terms of age group, language group, culture group, historical background, educational background, ethnicity, religion, computer literacy, language literacy, etc. (Sireci et al., 2006; Yu et al., 2004).

Further, when considering the influence of acculturation, globalisation, the Internet and international mass media, the issue of 'culture' is becoming much more complex than the assumption made in the classical test approach; namely that there is only one norm for each language or cultural group. The assumption of external generalisation within classical test theory cannot yield meaningful results because the original sampling population can never be equivalent to the current changing target population.

Guidelines to re-validating a test against the local culture have been provided by many government testing institutions, such as HPCSA, BPS, and APA (HPCSA, 2006*a*, 2006c, 2009). However, guidelines such as HPCSA's psychometric registration procedure would mostly be based on the classical model of 'one norm for each language/culture/age group' and the classical test theory (HPCSA, 2008). This study, therefore, intends to examine the possibility of using IRT theory for validating the quality of translated and adapted testing instruments.

The IRT assumption of 'uni-dimensionality' suggests the possibility of practical application in the cross-cultural setting. This aspect is especially important for dealing with individuals that provide different levels of ability due to culture differences. The 'uni-dimensionality' assumption suggests that item parameters are not dependent on the ability level of the respondents (Baker, 2001).

In other words, parameters retrieved from different ability groups of the same item would be the same; for example, a large census contains two groups of respondents (see Figure 1.9).

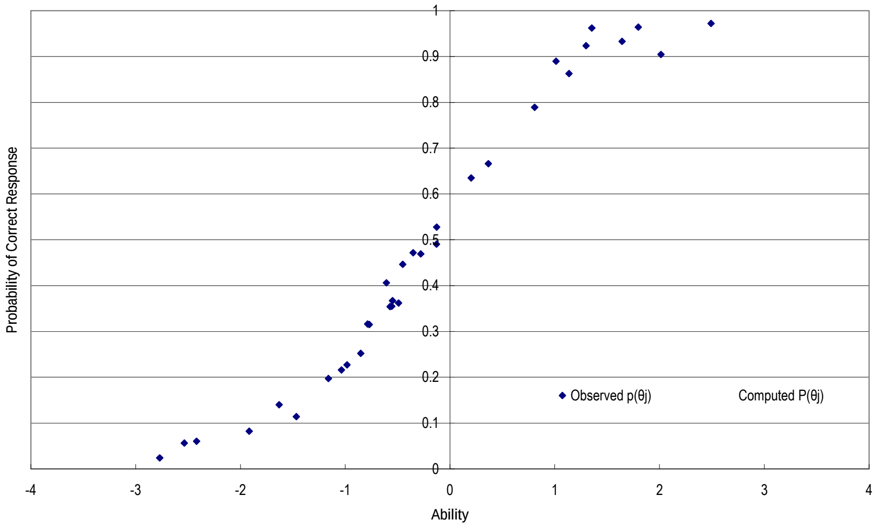


Figure 1.9 Observed proportion of correct response for mixed groups

If one runs this sample group against the 2PL model, two parameters that one would receive would be b=-0.2 (neutral to weak item), a=1.45 (good discrimination). When the two above parameters are plotted into an IRT curve, it would look like Figure 1.14).

If one separates two groups, and runs the same 2PL model, the parameter that one would get from the group with lower ability would be b=-0.2 (neutral to weak item), a=1.45 (good discrimination) (see Figures 1.10 and 1.11).

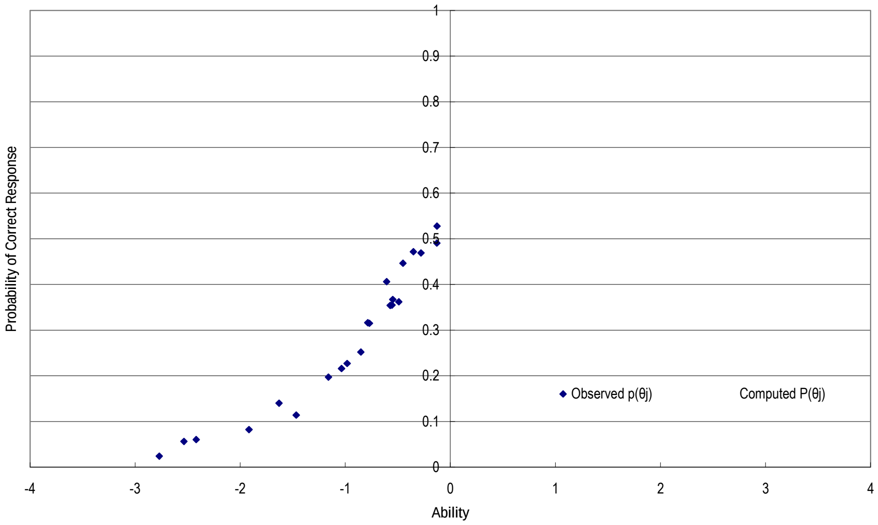


Figure 1.10 Observed proportion of correct response for lower groups

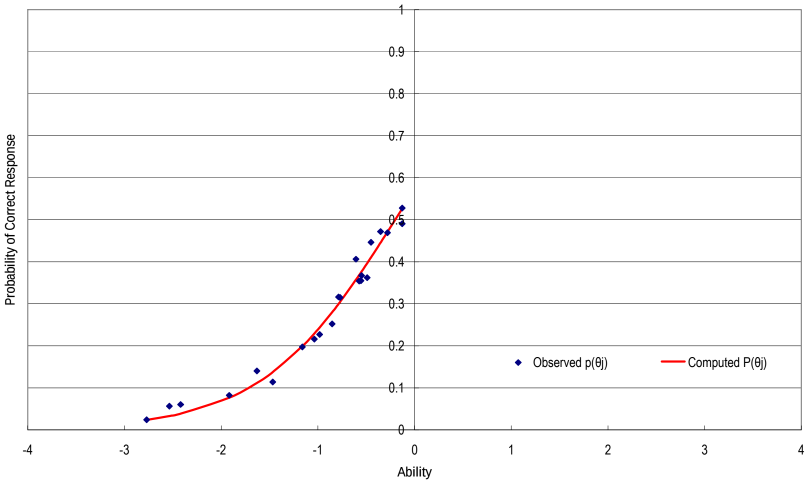


Figure 1.11 Item characteristic fitted for lower groups (b= -0.2, a=1.45)

If the same analysis is conducted in the higher ability group, the parameters would also be b=-0.2 (neutral to weak item), a=1.45 (good discrimination) (see Figures 1.12 and 1.13).

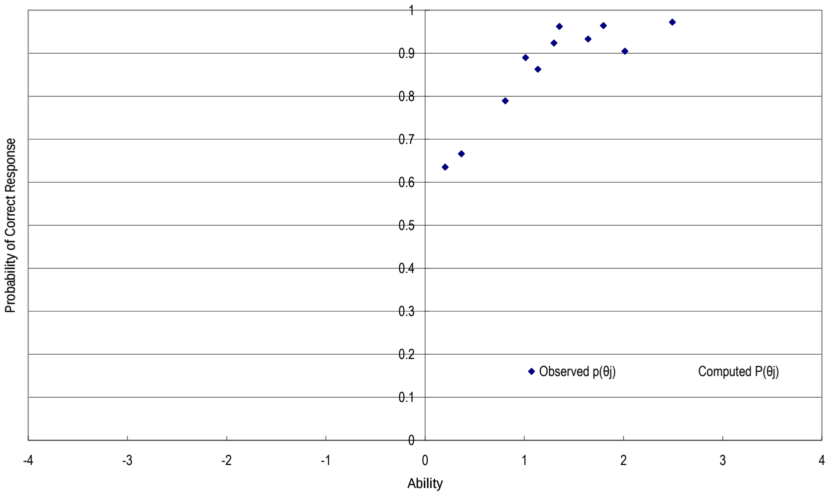


Figure 1.12 Observed proportion of correct response for higher groups

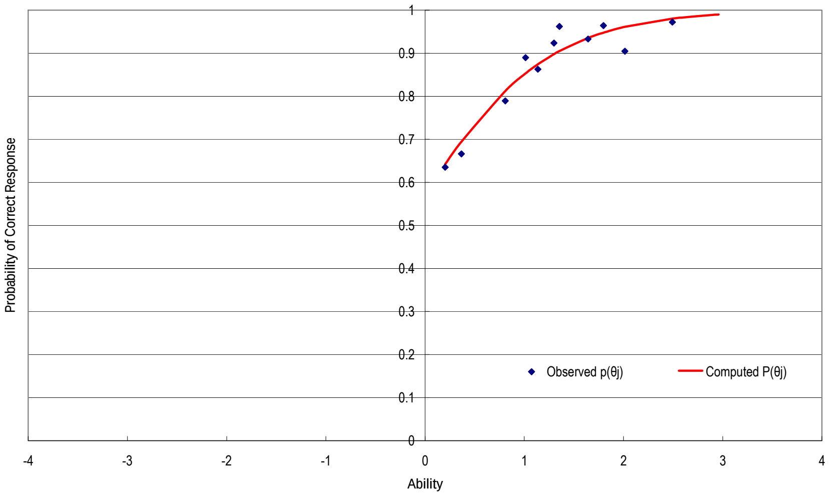


Figure 1.13 Item characteristic fitted for higher groups (b= -0.2, a=1.45)

The IRT theorists therefore suggest that item parameters (from the same item) generated from different samples would be invariant (see Figure 1.14). The individual groups and mixed group would generate the same parameters, b=-0.2 (neutral to weak item), a=1.45 (good discrimination), despite different ability levels. This is the IRT assumption of ‘uni-dimensionality’.

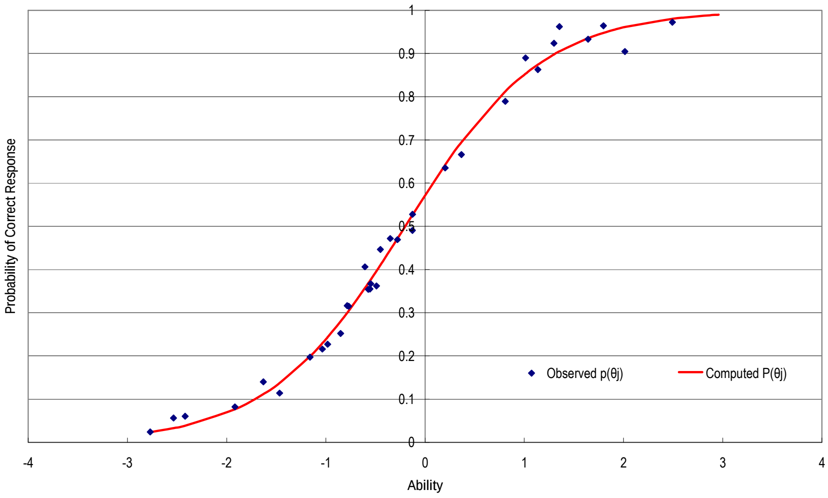


Figure 1.14 Item characteristic fitted for mixing high and low groups (b= -0.2, a=1.45)

The implication of the uni-dimensionality assumption in international research is: if the item is measuring one construct only in different groups, the item parameters would be the same for all sub-groups in a different culture. This would increase the applicability and external generalisation of tools across different age, culture, gender, and educational groups. It would also extend the usage of a single item towards different groups in international research. The IRT theory would have higher capacity to extend the application towards different groups, making it more suitable for cross-cultural/ international research.

However, in reality, the uni-dimensionality is difficult to achieve. For example, a psychometric test designed for measuring mathematic ability in group A, could end up reflecting more on English ability in a non-English speaker in group B. Under such conditions, the IRT is better than the classical model because of the ability to provide more detailed analysis for each individual. This includes different standard errors of measurement (SEM) across different ability levels (see Figure 1.15), instead of the single SEM per sample group as the classical model would suggest. The IRT theory would not assume the measuring ability of an item to be the same across all ability levels.

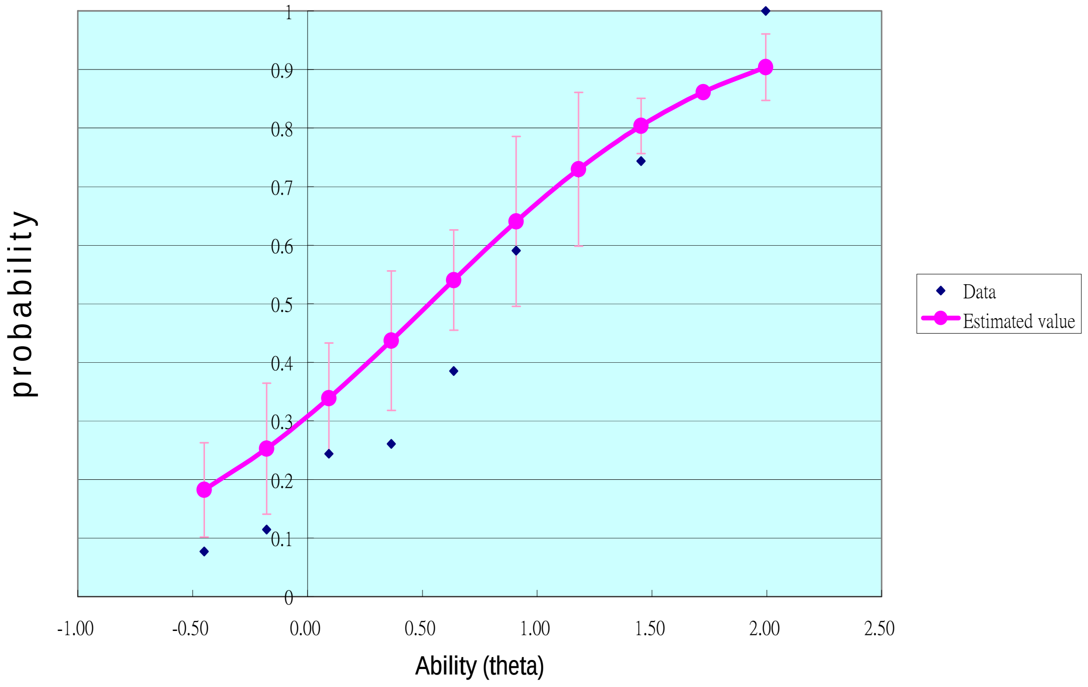


Figure 1.15 Illustration of different error estimation across different ability levels

The other aspect that makes the IRT model better than CTT in international research is the capability of ability estimation. The IRT model is more accurate in ability estimation than the classical 'total score' model. The classical 'total score' model often adds up all responses from one construct as the representation of ability. This model is often applied in different cognitive and educational testing. However, the IRT model takes a totally different approach, for example, given five candidates who scored differently across five items with different item parameters (see Table 1.1 on the following page).

According to the classical model of scoring, candidate 2 and candidate 3 would receive the same total score of 40% (two correct responses out of five). Under the IRT model, however, they would score differently due to the variation in item difficulty (see Table 1.1 on the following page).

Table 1.1 Five candidates' responses across five items (Yu, 2007)

	3PL Item parameters			Actual Response				
Item	a	b	c	candidate 1	candidate 2	candidate 3	candidate 4	candidate 5
1	1.27	1.19	0.1	Correct	Correct	FALSE	FALSE	FALSE
2	1.34	0.59	0.15	Correct	FALSE	FALSE	Correct	FALSE
3	1.14	0.15	0.15	Correct	Correct	FALSE	Correct	FALSE
4	1	-0.59	0.2	FALSE	FALSE	Correct	Correct	FALSE
5	0.67	-2	0.01	FALSE	FALSE	Correct	Correct	Correct
Total score				60%	40%	40%	80%	20%

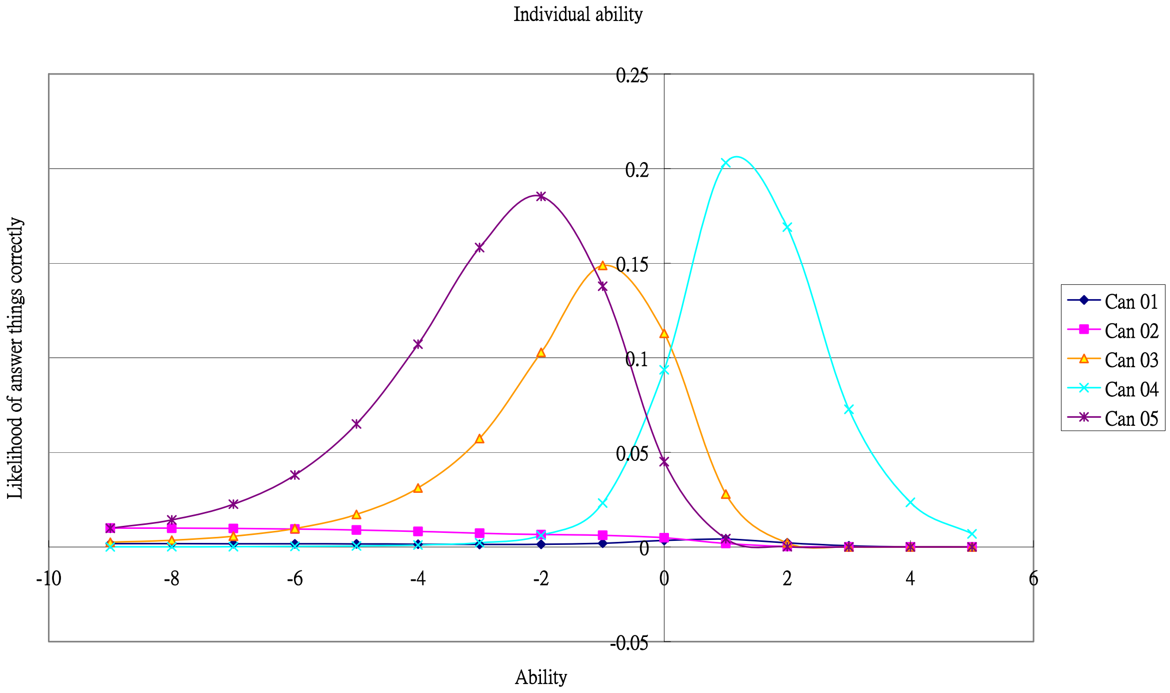


Figure 1.16 Illustration ability estimation

Although both candidates score two out of five, candidate 3 scores three relatively difficult items incorrectly (see Table 1.1, item 1. b=1.19, item 2. b=0.59, item 3. b=0.15) but scores the other simple item correctly (see Table 1.0, item 4. b=-0.56, item 5. b=-2).

The scoring pattern of candidate 3 is reasonable. Therefore following the IRT ability estimation method, it is estimated that candidate 3's ability is between -2 to 0 (see Figures 1.16 and 1.17*),* ability estimation is indicated by the peak of the curve).

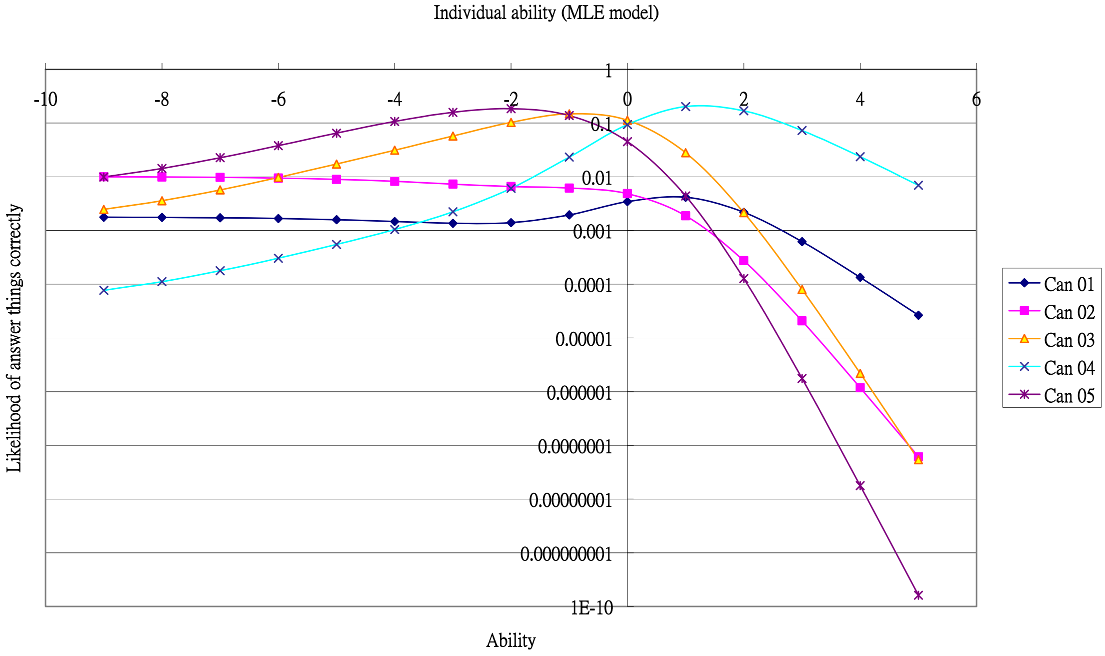


Figure 1.17 Illustration of Most Likelihood Estimation (MLE) of ability (Yu, 2007)

Candidate 2, however, also scores two out of five, but scored the difficult items correctly (item 1. b=1.19, item 3. b=0.15), and the simple items incorrectly (item 4. b=-0.59, item 5. b=-2). In such a case the results for candidate 2 would not be meaningful. It is more likely that this candidate would score item 1 and 2 correctly due to response bias (see Table 1.1, p. 25).

This model is useful in international research in examining the detailed structure of a construct. For example, candidate 2 could be a 'high ability' individual who was simply unable to read items 2, 4, and 5 correctly. From this result, the researcher can separate such individuals from the group and run the analysis separately to explore the possible external variables that influenced the results. This method could give the international researcher more leverage to investigate items (see Table 1.1, p. 25).

Overall, the IRT could provide individual ability estimations and general item information for different ability groups and individuals. It is a good tool for exploring item functioning in a cross-culture sample and provides a rich source for information.

1.4. Operationalisation

The focus of this research is on FC and the appropriate research methods to use with this type of instrument. The FC is operationalised as the PPA instrument. The correct research method has been operationalised as the IRT, CTT and other alternative statistical methods such as RICC, IRTCI and FCMCQ (see list of definitions). This research would explore all the above methods in order to reach a conclusion of the most accurate and efficient protocol that could be used in future FC research.

Due to the fact that this research takes place in an international setting, the cultural variable unavoidably plays a large part in this study. The typical factorial method cannot be applied due to the nature of Forced Choice, and this relationship is therefore explored via IRT methods.

1.5. Outcome summary

This research contains two parts. Part I is aimed at exploring which items should be amended (named Amend A in this research); and Part II investigates the 'quality' of the result yielded by this amendment (Amend A). The final output of the two studies is a protocol for FC-IRT research. Part I of the research suggests 16 items are necessary for amendment. However, 19 items also need amendments to maintain CPPA's construct integrity.

The results from Part II are successful. The 16 items recommended for amendment show better results (a better ICC curve), which means that the method has good practical applications. The results of this study seem to suggest that the combination of GRM-IRCCC and KTB is an acceptable method for FC psychometrics. However, a more suitable method still needs to be created for ordinal FC psychometrics.

The present study provides a research protocol for PPA-IRT research that was generated from current experience. It is supported by the results suggesting that GRM is applicable to the PPA. The research protocol for PPA would be:

1. Textual analysis for checking simple terminology error.

2. Functional analysis for checking the item discrimination index. The negative or low discriminating index would suggest poor item construction. Methods such as GRM-IRCCC or CPE-IRCCC can be used for this analysis.

3. Dynamical analysis is especially necessary for forced choice scales. Item contamination would suggest items are poorly aligned within a set. Methods that prove useful are KTB, RICC, FC-MCQ, and GRM-IRCCC.

1.6. Chapter summary and discussion

Previous research suggested that self-report measures are unavoidably affected by socially desired responding (SDR) style and bias, the results of which are exacerbated by culture differentiation. The forced choice (FC) method has been used to decrease SDR in previous research. However, due to its complexity and difficulty in exploration by classical test statistics, FC is seldom applied. The new development of item response theory (IRT) could open a fresh approach to the use and understanding of FC. It is with this objective in mind that this study investigates the application of item response theory in a forced choice psychometric instrument such as Thomas International's psychometric Personal Profile Analysis (PPA). The final product of this research is intended as a model for application of IRT in FC.

The three components of this study are (1) the development of a parameter estimation method, (2) item construct estimation, and (3) cross-cultural applicability. GRM and CPE would be used for parameter estimation, FCMCQ, RICC, and KTB function as a supporting method for parameter estimation and construct interpretation (IRTCI).

Based on analysis of Part I (n=650) of the research, it is suggested that 16 items were necessary for amendment. However, 19 items might also need to be amended to maintain CPPA's construct integrity. It was decided to change only 16 items and return it for the Part II research (n=307) to evaluate the effectiveness of the IRT amendment method.

The results from Part II were successful. The 16 items recommended for amendment show better results (i.e. a better ICC curve), which means that the method has good practical applicability.

This study suggests that the combination of GRM-IRCCC and KTB could be a suitable method for FC psychometrics. However, more research is required to deal with ordinal FC applications.

The protocol of PPA-IRT research should be carried out in sequence – textual, functional (GRM-IRCCC/CPE-ICC), dynamical (KTB/ FCMCQ), and constructual (GRM-IRCCC) – for better efficiency and quality.

To view this research critically, the final result only suggests IRT's applicability to FC measurements, and seems not to relate to SDR at all. This research can only provide an efficient model for FC creation via IRT methods, and could promote the use of FC, while indirectly reducing SDR. However, such a link is not direct and, as such, is questionable. Therefore, to ensure the firm link between FC and SDR, it is suggested direct FC-SDR research be conducted for future study.

CHAPTER 2. LITERATURE REVIEW

2.1. Introduction

The literature review covers the background of four areas. Thomas Personal profile analysis (PPA), forced choice method (FC), Classical test theory (CTT), and Item response theory (IRT).

2.2. Personal Profile Analysis (PPA)

This research used Thomas Personal Profile Analysis (PPA) as the main psychometric instrument for exploring various IRT methods. The PPA is a personality forced choice psychometric instrument that contains 96 items, or 24 item groups. It normally takes 15 - 30 minutes to complete. The PPA was developed from Marston's DISC model (by Hendrickson, 1996) and involves the four constructs; namely Dominance, Compliance, Influence, and Steadiness (DISC). The PPA explores an individual's behavioural potential in working environments (socially preferable expressions of self), under pressure (self image that can be maintained under stressful conditions), and real self image (theoretical 'true self') (Irvine, 2003).

2.2.1 Historical background

The early 1900s to 2000s could be named as the time of rising reductionism (Harth, 2004; Hooft, 2001; Peele, 1981) that gave birth to Marston's DISC theory. The academic world was amazed by the possibility of finding the basic, universal elements of various phenomena (Harth, 2004). Thus, it was proposed that entities, laws, or elements, such as electrons, protons, or neutrons, could function as the building blocks of the complex physical world, and serve to explain reality (Irvine, 2003).

In 1921, Albert Einstein was awarded the Nobel Prize in Physics for his discovery of the 'law of the photoelectric effect' (as cited in Szöllösi-Janze, 2009).

Seven years later, Frederick Griffith discovered the existence of DNA (as cited in Lorenz & Wackernagel, 1994). These developments in philosophy and physics also impacted on the psychological world. Spearman proposed the popular concept of 'g' as the fundamental factor for cognitive ability (Williams, Zimmerman, Zumbo, & Ross, 2003a), and Marston proposed the concept of 'psychon', the element for emotional energy, which is rooted in reductionism (as cited in Irvine, 2003).

Marston's 'psychon' are emotional energies that have its polarities. These energies stand in opposition to each other. The same energy might manifest in different kinds of 'emotional response'. According to Marston, these energies, or forces, originate from the biological structure of humans, and when interacting with the environment, manifests into four pathways (or polarities) taking four different directions (in Marston's original terms). Dominate, Influence, Submit, and Compliance (DISC) (as cited in Irvine, 2003).

Marston's DISC model was best explained by Berry's (1976) social leadership style interpretation. In a hostile environment, such as a migratory (hunting-gathering) society, the leader tends to possess 'Influential' qualities that enable him to interact on a friendly basis with most of his subordinates. The leader needs to be able to maintain/construct the social bond very quickly and successfully. In contrast, the subordinates tend to be 'submissive,' which is defined as 'willingness to support'. Subordinates, such as the tribe's doctor or hunters, are able to self-sustain.

However, it is only due to the friendship the leader provides that binds them together. The social hierarchy structure of such a system is not particularly clear. The leader and the subordinates keep within relatively short distances of each other, so that when the external environment is stressful, the leadership style tends to be relatively stress free due to the support received from the subordinate structure. I-S interaction is therefore popular among egalitarian hunting groups (Berry, 1976, 1980; Berry & Associates, 1986; Berry & Irvine, 1986).

In contrast, in a sedentary (agricultural, high population density) society, leaders tend to exercise 'dominance' (power, hierarchy, distance) via many social techniques. These would include retaining a little mystery, along with the creation of formalism, religion, mannerism, tradition, and distance. In such a culture, unquestioned patriotism (following unexplained cultural rules, traditions, and social disciplines) is normally a prerequisite. This social organisation can be defined as 'compliance' (Berry, 1976, 1980).

Individuals tend to adopt the interaction system that they are culturally most at ease with. PPA is an attempt to find out an individual's social interaction style they are comfortable with, I-S or D-C. And from the passive or active direction to determine individual's inclination towards I (active)-S (passive) or D (active)-C (active).

The present PPA models are not only based on Marston's original DISC terminology and definitions but they have also been shaped by other developments in psychometrics. Since 1928, the definition of DISC had been reshaped through the influence of different theorists and psychometrics. For example Allport's definition of Ascendance-Submission dimension in his Trait Theory in 1937 (as cited in Irvine, 2003), Murray's Need Theory in 1943 (as cited in Irvine, 2003), Jackson's Personality Domain in 1984 (as cited in Irvine, 2003), and Hendrickson's construct validity research in the USA, and the formulation of the PPA in 1958 (as cited in Irvine, 2003). The current definitions of the DISC descriptive terms are (Hendrickson, Undated/1958; Irvine, 2003).

- **Dominance (D)**: Assertive, Competitive, Direct, Driving, Forceful, Inquisitive, Self-Starter, Aggressive, Blunt, Egocentric, Daring, Decisive, Demanding, Dominating, Overbearing, Self-assured, Self-indulgent, Venturesome.

- **Influence (I)**: Communicative, Friendly, Influential, Persuasive, Positive, Verbal, Affable, Charismatic, Charming, Confident, Effusive, Generous, Gregarious, Optimistic, Participative, Poised, Promoter, Self-promoting, Sympathetic, Trusting.

- **Steadiness (S)**: Amiable, Deliberate, Dependable, Good Listener, Kind, Persistent, Accommodating, Easy-going, Industrious, Lenient, Non-demonstrative, Patient, Passive, Predictable, Relaxed, Self-controlled, Serene, Soft-tempered, Steady.

- **Compliance (C)**: Accurate, Careful, Compliant, Logical, Perfectionist, Precise, Systematic, Adaptable, Cautious, Conservative, Conventional, Diplomatic, Disciplined, Evasive, Open minded, Overly Dependent, Rational, Self-effacing, Worrier.

2.2.2 Psychometric quality - psychometric theory behind PPA, construction of test, and application of the test

2.2.2.1 Early origin in USA - creation of first form 1958

The Thomas PPA was constructed by Thomas Hendrickson in 1958 (as cited in Irvine, 2003). The original research selected various DISC terminologies according to Marston and Hendrickson's operational definition of DISC. Terminologies in four dimensions were then used to construct a preliminary form, which was given to 115 subjects (67 males / 8 females) for generating frequency responses. The occupational distribution used was. 46 college students, 17 teachers, 27 supervisors, 16 other professionals, 13 office workers, 6 miscellaneous (Hendrickson, 2007; Irvine, 2003).

2.2.2.2 Empirical test of first form, general reliability

The second version matched frequencies for each of the four words representing the four different dimensions into tetrads (item sets). In 1958, in the USA, the revised form was administered to a larger and more representative sample groups of 500 (388 males/112 females) divided between the following occupational groups: 212 managers, 128 professionals, 62 clerical, 38 salespeople, 34 machine operators, 36 miscellaneous.

A random sample of 100 was drawn from the above group to determine split-half reliability and inter-correlation among the four factors.

The results indicated that the Personal Profile had a satisfactory internal consistency when assessed in this way (see Table 2.1).

Table 2.1 Hendrickson's initial scales inter correlation and reliabilities (as cited in Irvine, 2003)

Scales	**D**	**I**	**S**	**C**
Inter-correlations		0.60	-0.42	-0.11
			-0.02	0.25
				0.34
Reliabilities				
Try Out Split-half	0.93	0.78	0.84	0.72
Internal	0.86	0.89	0.80	0.81
Re-test	0.84	0.70	0.77	0.87

To eliminate non-discriminating items from the scoring key, an item analysis was initiated. A random sample of 185 (130 male/55 female) was drawn from a population of 1 200 to re-test the quality of the new form (see Table 2.1). This sample had an occupational distribution of 89 managers, 35 technicians, 26 office workers, 12 engineers, 12 sales people, 6 staff and 5 miscellaneous (Irvine, 2003). The general results indicated that 75% of the D (Dominance) and S (Steadiness) items had good correlations (>.60). The I (Influence) and C (Compliance) items had 75% median inter correlations (>.40) and 30% good correlations (>.60)

2.2.3 Change of construct definition and new scoring key

The internal consistency was confirmed and the scoring key adjusted (see Table 2.1). At this stage, the Marston dimension of SUBMISSION was changed to STEADINESS and the Marston dimension of INDUCEMENT changed to INFLUENCE (Hendrickson, Undated/1958). A random sample of 100 (75 males / 25 females) was selected from the previous 1 200 group to test the new scoring key, and the results were correlated against the original trials. The correlations obtained ranged from .99 to .87, with an average of .96 (Hendrickson, Undated/1958).

2.2.4 Establishing the PPA graphs of the new form

Another sample of 1000 was drawn (752 males/248 females, 43% managers/supervisors, 18% salesman, and 11% clerical workers). The scores obtained via the new scoring key were converted to a percentile system that later evolved into the graphical reference, and later presented into three masks; namely work, under pressure, and self-graphs.

This research was conducted in the early 1960s and the final version led to an extensive report on the issue of behaviour in the work place. The report was submitted to the American Psychological Society, and is regarded as important both in terms of the results the PPA achieved, and the methodology of the research (Irvine, 2003).

In each item group (tetrad), four words are claimed to have equal response frequencies. High response words were grouped together with other high response words, low response words with other low response words. The 76 of the original 96 words were absorbed in this manner and five extra tetrads were constructed to bring the total once more to 24. Of the words retained, 39% are the same as in Marston (Hendrickson, Undated/1958). The PPA ended up using 24 tetrads, and 96 items to measure the four constructs (Hendrickson, Undated/1958).

2.2.5 UK introduction in 1981 and current form

The PPA questionnaire, as derived by Hendrickson, was introduced into the UK in 1981 following adjustments to take into account different perceptions between US and UK uses of the English language. The revision was completed to allow for contemporary attitudes to equal opportunities and gender neutrality (Hendrickson, Undated/1958).

2.2.6 Registration

The PPA is currently registered with the British Psychological Society (BPS) (Hendrickson, 2007) and the Health Professions Council South Africa (HPCSA) (HPCSA, 2006b; SATP, 2003).

2.2.7 Reliability and validity

2.2.7.1 Reliability

A test is said to be reliable if it provides the same score for each subject on different occasions. Thomas International recommends that the PPA be given at intervals of no less than three months (Irvine, 2003). The minimum acceptable coefficient for test reliability is 0.7 (Kline, 2000, DeVellis, 2003; Thompson, 2003). High test/retest reliability has been shown for the PPA in retesting, and the UK data are regularly reviewed. As an example, one test/retest reliability study involved 72 people, (47 males/25 females), all employed in executive or professional positions. Retest intervals ranged from 3-12 months with a mean of six months (Hendrickson, Undated/1958). The test/retest reliability coefficients of the PPA dimensions of DISC were as follows (see Table 2.2) These results suggest that the PPA is a reliable measure, and that it has stability over time.

Table 2.2 Test-retest reliability of PPA

	R
D	0.84
I	0.70
S	0.77
C	0.87

2.2.7.2 Validity

Early data suggested that the PPA and its interpretations from the DISC factors achieved a predictive (test-retest) validity of more than 85% (Irvine, Mettam & Syrad, 1994). The research showed that the PPA gives good predictive validity when objective and verifiable criteria are used. It showed clearly distinguishable profiles for different job types and also differences within profiles for successes and failures in these jobs.

The PPA's construct validity was investigated by correlating it with the Guilford-Zimmerman Temperament Survey (GZTS), 16PF and OPQ, and high correlations with these scales were obtained (Hendrickson, Undated/1958) (see Table 2.3).

In terms of construct structure, the PPA generates a high correlation between the DISC subscales, and the pattern remains stable in international research. Throughout international research, Russia (n=600), Holland (n=127), Turkey (n=214), Denmark (n=539), USA (n=1512), UK (n=4083), SA (n=5655) and China (n=650) all generated a similar result. In terms of reliability, the test/retest reliability was D=0.84, I=0.70, S=0.77, C=0.87 (Irvine, 2003) (see Tables 2.4 and 2.5).

Table 2.3 DISC construct comparison with G-ZTS (Hendrickson, Undated/1958)

Dominance	**Inducement Influence**	**Steadiness Submission**	**Compliance**
G-ZTS(+)	G-ZTS(+)	G-ZTS(+)	G-ZTS(+)
General activity	Social ascendance	Restraint	Restraint
masculinity	Sociability	Emotional stability	Reflectiveness
	Masculinity	Objectivity	
		Personal relations	
G-ZTS(-)	G-ZTS(-)	G-ZTS(-)	G-ZTS(-)
Restraint	Restraint	General activity	Sociability
	Reflectiveness	Reflectiveness	Emotional stability
			Masculinity
Supervisors	Supervisors	Supervisors	Supervisors
Ratings (+)	Ratings (+)	Ratings (-)	Ratings (-)
Technical	Directing	Work quality	Ambition
Qualifications			
Planning	Management	Judgment	
	Relations		
Judgment	Ambition	Creativity	
Creativity		Ambition	
Company Knowledge		Interest	
Ambition			

Note: (+) indicates positive correlation above 0.3, (-) indicates negative correlation below -0.3. Detail also contrasted with Supervisor's rating on candidates in DISC groups (Hendrickson, Undated/1958).

Table 2.4 Hendrickson's (1958) scale inter correlations and reliabilities (n=500, 388M, 122F) (Irvine, 2003)

Scales	D	I	S	C
		0.6	-0.42	-0.11
Inter correlations			-0.02	0.25
				0.34
Reliabilities				
Try Out Split-Half	0.93	0.87	0.84	0.72
Internal	0.86	0.89	0.8	0.81
Re-Test	0.84	0.7	0.77	0.87

Table 2.5 Similarity between item structuring and international research.(Pearson correlation) (Hendrickson, Undated/1958; Irvine, 2003)

Russia n=600	**D**	**I**	**S**	**C**
Dominance		0.03	-0.65	-0.51
Influence			-0.50	-0.58
Steadiness				0.47
Compliance				
Holland n=127	**D**	**I**	**S**	**C**
Dominance		0.11	-0.70	-0.60
Influence			-0.31	-0.57
Steadiness				0.38
Compliance				
Turkey n=214	**D**	**I**	**S**	**C**
Dominance		0.04	-0.72	-0.47
Influence			-0.32	-0.52
Steadiness				0.28
Compliance				
Denmark n=539	**D**	**I**	**S**	**C**
Dominance		-0.09	-0.73	-0.50
Influence			-0.31	-0.44
Steadiness				0.32
Compliance				
USA n=1512	**D**	**I**	**S**	**C**
Dominance		0.05	-0.78	-0.54

Influence			-0.45	-0.61
Steadiness				0.46
Compliance				
UK n= 4083	**D**	**I**	**S**	**C**
Dominance		-0.15	-0.75	-0.49
Influence			-0.21	-0.40
Steadiness				0.24
Compliance				
China* n=650	**D**	**I**	**S**	**C**
Dominance		-0.127	-0.606	-0.450
Influence			-0.398	-0.368
Steadiness				0.121
Compliance				

*China study is the current study, 2008

2.2.8 Summary and discussion

This research study utilised the PPA psychometric test of Thomas International. Thomas Hendrickson designed the test on the basis of Marston’s DISC theory in 1958, which again was influenced by the reductionism prevalent during that era. The DISC theory postulates that there are four different behavioural manifestations of psychological energy. Two axes of polarity are postulated as positive-negative and active-passive.

From the two axes, four quadrants of energies are named as: ‘Dominant (active-negative),’ ‘Influence (active-positive),’ ‘Submission (passive-positive),’ and ‘Compliance (passive-negative)’. The four quadrants attempt to describe the habitual direction that an individual would tend to ‘behave’ in when encountering social interaction.

Berry (1976) further interprets the framework in the social leadership style theory, arguing that IS is the product of an egalitarian migratory society, while DC is the product of a hierarchical sedentary society. It is also suggested by Berry (1976) that individuals would be more comfortable operating in a leadership style that is closer to their heritage.

The DISC model was later modified and operationalised by Hendrickson (Undated/1958), and with the aid of various researchers, the PPA psychometric test was developed. The reliability of PPA is generally acceptable (Test/retest, D=.84, I=.70, S=.77, C=.87). The validity of PPA has been conducted by means of construct validity studies with the Gilford-Zimmerman Temperament Study, 16PF, and the OPQ. The PPA's constructs were examined by means of the international comparison studies conducted in Russia (n=600), Holland (n=127), Turkey (n=214), Denmark (n=539), USA (n=1512), UK (n=4083), and China (n=650). These studies confirmed the validity of the PPA constructs (Ivrine, 2003).

However, although past research has suggested good test/retest reliability (all above .70), no duplicate study was conducted in the Chinese sample pools. There are two issues that can be raised. The first concerns the suitability of the PPA system for international usage, and the second issue concerns its internal consistency.

Hendrickson (Undated/1958) and Irvine (2003, 2007) provided a macro international study for PPA construct via correlational research (see Table 2.5), although the result can be interpreted as universally similar. Due to the nature of FC, it is suggested that all correlational based calculation methods for FC should be used cautiously. This is because, with the FC measurement, it is still questionable whether the correlational structure would give more information about the scoring method than about the actual constructs.

Yet again IRT is selected, for the very reason that it is not a correlational-based statistic. Although the GRM model shows a good result with SA PPA research, it should be noted that current research is in a highly experimental stage.

2.3. Forced choice psychometrics

2.3.1 Historical overview

2.3.1.1 Initial stage 1940s

Forced choice (FC) measuring appeared in the psychometric world around the 1930s. The first FC psychometric test that appeared was the Humm and Wadsworth Temperament scale (as cited in Humm, 1939a, 1939b; Humm & Wadsworth, 1933; Kruger, 1938; Nederhof, 1985) although initially it received little attention. A similar idea for the FC technique was developed by Horst and Wherry while working on personality measurement (as cited in Travers, 1951; Zavala, 1965). Later, the staff of the Personnel Research Section of the Adjutant General's office applied the technique to the problem of rating officers, which resulted in the production of the Army Efficiency Report (Sisson, 1948; Travers, 1951).

The technique of FC construction was later presented at the American Psychological Association's 54th annual meeting (Staff, 1946). It received good recognition and soon became one of the most popular methods to combat Social Desirable Response among the measurement society for the next two decades (Ford, 1964; Zavala, 1965).

2.3.1.2 Increased popularity 1950s -1960s

The forced choice method was created with the intention of decreasing the respondent's ability to manipulate results (Staff, 1946). The FC method received good recognition after the APA conference and was used for a variety of purposes. The FC technique has also been used by the US military to rate servicemen and general efficiency (Falk & Bayroff, 1954; Sisson, 1948; Wherry, 1959), engineers (Lepkowski, 1963), academic performance (Schutter & Maher, 1956), general personality (Denton, 1954; Gordon, 1951) , personnel selection (Bass, 1957; Ghiselli, 1954; Gordon & Stapleton, 1956), empathy measurement (Denton, 1954), and psycho-physical measurement of sensory function (Blackwell, 1952).

As interest in the FC method increased, some older tools were adapted into the FC format, such as various anxiety measurements (Bendig, 1956; Christie & Budnitzky, 1957; Heineman, 1953; Howe & Silverstein, 1960; Taylor, 1953).

2.3.1.3 Interest wanes - 1970s

Interest in the FC method, however, began to wane in the 1970s. Many researchers claimed that its methods were deeply flawed in various respects, which can be classified in terms of three aspects: construction, calculation, and interpretation (Brown & Harvey, 2003; Martinussen, et al., 2001; Nederhof, 1985).

Construction

Researchers have claimed that selecting item groups of equal preference is a theoretical concept that is extremely difficult to operationalise. Item selection statistical processes such as discrimination index, general factor loading, and magnitude of the group factor loading (Wherry, 1959) are most commonly used. However, these processes cannot guarantee that the items selected are equally preferable throughout differing samples, testing conditions, time, and cultures. Even after successful grouping, an item grouped according to equal preference would face another problem, i.e. unrealistic choice. FC methods often involve alignment of two or more constructs. To select one construct from the options provided does not resemble a real choice reality (Scott, 1963). In reality, humans do not need to give up one value for another. This makes the generalisationability of FC questionable.

Calculation problem

Theorists mostly criticise FC for its inability to use classical test theory statistical methods (such as Cronbach's Alpha) for internal consistency and factor analysis for construct validity (Bartram, 2007; Martinussen et al., 2001), an aspect that will be explored in a later sections.

Interpretation

Due to difficult and unrealistic choice options, it is commonly observed that missing responses occur when respondents become too frustrated to make a choice (Edwards & Diers, 1962). Also, there are no normative comparisons for the result, making it difficult to further differentiate individuals who possess similar profiles (Hicks, 1970; Johnson, Wood & Blinkhorn, 1988; Tenopyr, 1988). After all, FC is still found to be 'fakeable' because respondents can conclude what is required for specific professions (Vasilopoulos, Cucina, Dyomina, Morewitz & Reilly, 2006).

2.3.2 Debates in forced choice: Calculation

2.3.2.1 Lack of normative result, artificial correlation

There are many debates around FC psychometrics, mainly in relation to the statistical issue (Christiansen, Burns & Montgomery, 2005; Hicks, 1970; Johnson et al., 1988; Zavala, 1965). Ault and Barney (2007) suggest that FC is limited by its form, unlike the Likert scale, and FC would lead to certain types of relationship among constructs, such as artificial positive or negative correlations. The effect is constructed mainly because the forced choice item would only allow the subject to select between a limited number of options. As an example, in the Myers-Briggs Type Indicator (MBTI)'s forced choice, candidates are forced to select between 'thinking' or 'feeling' as a sub-option within an item, which leads inevitably to negative correlations between thinking and feeling. However, such correlations are 'artificial' (Ault & Barney, 2007).

2.3.2.2 Internal consistency cannot be tested

As mentioned above, due to the potential problem in FC of creating 'artificial' correlation, the integrity of reliability is also endangered, such as in Cronbach's Alpha and split-half reliability. According to Ault and Barney (2007), FC is relatively un-interpretable via normal classical test theory (CTT) psychometric reliability and validity indicators, such as Cronbach's Alpha or split-half reliability (Ault & Barney, 2007; Hicks, 1970; Tenopyr, 1988).

2.3.2.3 Factor analysis

Due to the fact that FC scales could distort the correlation result, early researchers had intuitively suspected that FC could not be applied to factor analysis (Edwards & Walsh, 1964). The suspicion was later confirmed in various studies (Clemans, 1996; Closs, 1996; Gordon, 1976; Johnson et al., 1988; Saville & Wilson, 1991). The 'artificial correlation' (as this paper terms it) is termed 'constrain in correlation' which would result in high or low in factor systems, thus making factor analysis unsuitable for FC measurements.

2.3.2.4 The other side of the problem with Likert scales

Some theorists highlight Likert scales' inability to counter pseudo and extreme response styles, and Underhill, Lords and Bearden (2006) emphasise the 'fake resistance' ability of FC methods in contrast with Likert scale psychometrics.

2.3.3 Contemporary view

Underhill et al. (2006) operationalised faking through comparing the mean difference between honest response results and optimised response results in two equivalent Likert and FC forms (2x2). They argued that if the difference is insignificant, this would imply that such psychometric instruments can be defined as difficult to 'fake'. The researchers used university students in the sample (n=172) and found significant differences (T test) in 5-point Likert scale psychometrics, but no significant difference in the FC format (Underhill et al., 2006). The researchers further concluded that FC is more difficult to 'fake' in socially preferable ways due to its unique nature (Underhill et al., 2006).

Research also indicates that FC data might provide problematic reliability (Baron, 1996) and factor analysis results (Cornwell & Dunlap, 1994). However, in contrast to the Likert scale, Baron (1996) also suggests that FC is closer to a realistic environment. Unlike the Likert scale, which assumes that constructs will not influence each other, FC presents a platform that forces candidates to select between constructs, which is more realistic (Baron, 1996).

Researchers such as Matthews and Oddy (1997) suggest that both FC (Ipsative scale) and Likert Scale (Normative Scale) have limitations. The Likert scale is endangered by pseudo/extreme response patterns, and FC by reliability. Therefore both should be used in practical assessment.

2.3.4 Summary and discussion

The forced choice psychometric method appeared around the 1930s. The FC concept was initiated from the clinical measurement researchers such as Humm-Wadesworth team (as cited in Humm, 1939a) and the Horst - Wherry team (as cited in Travers, 1951; Zavala, 1965). It received favourable recognition post the 54th APA annual meeting in 1946 (Staff, 1946). FC reached its highest popularity between the 1950s and 1960s, but its star began to wane in the 1970s. This was due mainly to the fact that FC is difficult to construct, interpret and calculate according to classical test theoretical (CTT) statistical protocols. Many researchers considered that FC could not be examined under common psychometric procedures, such as reliability/internal consistency and factorial/correlation statistics. Some theorists suggest that the FC method is not free of SDR, as is claimed.

Contemporary researchers looked back to the FC format with alternative tools, such as item response theory, which would suggest that FC can be interpreted and standardised effectively. With the help of contemporary technologies and calculations, such as CAT and IRT, FC might shed new light on the common SDR puzzles.

Early researchers on the Ipsative did not mark the difference between types of tools clearly. It might have been due to the fact that researchers often discuss around the tool without revealing the actual item and date on which it is used, which could delay publication (view the debates between Baier, 1951; Richardson, 1951; Sisson, 1948; Staff, 1946; Travers, 1951).

The FC items often come in item sets of 2 or 4, i.e. the two selections FC (dyad) sampled by MBTI and the four selections FC (tetrad) sampled by PPA. However, FC models are not fully defined or understood, which leads to varied interpretations and criticisms. It needs to be noted that much of the research centred on the 'forced choice' nature of the FC scale in spite of the fact that they are differentiated in various types and quality (Berkshire & Highland, 1953).

Although all FC tools have an FC element, the difference in form leads to a completely different statistical approaches, and cannot be considered as a single format. This implies that some of the criticism targeting FC cannot always be applied to all its forms. Research using an FC umbrella term to generalise and criticise FC could lead to over generalisation. For example, an early paper of Travers' (1951) considered the creation of FC as simply an alignment of equally preferable descriptive texts from two opposite sides of one construct.

In Travers' interpretation, putting two different constructs (two dyads) into one item set is the method of creating a tetrad (Travers, 1951). Travers' idea originated from Sisson's (1948) renewed FC Army scale (WD AGO Form 67-1 Part II). In Sisson's model, subjects are faced with two socially preferable and two un-preferable options. In Sisson's (1948) consideration, most 'friendly' assessors would tend to mark one item as 'Most' within one of the Positive items (see Table 2.6 below), i.e. item 01 between option a and b).

Table 2.6 Section IV of WD AGO from 67-1 Part II (as cited in Sisson, 1948)

	Actual item	**Type of item**
	Item 01	
a	People work for & with him because of his personality	Positive
b	Never rank-conscious	Positive
c	Thinks only of himself	Negative
d	Worries a great deal	Negative

While on the other hand they would choose one item as ‘Least’ in one of the negative items, individuals who are marked ‘M’ on the positive item, and also ‘L’ on the negative item would receive one positive mark. An item set (tetrad) could generate a maximum of two positive marks (Sisson, 1948).

However, in contrast to the later development, such as PPA, Sisson’s (1948) items possess easily interpreted polarities, and therefore might not have the ability to suppress SDR. The assessors could simply ‘spot’ the positive and negative items and select accordingly. In contrast, PPA items do not have any specific polarity, making it harder to evoke SDR response. Using Sisson’s (1948) scale as an example, it is suggested that not all FCs are constructed in the same way as Staff (1946) would suggest. Therefore to criticise all FCs for the same defect is an oversimplification.

2.4. Psychometric theory, from Classical Test Theory to Item Response Theory

2.4.1 Historical overview

The origin of psychological testing can be traced back to the mid-1800s. In the perceptual laboratories of Leipzig, Germany, 1879, Wilhelm Wundt and colleagues were the first to recognise the importance of obtaining psychological measurements under carefully controlled conditions. Wundts’ work mostly involved behavioural measurements such as reaction time, perceptual stimulation, auditory discrimination, or the estimation of the relative weights of objects. Wundt established the standard control condition for data collection (Crocker, 1986; Eid & Diener, 2005; Jen, 2001).

In 1883, British psychologist Sir Francis Galton, inspired by Charles Darwin’s evolutionary theory, shifted his focus to individual differences. Galton later developed various quantitative methods for analysis data. The research methods later inspired statistician Karl Pearson to develop correlation methods. Galton also inspired Charles Spearman to develop the well-known theory of intelligence ‘g’ and advanced correlational procedure factor analysis.

Later in 1904, in Spearman‘s attempts to explain fallible measures and true objective value, he laid the foundation of the classical true score model (equation 3-1) (Crocker, 1986; Williams, Zimmerman, Zumbo, & Ross, 2003b).

Despite many contributions by German and British psychologists, two Frenchmen first created a psychological test in the contemporary definition. In 1905 Alfred Binet and Theophile Simon, at the request of the French Education Department, developed a tool for identifying mentally deficient children. The Simon-Binet Scale is therefore the first intelligence psychometric test (as cited in Jen, 2001).

The terminology of ‘mental testing’ was later established in US by psychologist James McKeen Cattell’s classical text ‘Mental tests and measurements’. Cattell further established the concept of ‘normative’ interpretation of testing results. In 1904, E.L. Thorndike’s publishing of ‘An Introduction of the Theory of Mental and Social Measurements’ set forth the first systematic discussion of measurement problems and differing aspects of testing theories.

The concept of psychological testing was well received by the psychological world after the 1900s. It reached the height of its popularity between 1915 and 1930. From 1930 to 1945 the classical model emphasised the concept of reliability and validity, and various common statistical constructs, such as the standard error of measurement (SEM). Test/retest reliability was also developed during this period. As a result of World War II, the application of psychometrics moved from academic study to mass military human resource classification, management, and personnel selection during the period from 1945 to 1960.

Since 1960 till the present time item response theorists began to re-examine various CTT assumptions, using new statistical models and newly available artificial intelligence techniques, and they further refined scaling and measurement approaches (Crocker, 1986; Jen, 2001). The current period is therefore named the era of ‘modern test theory’ by item response theorists.

2.4.2 Theoretical overview

The concept of classical test theory was mostly explained in the early publication of Harold Gulliksen's 'Theory of Mental Test' (1950). Presentation of the terminology 'CTT' within contemporary texts could imply that different aspects exist within the measurement society. Using the term 'classical test theory' (CTT) implies the existence of 'modern test theory'. Some University textbooks do not even use the term 'classical test theory'. The content of classical theory such as true score, reliability, and validity are included in some books, but the term CTT is not used (Aiken, 1991; Barclay, 1991; Cohen, Swerdlik, & Smith., 1992; Cronbach, 1990; Dunn, Mehrotra, & Halonen, 2004; Gregory, 1992; Haynes & O'Brien, 2000; Kaplan & Saccuzzo, 2005; Kline, 2000; Murphy, 1988; Murphy & Davidshofer, 1998; Tallent, 1992; Walsh & Betz, 1990, 1995; Zechmeister & Posavac, 2003).

The initial exposition and discussion of this approach used the term 'classical true score theory' instead of 'classical test theory (CTT)' (see Allen & Yen, 1979). The term 'CTT' is used mostly by item response theorists to contrast it with IRT principles (Baker, 1992; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980a; Yu, 2007).

There are only a small minority of textbook writers who would use CTT/IRT classification (Aiken, 2000, 2003; Crocker, 1986; Eid & Diener, 2005; Hogan, 2003, 2007; Kline, 2005). It is suggested that not all researchers have reached consensus about item response theory as the 'modern' theory', as Crocker and Algina (1986) suggest in their 'Introduction to Classical and Modern Test theory'. It is also possible that the term CTT is considered too technical for introductory level texts.

CTT is commonly referred to as an umbrella terminology for many test methods (Kline, 2005). CTT includes most of the popular analysis techniques, and refers to all the measurement methods designed around Spearman's true score theory. The CTT assumes that the raw score (X) obtained by any subject is made up of a true component (T) and a random error (E) (Crocker, 1986; Gulliksen, 1950; Kline, 2005; Williams et al., 2003b).

This can be expressed as:

$$X=T+E \tag{3-1}$$

In terms of variance, CTT would derive from equation 3-1

$$VAR(X)=VAR(T) + VAR(E) \tag{3-2}$$

Further, the concept of reliability (R) is therefore defined as:

$$VAR(T)/VAR(X)=R \tag{3-3}$$

Which can be expressed as:

$$VAR(T)=R \times VAR(X) \tag{3-3.2}$$

Due to (3-1), R can also be expressed as:

$$R= 1- [VAR(E)/VAR(X)] \quad \text{or} \quad 1-R=[VAR(E)/VAR(X)] \tag{3-3.3}$$

Also, the Standard Error of Measurement (SEM) formula is

$$SEM= SD \times \sqrt{}\,(1-R) \tag{3-3.4}$$

When expressing the formula with (3-3.3) formula it should be:

$$SEM=SD \times \sqrt{}\,[VAR(E)/VAR(X)] \tag{3-3.5}$$

The basic assumption of true score theory is that the subject would get the same test score if he had an infinite number of testing sessions.

This assumption gave rise to many psychometric scales as well as various techniques such as descriptive statistics, mean/variance/standard error of measurement (SEM) (equation 3-3.5), measurements for types of reliability (equations 3-3~3-3.5), validity (KR21 and Cronbach's Alpha, factor analysis), and item analysis as the item benchmark tools (p value, D value). For examining item bias, the common CTT practice is to use the Mantel Haenszel model of Differential Item Functioning (DIF) analysis (Zumbo, 2007).

2.4.3 CTT assumptions and limitations

Popular theorists such as Lord (1980a), Baker (1992), Hambleton-Swaminathan (1985) and Kline (2005) indicate several limitations or short-comings arising from CTT's assumptions. The current study attempts to summarise them in terms of four aspects. The two commonly referred to aspects are: sample dependence and parallel-form reliability aspects, and the two uncommonly referred to are: statistic and realistic aspects.

2.4.3.1 Sample dependents and parallel-form reliability

Most texts would focus their discussion on these two topics, i.e. sample dependence and parallel form reliability (Alagumalai & Curtis, 2005; Baker, 2004; Fan, 1998; Hambleton & Jones, 1993; Hambleton & Swaminathan, 1985; Lord, 1980a; Yu, 2007).

Sample dependence is the most common limitation when referring to CTT. It is in regard to various parameters originating from true score theory that are particularly specific to the sampling/norm group. For example, difficulty (p-value) value, discrimination index (r or d value), and all types of reliability measures are highly dependent on sample groups - when the sample group changes, they also change. The essence of sample dependence is that a particular research study is highly dependent on the sample group itself, and therefore almost impossible to fully duplicate, so that it is difficult to generalise results. For example, the p-value (difficulty) would be high for a high-ability group, but low for a low-ability group.

The d-value (difficulty) would be lower for a homogenous group, but higher for a heterogeneous group.

The variances generated by different sample groups can also expected to be different, so that the reliability (r) changes according to the group. Sample dependence leads to the common practice of using different norm groups for interpreting results. However, due to the continuous changing nature, historical effect and sampling effect, it is very hard to ensure the stability of the above parameters. It is also difficult to truly compare candidates from different norm groups (Crocker & Algina, 1986; Hambleton & Jones, 1993; Hambleton & Swaminathan, 1985; Kline, 2005; Schumacker, 2005).

Parallel form and reliability is another commonly discussed topic. Most texts separate issues of parallel form and reliability, but they should actually be considered together because they have the same origins. Parallel form is one of important assumptions of true score theory. In CTT, if a candidate is able to complete an infinite set of instruments that measure the same construct and possess similar d, r, and p parameters, one is able to measure this candidate's 'true score'. The notion of an 'infinite' set of items was later defined as different 'parallel forms'.

However, when encountering actual practice, it is impossible for the 'parallel from' to be operationalised. It is not practical to request all candidates to complete infinite parallel forms in common social measurement practice (clinical, industrial, educational, or social). Most social measurements do not have parallel forms, and if they do, utilising parallel forms and assuming there are no test/retest effects (test results do not correlate with each other) is also illogical.

Reliability measurement is based on the parallel form assumption, which could be potentially problematic. Reliability parameters that base their foundation on test/retest, split-half, and internal consistency all originate from the parallel form assumption. This approach to reliability can therefore be questioned because this assumption of correlation is limited in reality (Hambleton & Swaminathan, 1985; Kline, 2005; Lord, 1980a; Yu, 2002).

Statistical issues commonly appear with IRT theorists contrasting IRT functions with CTT theory. This can be separated into three sub-categories: measurement, correlation, and total score issues. CTT measurement works only down to test level; it is very difficult to find out details at the level of items in regard to individual candidates. This would normally reflect on the standard error of measurement SEM. CTT assumes the sample SEM for all candidates across all ability levels, which is unrealistic and illogical.

For example, for two individuals with different levels of ability to receive the same SEM is not reasonable. In terms of IQ, if the SEM is 5, two candidates scored 100 and 115, to calculate the fiducial interval (confidence interval for individual score) at a confidence level=.05, one would need to add 2 SEM for each value. For the lower candidate, the score interval would be 90-110. For the higher candidate, the score interval would be 105-125 (Jooste, 2003). However, such a calculation is unrealistic. The lower value 100, which is closer to the mean, should therefore have most of the data and with a smaller fiducial interval. For the larger value, with relatively less data, the fiducial interval should be bigger. CTT would assume, unrealistically, that all ability levels would have the same SEM (Kline, 2005).

Correlation techniques originated from the work of Pearson and were followed by Charles Spearman. However, since Spearman plays such an important role in the creation of CTT, it is hard not to see a trace of correlation in all CTT methods. According to Wilson (1977), all CTT methods rely heavily on the Pearson Product Moment (PMM) correlation concept in definition of assumptions. The correlation concept is within the basic true score concept, reliability, and validity. Also, the fundamental true score concept that ‘there is no correlation between true score and error’ is considered difficult to prove in practice (Lord, 1980a; Zimmerman, 1980).

Total scores from CTT theory are commonly used as the indication of the construct or sub-construct. Given a situation where two candidates have the same total score from different items, this would not necessarily indicate that they are truly equal to each other. Furthermore, it is quite possible to generate the same total score from different response patterns (or social response bias from the Likert scale) or items.

Under such conditions it is difficult to make absolute assertions that both scores represent exactly the same construct. This issue would further lead to the reconsidering of various techniques based on total score related statistics, such as reliability, that would highly affected by item-total correlation (Hambleton & Swaminathan, 1985).

The **realistic** aspect mainly focuses on the practical aspect of measurement. It is suggested that the practical application of the CTT-based test procedure is relatively narrow. Due to the sampling norm principle CTT items cannot accommodate every level of ability, only the level that it was researched beforehand. It is also difficult to equate the results between tests if they are not a parallel form. Furthermore, the CTT aspects relating to the item bias/norm system are based on the assumption that there are no true differences between groups (Hambleton & Swaminathan, 1985).

2.4.4 Advantage of CTT

CTT is relatively easy to comprehend, construct, and apply. CTT's theoretical assumptions are easier to satisfy with real test data, and CTT requires a smaller sample size for initial research. In contemporary practice, it is still the most popular form of psychological testing (Hambleton & Jones, 1993; Lord, 1980a).

Table 2.7 Comparison between IRT and CTT (Hambleton & Jones, 1993)

Area	**Classical test theory**	**Item Response Theory**
Statistic Model	Linear	Non-Linear
Level of Assumption	Weak (i.e. easy to meet with test data)	Strong (i.e. more difficult to meet with test data)
Item-ability relationship	Not Specified	Item characteristic functions
Ability	Test Scores or estimated true scores are reported on the test-score scale (or a transformed test-score scale)	Ability scores are reported on the scale -∞ to +∞ (to a transformed scale)
Invariance of item and person statistics	No- item and person parameters are sample dependent	Yes- item and person parameters are sample independent, if model fits the test data
Item statistics (parameters)	p-value (difficulty), (r or d) discriminate index value, and reliability (r, Alpha, KR20/21, and Spearman-Brown formula)	a, b, and c (for three parameter model) plus corresponding item information functions.
Sample size	200 to 500 (in general)	Depends on the IRT model but larger samples, i.e., over 500, in general are needed

2.5. Item Response Theory

2.5.1 Historical overview

2.5.1.1 1940s: The beginnings

IRT was born during the time that CTT was at its most popular, and as a result of various debates and discussions regarding the true implication of reliability and validity. Early attempts to counter CTT shortcomings could be traced back to the work of Richardson (1936) and Lawley (1943). However, IRT really began with Tucker’s (1946) attempt to apply the term ‘item characteristic curve’ (ICC) in his text. Early works from Lawley (1944) and Lazarsfeld (1950) were influential in the work of Frederic Lord, who is considered to be the main founder of item response theory (IRT), or modern test theory.

2.5.1.2 Model evolution

Lord's publications established the first item response model and associate methods for parameter estimation and contributed to the application of this model (the normal ogive model) on real achievement and aptitude test data (1952, 1953a, 1953b). Lord initiated IRT with two parameters similar to a 'normal ogive' model, because this early model was equipped with only difficulty (position) and the discrimination index (slope) as parameters.

The design was similar to CTT's difficulty (p value) and discrimination index (d or r value), but presented in a comprehensive visual form that could encompass all levels of ability. However, this model was later found to be difficult to duplicate in practice.

Birnbaum (1957; 1958a, 1958b) subsequently substituted the ogive model with the now popular logistic model (two parameter logistic model. 2PL). Birnbaum (1968) also created the three parameter logistic model (3PL) that added a new guessing component (c value, guessing index) to the model. During the same period, Danish psychometrican George Rasch (1960, 1966a, 1966b) formulated the Rasch model, which is conceptually similar to the one parameter logistic model (1PL) with only difficulty (position of curve) remaining. Unlike 2PL or 3PL models, the Rasch model does not simulate how an item functions, but rather explores how it can act as a benchmark for how items 'should' work (Rasch, 1960, 1966a, 1966b). The Rasch model was very influential in IRT item development (Rasch, 1960, 1966a, 1966b).

The progress and recognition of IRT was initially slow in measurement society during the 1950s-1960s. This slow development was mainly due to the complexity and intensity of mathematical calculations required by the IRT model. During the 1970s, thanks to the advent of personal computers and various statistical softwares, the development gradually speeded up. Samejima (1969, 1972, 1973a, 1973b) expanded the original dichotomous model designed by Lord (1952), Birnbaum (1968) and Rasch (1966a), and this led to the development from ability tests with true/false responses to a polytomous and continuous response GRM suitable for the popular Likert format in affective psychometrics..

In the 1980s, IRT theorists further established the methods for test score equating (Lord, 1980a; Rentz & Bashaw, 1975; Wright & Stone, 1979), computer adaptive testing (Lord, 1974, 1977, 1980b; Weiss, 1976, 1982, 1978, 1980a, 1983), and test design/evaluation (Lord, 1980a; Wright & Stone, 1979).

2.5.2 IRT model characteristics/assumptions

IRT is based on the following assumptions/characteristics (Hambleton, 1989; Hambleton & Cook, 1977; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980a; Yu, 2007).

- **Local independence, IRT item parameters are sample free:** All parameters in 1PL, 2PL, or 3PL models are free of the sampling effect. Parameters should not differ from sample to sample.
- **Examinee error measurement**: Instead of total standard error of measurement, individual error measurement can be calculated.
- **Test-free ability**: The abilities are test free, due to the fact that items between different tests can be compared meaningfully.
- **Item/test information**: IRT's new measurement concept as the third gold standard other than CTT's famous validity and reliability, namely item information. Item/test information indicates the amount of information that an item/test combination could generate for measurement. Items that are associated with a unified response yield low information.
- **Individual difference**: With the help of the item set and item parameters, IRT can generate different scores for candidates with the same raw/total score.
- **Model checking**: IRT uses 'goodness of fit' to recheck a model's fitness and does not assume that the current model is an absolute fit.

2.5.2.1 Practical limitations for IRT model

Although IRT possess various positive traits, it is still not the most popular method in measurement society. According to Yu (2007) the ‘un-popularity’ of the IRT model is mainly due to the following.

Difficulty: IRT model is built on complex mathematical modelling, which is difficult to comprehend and apply in practice. It is therefore difficult to access for psychologists and educational researchers who have limited mathematical training.

Lacking application: Most IRT theorists have a statistical or mathematical background. However, this would imply that most of the IRT researchers would have a higher theoretical interest than concern for its practical application.

Calibration complexity: IRT cannot rely on manual calculation, the calculation that it involves would require, at least, the operation of a personal computer. Therefore the popularity of the IRT model only increased after the advent of the personal computer.

Lacking support from CTT theorists: CTT questions IRT’s applicability in a practical situations. For example, it is very difficult to attain the large sample size that is an IRT requirement in most practical research. Also, many IRT results are not used as a standard in psychometric society, and some of the associated statistics have a purely academic rather than practical focus.

Large sample requirement: For various published CTT validity and reliability research studies, the sample sizes vary between 200 to 500. In contrast, IRT’s entry sample size is 500, which implies that the approach cannot be used when research funds are limited.

2.5.3 Summary and discussion

Contemporary testing theory initiated during the mid 1880s. Inspired by various researchers before him, Spearman laid down the fundamental theory and model for Classical test theory, such as true score and correlation/regression theory in the 1900s.

During the two World Wars, psychometric theory expanded due to its ability for aiding selection of military human resource. However, during the 1940s and the 1960s, arising out of discussions on some theoretical limitations of Classical test theory (CTT), various measurement specialists with mathematical/statistical backgrounds started to suggest a new model, and this led to the creation of item response theory (IRT).

Lord is considered as one of the most influential individuals of this movement. IRT targeted various CTT limitation in true score theory, such as overemphasis on correlation/linear model, highly sample dependent, incomprehensive concept of standard error of measurement, and unrealistic parallel forms model/reliability. The IRT model claims to be much more powerful than its predecessor by suggesting a number of new concepts such as item information. These include a non-linear logistic model, local independence, individual error measurement, more individual differences, test free ability, and model re-checking procedures.

However, IRT lacks support from the general measurement society, mostly due to its mathematical complexity and large sample entry requirement. The current era is facing various CTT limitations, and with the support of personal computer, software, and the web 2.0 network, the focus shifts back to IRT. After all, IRT could be the modern model for the measurement of the future.

The CTT and IRT have almost the same application quality, although IRT offers a few additional qualities (that many researchers would not know how to apply). Questions are often raised regarding this issue. If both CTT and IRT's application aspects are similar, why should researchers study the much more complex IRT model instead of just CTT? Applying Ockham's Razor would suggest an inevitable conclusion because, after all, simpler is better (as cited by Encyclopedia Britanica, 2009).

This set of entangled issues results in a situation that is similar to the one between Newtonian gravity and Einstein's general relativity. Both models attempt to explain 'gravity,' but Newtonian's *matter attraction* is much simpler than Einstein's *space-time curvature*. Moreover, in terms of practical application, Newtonian's theory can be applied to most of the daily cases with only minor errors. Nevertheless, Einstein's theory is more capable of explaining these errors, and is further able to explain the curvature of light and time distortion due to speed and gravity.

A similar relationship exists between CTT and IRT. CTT is applicable to most social assessments and requires large samples. However, CTT is still limited by its correlation assumption. CTT is simple to use and can provide single SEM and summary statistics, but is not able to go into the dynamic world of measurement. IRT, on the other hand, is able to investigate these dynamical details and provide much more insight.

Once again, when we look into the true meaning of Ockham's Razor '*pluralitas non est ponenda sine necessitate'* (plurality should not be posited without necessity) (as cited by Encyclopedia Britanica, 2009), CTT is actually making more assumptions, and assumptions of greater scope, regarding the nature of testing response than IRT. However, it is not the intention of this paper to discard CTT. CTT is still very useful and reasonable to deal with common, small-scale applications. The suggestion is that IRT should be equally 'respected' because of its applicability in large-scale research projects.

CHAPTER 3. POSTULATES

3.1. Introduction

The current study is based upon many assumptions and postulates. Indeed it would not be possible to conduct the research and the associated analyses without making some assumptions, and the reliability and validity of the results of the research study are ultimately dependent on these assumptions. It is therefore necessary for readers to explore these assumptions, as they could point to fundamental mistakes in the research approach taken in this study.

At this stage the research focus should be revisited. Figure 3.1 states the basic research question of this study and three areas of involvements. The research focuses on **'Can item response theory (IRT) be used in forced choice psychometric (FC) adaptation?'** The three major areas of involvement would be **IRT issues, FC issues, and adaptation issues.**

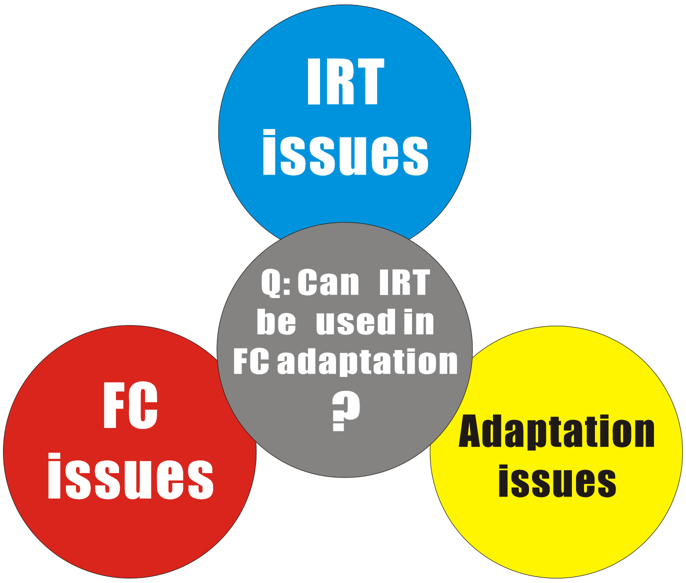


Figure 3.1 Basic research question and areas of involvement

However, when operationalised, the three areas would convert to. IRT to contrasting statistical methods, FC to personal profile analysis issues, and adaptation to Chinese cultural issues, as these three areas are what this research used to explored the research question (see Figure 3.2)

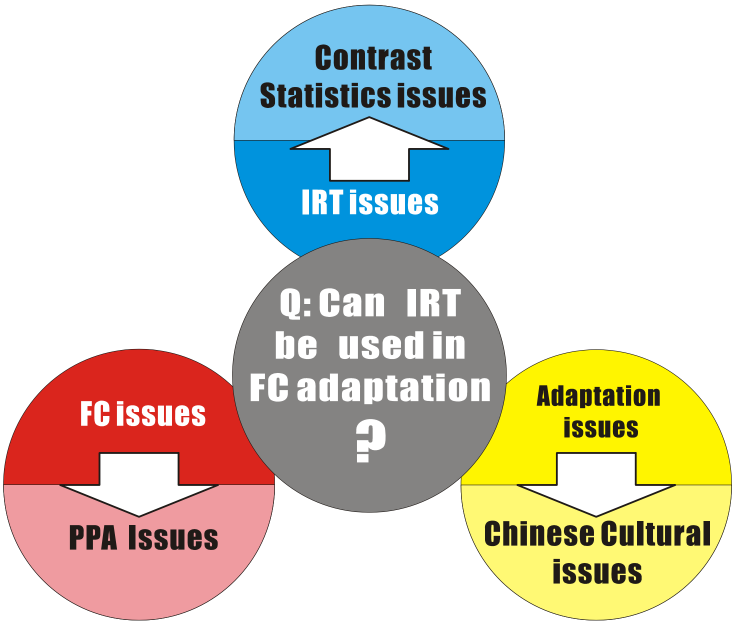


Figure 3.2 Research question operationalised and areas of involvement

In the current operationalisation, the three areas are not mutually exclusive. The interaction between three areas actually generates three more areas of issues. **'Can FC data be used in all stats? Can stats represent responses in another language accurately? Do PPA constructs exist in current Chinese culture?'** (See Figure 3.3.)

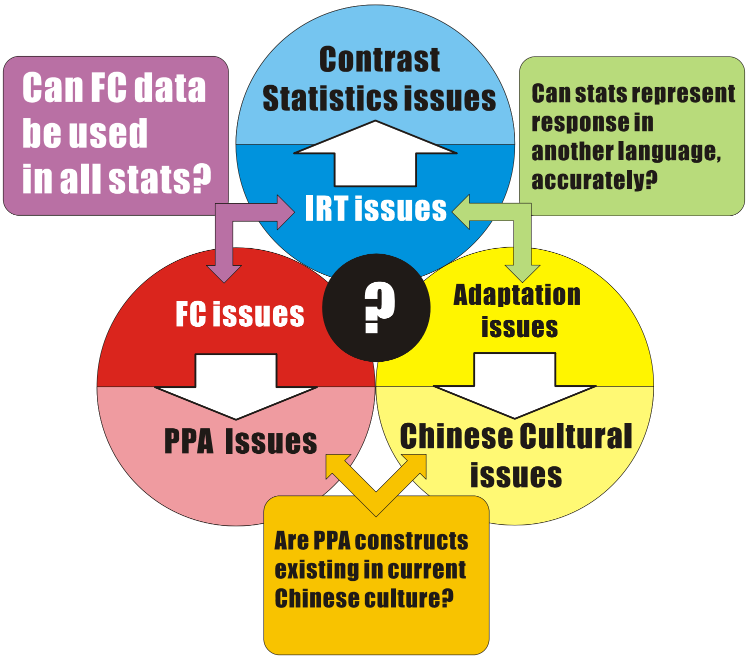

Can FC data be used in all stats?
Contrast Statistics issues
Can stats represent response in another language, accurately?
IRT issues
FC issues
?
Adaptation issues
PPA Issues
Chinese Cultural issues
Are PPA constructs existing in current Chinese culture?

Figure 3.3 Three operationalised areas and their interactions

These operationalised three main areas and three interaction areas contain many issues that cannot be covered by the limited scope of the current research. Therefore, this study can only assume the positivity of these areas in order to answer the main research question. The following chapter is dedicated to clarifying the following assumptions. The total six areas are designated in Figure 3.4.

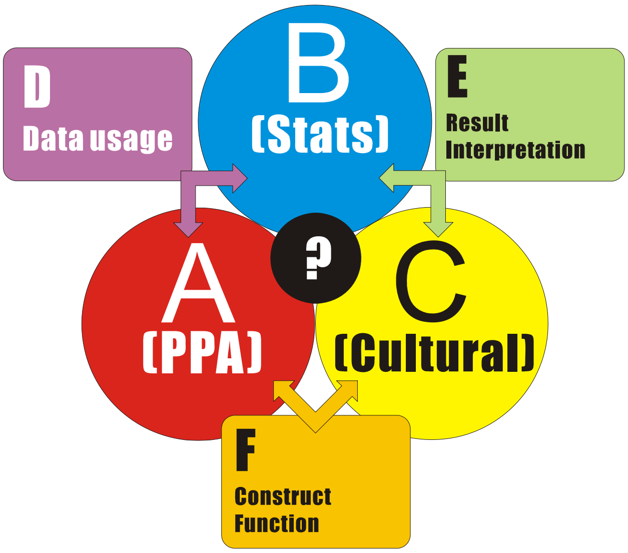


Figure 3.4 Six areas of assumptions in the current research

3.2. Area A: Basic assumption within personal profile analysis (PPA)

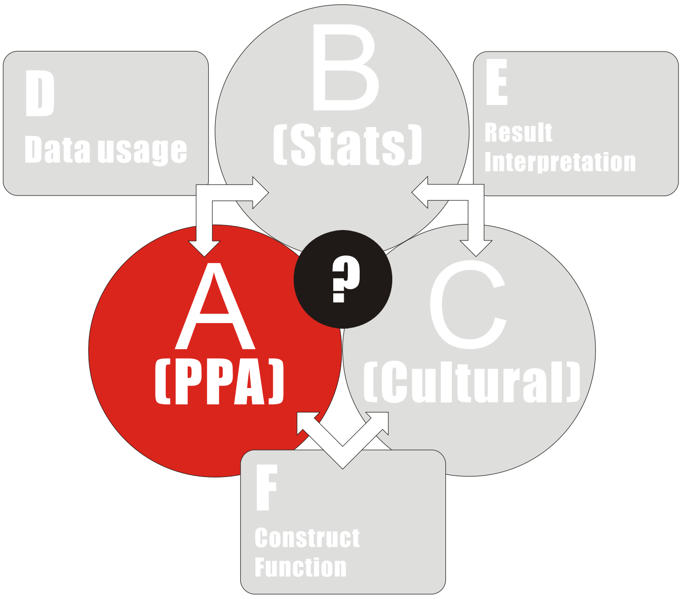


Figure 3.5 Area A: Basic assumption within Personal Profile Analysis (PPA)

PPA (personal profile analysis) is a HPCSA registered tool. However, to conduct the current study, this research still needed to assume the full functionality of various PPA construct structures, calculation methods, and current standardised administrative method are correct and functional. The following section introduces the fundamental assumptions such as IEC (Item Express Construct) and PAC (Pre-Assigned Construct); and Hi (high) and Lo (low) markers.

Table 3.1 Pre-assigned construct (PAC) of Personality Profile Analysis (PPA)-English (Hendrickson, Undated/1958; Irvine, 2003)

Tetrad no.	PAC	Item Terminology	PAC	Item Terminology	PAC	Item Terminology	PAC	Item Terminology
1	S	Gentle	HI I	Persuasive	C	Humble	Lo D	Original
2	I	Attractive	C	Dutiful	D	Stubborn	Lo S	Pleasant
3	Lo C	Easily Led	D	Bold	Hi S	Loyal	I	Charming
4	Hi C	Open minded	S	Obliging	Lo D	Will power	I	Cheerful
5	Lo I	Jolly	C	Precise	Lo D	Courageous	S	Even-tempered
6	D	Competitive	S	Considerate	Lo I	Happy	Lo C	Harmonious
7	Lo C	Fussy	Hi S	Obliging	D	Won't be beaten	I	Playful
8	Hi D	Brave	Hi I	Inspiring	Lo S	Willing to submit	Lo C	Timid
9	I	Sociable	S	Patient	D	Independent	Hi C	Soft-spoken
10	D	Adventurous	Hi C	Receptive	Lo I	Polite	S	Moderate
11	I	Talkative	S	Controlled	Lo C	Go with the flow	D	Decisive
12	Lo I	Polished	D	Daring	Hi C	Diplomatic	S	Satisfied
13	Hi D	Aggressive	I	Life of the party	S	Soft-touch	Lo C	Fearful
14	C	Cautious	Hi D	Determined	I	Convincing	Hi S	Good-natured
15	Hi S	Willing	no (I)	Eager	C	Agreeable	Lo D	High Spirited
16	HI I	Confident	Lo S	Sympathetic	Lo C	Tolerant	D	Assertive
17	Hi C	Well-disciplined	S	Generous	Lo I	Dramatic	D	Persistent
18	Hi I	Admirable	Hi S	Kind	Lo C	Resigned	D	Force-of-Character
19	Hi C	Respectful	D	Wants to be in the lead	I	Optimistic	S	Accommodating
20	D	Argumentative	Hi C	Adaptable	Lo S	Easy going	I	Light-hearted
21	Hi S / Lo I	Trusting	Lo S	Contented	D	Positive	C	Peaceful
22	I	Good-mixer	Lo C	Cultured	D	Vigorous	S	Caring
23	I	Companionable	Hi C	Accurate	D	Outspoken	Lo S	Restrained
24	D	Restless	S	Neighbourly	I	Popular	C	Faithful

Note: Hi – items are only scored when items are marked as ‘M’ / Lo – items are only scored when items are marked as ‘L’

IEC (Item Express Construct) and PAC (Pre-Assigned Construct)

This study assumes that if a PPA item is functioning according to its PAC (which can be expressed as IEC=PAC), it is considered a functional item. Background information regarding the postulation between Pre-Assign construct (PAC) and Item Expressed construct (IEC) will be introduced in the following sections.

3.2.1 Area A.1: PAC in PPA

The PPA items are pre-assigned to constructs (see Table 3.1). The ‘descriptive words’ originate from Marston’s theory; Hendrickson (Undated/1958) used Allport’s descriptor theory and drafted the first group of trait- terms in response to four constructs (as cited in Irvine, 2003). These trait terms were later embedded in the preliminary form, which was then administered to 115 subjects for investigating the response frequency for each trait term. The second version matched frequencies (as indices of equal response strength) for each of the words representing the DISC construct, and grouped them into 24 sets of tetrads (Irvine, 2003).

3.2.2 Area A.2: Hi - Lo markers

Another important mechanism that was embedded within the PPA is the Hi- Lo marker (see Table 3.1). This means that items are only scored when marked with ‘M’ or ‘L.’ Most of the PPA items are marked in both M and L situations.

Hi items are the ‘non-socially preferable’ items within the tetrad. This means that when placed within the tetrad, it is very likely that they would be marked as ‘L’ (least). Such items normally contain extreme descriptions of certain traits of the target culture, and they are therefore very likely to be marked as ‘L’ (least). Examples of such items are ‘Easily Led’, ‘Obliging’, ‘Aggressive’ or ‘Determined’. It is unlikely that ordinary individuals would describe themselves in these terms (i.e. marking them as ‘M’).

However, it is very common for subjects to deny that they possess these traits (i.e. by marking an item as ‘L’). Since scoring ‘L’ would decrease the spread due to the reason that most of the respondents would use such an option, that option is excluded. The item is scored only when it is marked with ‘M’ (most). This is the rationale behind **Hi item.**

Alternatively, **Lo items** are the ‘socially preferable’ items, which means that they are very likely to be marked as ‘M’ (most). Examples of such items are ‘Pleasant’, ‘Original’, ‘Open minded’ or ‘Polite’. These items are common terms that individuals would use in their daily lives. Therefore it is very likely that they would be used as self-descriptors. Scoring all the ‘M’ responses would also decrease the spread of the data. Therefore they are only scored when they are marked as ‘L’. This is the rationale behind **Lo item.**

The other items are neither popular nor un-popular in terms of ‘M’ or ‘L’ within their tetrads, and no Hi/Lo scoring structure is therefore assigned. However, this leads to a very important issue when translating the PPA. In short, the social preference of terminologies is the core of the PPA scoring method. It is very likely that the translated items would not function in the same way in another culture, because the social preference of the translated terminology would not be equivalent to the original. If so, the question remains as to whether the translated items would still function in the same way in the new culture.

3.3. Area B: Basic assumptions within the statistical methods, contrasting Item Response Theory and other statistical methods

Various methods are used to contrast the functionality of item response theory. In all these methods, the assumption must be made that they are functional and are correct in order for them to be applied in this particular context. This research also assumed that they can be compared with one another.

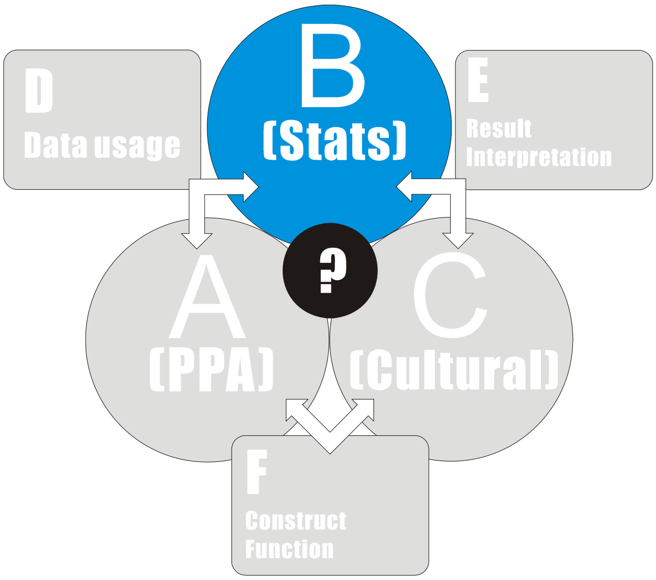


Figure 3.6 Area B: Basic assumptions within statistical methods

3.3.1 Area B.1: Face validity of the level of measurement conversion

In this study it was decided to present the Item Characteristic Curve (ICC) in its raw form as the main theme to contrast with other forms. This means that no parameter estimation was used in the ICC. Parameter estimation methods was only excluded from the ICC, but the other methods, such as general response model (GRM) and correlational parameter estimation (CPE), still required parameter estimation methods.

It is further assumed that this method is the best representation of the PPA data; the main reason being the complexity of FC data. However, to reconfirm the data, the general response model (GRM) family, Kendal's Tau B (KTB) model, and forced choice to multiple choice question conversion (FCMCQ) would be used to contrast the difference. This research would also need to assume all above methods' functionality and can be applied and contrasted with one another in the current research. More detail of how PPA data are used in all the statistical methods would be explored in area D. Detailed assumptions on how the results can be interpreted from all the statistical methods would be explored in area E.

3.4. Area C: Basic assumptions within Chinese Culture and psychometric adaptation

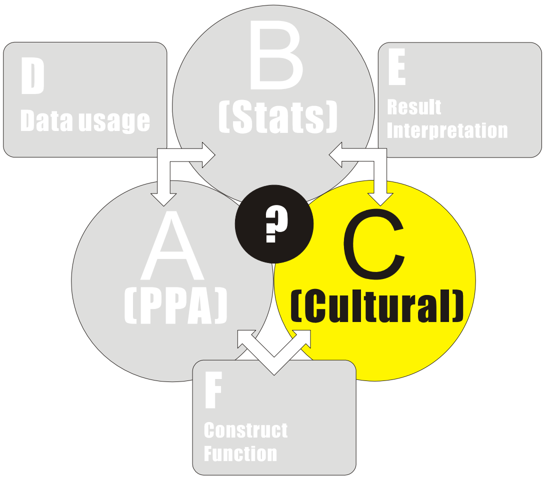


Figure 3.7 Area C: Basic assumptions within cultural issue

Psychometric adaptation is a complex subject matter. It cannot be conducted by translation only. When cultural and historical contexts are changed, many of the terminologies no longer yield the similar meaning, preference, and psychological responses. All these factors are important to the functionality of a psychometric test.

Chinese culture is currently going through a rapid phase of change. This changing phase would quickly shift the valued and system of terminology towards global values and standards. The basic assumption of this area is that conducting adaptation research in the current Chinese culture can yield stable results that can be generalised to the current Chinese population; at least valid for an acceptable time period. This is a very important assumption due to the reason that the changes within Chinese culture are very fast and complex. This research also assumes that the results yielded from this adaptation research derive from the treatment, and not from the natural cultural/historical changes.

Further, In terms of adaptation, it is also important to question the functionality of the PPA's constructs. The detailed assumptions of the functionality of various PPA constructs would be explored in area F.

3.5. Area D: Basic assumptions within data usage

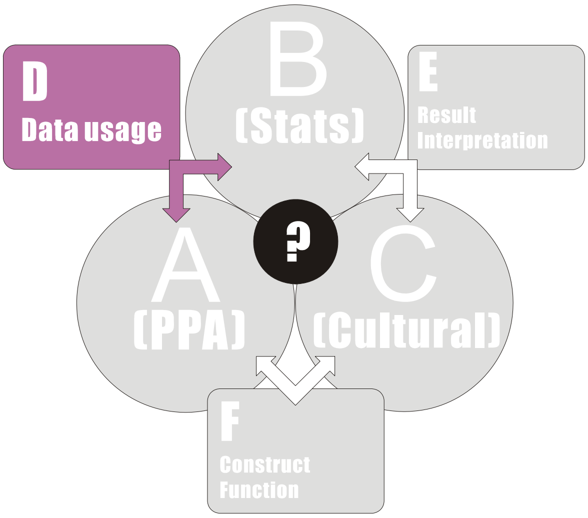


Figure 3.8 Area D: Basic assumptions within data usage

The fundamental assumption of this area is that forced choice (FC) data can be used in various statistical methods. The FC format would generate data that lies somewhere between ordinal and nominal. The nature of this data would lead to difficulty in the selection of the correct statistical methods.

Due to the reason that IRT is not widely used in FC psychometrics, especially not with Personal Profile Analysis, this study postulates that various models such as raw ICC, the three-parameter logistic model and interpretation, GRM (Samejima, 1999), KTB, FCMCQ and plotting methods are applicable for PPA.

3.5.1 Area D.1: Is PPA forced choice format nominal or ordinal and what parameter estimation method should be used?

Various bodies of literature suggest that in order to interpret item response results, a particular parameter estimation method should be used to estimate parameters. For ordinal or nominal tests, it is suggested that the GRM be applied (Samejima, 1999). However, this model was designed for Likert scale items, which makes it questionable when applied to FC items.

The PPA items have some similarities to the ordinal (Likert) format, so that an ordinal rating can be given to the four items in a tetrad. This means that 'M', 'L', and 'blank' can be treated as an ordinal response per 'descriptive word'. However, the PPA's FC ordinal interpretation is an assumption, and as such it is questionable because all the 'descriptive words' within one item group (tetrad) are not mutually exclusive.

The non-mutually exclusive condition implies that if a candidate selects an item with a tetrad as 'M/L', there would be no possibility of he/she also selecting another item as 'M/L'. This condition is largely different from that applying in the common ordinal Likert measurement situation (Hendrickson, Undated/1958).

Such a relationship can be summarised in a typical statistical combination of $_4P_2 = \frac{4!}{(4-2)!} = 12$ (choosing two words within four options, when the order of selection counts as difference). The total options would be 12 (possible combinations of each tetrad) x 24 (number of tetrads within the PPA) = 288 (different types of response). Such a result format is very different from a normal Likert tool. In the common Likert approach three (the three selection ranges for PPA item, M/blank/L) ordinal responses would typically be required. With 96 items (total number of PPA items is 96), the possible combinations would be 3^{96} for Likert items, which is much more than corresponding range for the FC questions (i.e. 288).

Also, FC results should be considered nominal instead of ordinal. This is because none of the responses have an ordinal relationship with each other (i.e. MDLS-*most D least S,* MSLI, MILD, has no ordinal relationship with each other) (Hendrickson, 1983).

This raises the question as to whether a researcher should consider the FC scale as ordinal or nominal? The answer is 'both'. This study is based on the following assumptions.

- When the scale is considered ordinal, the researcher can investigate the functionality per 'word' within a tetrad.
- When the scale is considered nominal, the researcher can investigate the functionality of the 'tetrad.'

In terms of ordinal, PPA's 24 item sets would be break into 96 separated ordinal items. The response would be coded as Most (M)= 3, Blank= 2, Least (L)= 1. This would enable the researcher to investigate how each 'word' functions with its item set (tetrad). The RICC (Raw Item Characteristic Curve), CPE (Correlation Parameter estimation method), GRM (General Response method), and KTB (Kandal's Tau B) would accept this assumption and process the FC data as ordinal data.

In terms of nominal, the 12 possible combination results from each item set would each be treated as 12 nominal responses; similar to multiple choice responses. They are treated as nominal due to reason that none of options is higher in strength than one another. The forced-choice to multiple-choice question (FCMCQ) flattening method would follow this assumption and process the FC data as nominal data. This enables the researcher to investigate which combinations are 'over popular' or 'over unpopular'. Following this assumption the researcher is able to investigate the dynamic relationship within the tetrad.

3.6. Area E: Basic assumption within data interpretation

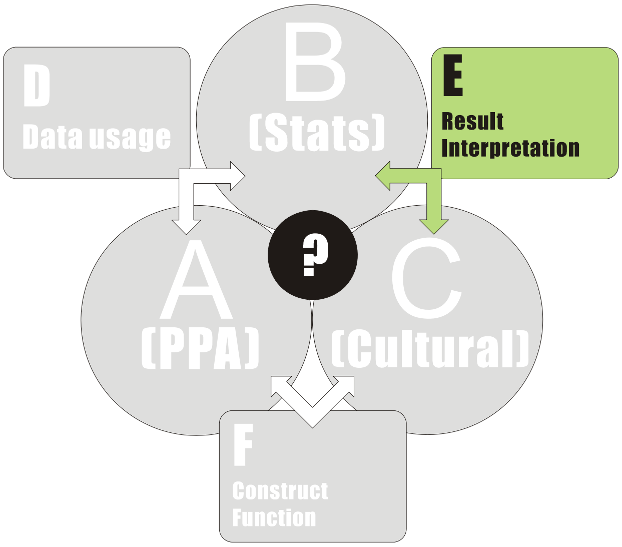


Figure 3.9 Area E: Basic assumption within data interpretation

Many assumptions are made in this area. The basic assumption is that all the data generated via current statistical methods can be interpreted meaningfully. The assumptions are made that these results reflect most to the treatment but not the instability nature of statistical methods or culture/historical effect from the sample group (see E.1, E.2).

The sub-assumptions are definitions of success and problematic responses. The success responses such as assuming the positive relationship between per-assignment construct (PAC) and item expressed construct (IEC) can be defined as success results (see E3). On the other hand, under what types of result would this research define that the results are problematic, or contaminated? (See E.4, E.5).

- Area E.1: IRT can be used to interpret cross-cultural psychometrics.

IRT can be used in an international setting. It assumed that local independence can also act as the benchmarking device to examine item functionality.

- Area E.2: Raw ICC can be used to explain multi-dimensional constructs of PPA.

The current study assumes that RICC, GRM, FCMCQ, and KTB can be interpreted meaningfully in the context of an FC instrument.

- Area E.3: Positive relationship between PAC and IEC could be defined as Positive ICC

This study applied item response theory to investigate the item expressed construct (IEC). IEC is operationalised as the 'true construct' that such an item 'expressed' when placed within a tetrad. This study postulates that if an item functions according to its PAC it should generate a positive ICC when defining the PAC as the 'ability'. This means that, if such an item functions according to its PAC, the probability of using such item should increase when the construct score (PAC) increases.

- Area E.4: If an item does not demonstrate complete synchronicity between PAC and IEC, such a tetrad is contaminated.

In terms of IRT, it is operationalised that if an item demonstrates a positive ICC towards constructs other than its PAC, such an item is defined as 'contaminated' (see below).

- Area E.5: Forced choice item group 'cross contamination'.

Table 3.2 Example of per assigned construct (PAC) within PPA system tetrad 02

I	Attractive	C	Dutiful	D	Stubborn	Lo S	Pleasant

This section introduces the possible cause of contamination in PPA items. However, various explanations of causality are not fully investigated, and can only be treated as assumptions. It is assumed that the dysfunction of one item within a tetrad leads to dysfunction of the entire tetrad. This study operationalised this phenomenon as *contamination (defined in E.4).*

This study postulates three types of contamination: *un-popular*, *over-popular*, and *vague item*s. The following section illustrates three aspects via PPA tetrad 02 (see Table 3.2).

Type 01: Un-popularity, for example; due to the un-popularity of the term 'stubborn' the probability of marking this item as 'M' would be distributed to three other items. In IEC, one could observe that other items are loaded with 'D' constructs distributed by the term 'stubborn'. In IEC, 'Attractive' would be ID instead of I, 'Dutiful' would be CD instead of C, and 'Pleasant' would be SD instead of S (see Figure 3.10).

Vice versa, the 'unpopular' item would also 'attract' all the negative response, the 'L (least)' response to itself. When examining the distribution of 'L', the probability of other items been marked as 'L' would be attracted to the item 'Stubborn'.

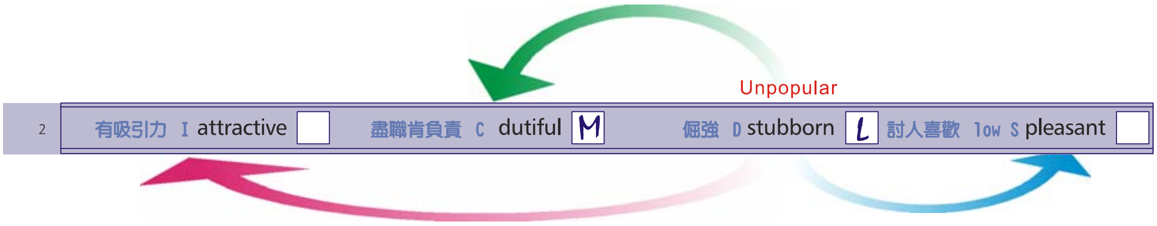


Figure 3.10 Type one contaminations: Unpopular items

Note: The probability of marking this item 'stubborn' as 'M' would be distributed to three other items.

Type 02: The over-popularity of an item within a tetrad could also be a potential cause of contamination. Over-popular items could attract responses from other items within the same tetrad, which can be observed by mixing the constructs. Due to the popularity of the term 'dutiful', the probability of marking this item as 'M' would be attracted from three other items. In IEC, one should observe that item C is loaded with constructs from the other items, there the IEC of the term 'dutiful' would become from C to ICDS. (see Figure 3.11).

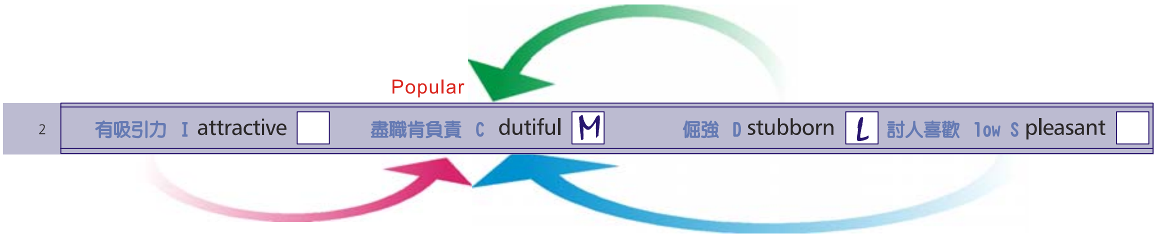


Figure 3.11Type two contaminations: Popular items

Note: The over-popular term 'dutiful' increases the probability of that item would be marked as 'M' over the three other items.

Type 03: Would be due to a vague item, a contaminated implication that leads to differentiation between IEC and PAC. When an item actually has the implication of other constructa it would also attract the relating construct. In IEC, the term 'dutiful' should mean C only. However, it also has an S connotation. Therefore the IEC of the item would be CS instead of S (see Figure 3.12). The 'mixing construct' occurs due to the reason that items existed exclusively within one construct as the researcher intended. Some items and constructs have more than one construct implication. This occurs quite often, such as many of the S items can have a C implication, and many of I item could have a D implication.

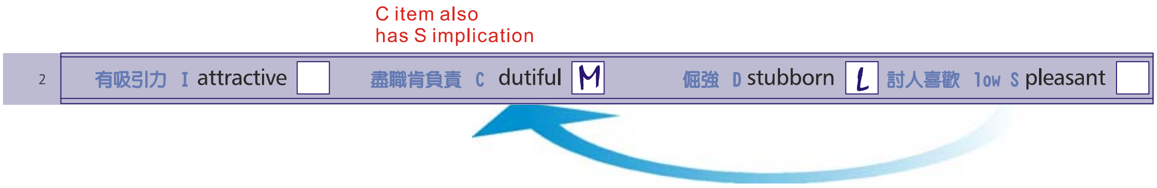


Figure 3.12 Type two contaminations: Vague items

Note: The term 'dutiful' should mean C only. However, it also has an S connotation. Therefore the IEC of the item would be CS instead of S.

The cause for contamination is complex. It could stem from the fact that the item cannot be interpreted in terms of its original construct (PAC) in the new culture. It could also stem from the fact that other items within that tetrad are not functioning correctly. The current study does not try to answer the question of how it originates, but simply attempts to provide an indication of the existence of 'contamination'.

3.7. Area F: Basic assumption within construct functioning

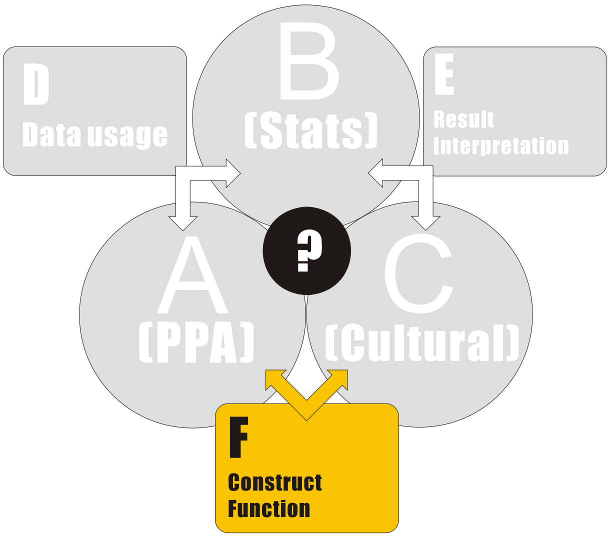


Figure 3.13 Area F: Basic assumption within construct functioning

This research is conducted based on the assumption that various PPA constructs still exist and are partially functioning within the current Chinese culture. The current Chinese trial version (A251 and C7) is a direct translation from the original item. These lead to the question of the functionality of the translated version. This paper assumes that this form is 'partially' functional and uses this trial version (A251 and C7) to collect the necessary information to create the better version.

3.7.1 Area F: Chinese PPA PAC, do they work?

The original items were constructed based on a USA sample in 1958. The Chinese trial version (A251 and C7) adopts a similar PAC structure in the translation. The old form (A251 and C7) text and PAC are (see Table 3.3).

Table 3.3 Pre-assigned construct (PAC) of Personality Profile Analysis (PPA)-Chinese (old version A251 and C7) (Hendrickson, Undated/1958)

Tetrad Order	PAC	Chinese terminology	PAC	Chinese terminology	PAC	Chinese terminology	PAC	Chinese terminology
1	S	溫和	HI I	能夠說服別人	C	羞怯	Lo D	做事與眾不同
2	I	待人友好	C	願意合作	D	固執	Lo S	溫柔可愛
3	Lo C	易被領導	D	勇敢	Hi S	值得信賴	I	喜歡與人交往
4	Hi C	開明	S	盡力取悅別人	Lo D	有意志力	I	快活
5	Lo I	非常風趣	C	辦事精細	Lo D	有膽量	S	性情平和
6	D	喜歡挑戰	S	體貼別人	Lo I	愉快幸福	Lo C	不喜歡衝突
7	Lo C	愛挑剔	Hi S	順從	D	好勝	I	喜歡嬉戲
8	Hi D	敢與參與	Hi I	激勵他人	Lo S	願意遵從	Lo C	膽小
9	I	好交際	S	有耐心	D	獨立自主	Hi C	說話溫和
10	D	喜歡冒險	Hi C	願意接受忠告	Lo I	謙虛	S	冷靜
11	I	健談	S	自制力強	Lo C	按常規辦事	D	遇事果斷
12	Lo I	文雅有禮	D	敢做敢爲	Hi C	圓通靈巧	S	心滿意足
13	Hi D	喜歡承擔責任	I	善於與人交往	S	易被利用	Lo C	不願冒險
14	C	避開麻煩	Hi D	專心做事	I	能說服別人接受自己的觀點	Hi S	樂意且真誠
15	Hi S	願意幫助他人	no (I)	爭切	C	討人喜歡	Lo D	有朝氣
16	HI I	自信	Lo S	有同情心	Lo C	會考慮他人的觀點	D	維護自己的權益
17	Hi C	嚴於律己	S	願意與人分享	Lo I	活潑	D	完成任務
18	Hi I	值得稱讚	Hi S	爲人和善	Lo C	依賴他人	D	決心取得成果
19	Hi C	敬重他人	D	喜歡冒險	I	凡事都很樂觀	S	不自私
20	D	好辯	Hi C	會變通	Lo S	隨和	I	喜歡逗樂
21	Hi S / Lo I	信賴他人	Lo S	知足	D	自信樂觀	C	平和
22	I	易於結交	Lo C	舉止得體	D	精力充沛	S	善解人意寬以待人
23	I	樂於交友	Hi C	做事追求正確	D	有話直說	Lo S	傾於深藏不露
24	D	易厭倦	S	樂於助人	I	希望被人喜愛與羡慕	C	忠實可靠

Note: Hi – items are only scored when items are marked as ‘M’; Lo – items are only scored when items are marked as ‘L’.

3.7.2 Area F: Problems of Chinese translation

The current Chinese trial version (A251 and C7) is a direct translation from the original item. However, such a translation process presents several serious issues. Firstly, although the items are translated according to the original English items and constructs, there is no guarantee that such items would function according to the same PAC.

Secondly, the Chinese Trial PPA's response frequencies have not been researched. It is therefore quite possible that some items with the Hi-Lo marker within a tetrad are no longer functioning appropriately. Thirdly, the tetrads' combinations and their effects have not been researched. Because items are grouped into tetrads, dysfunction of one term within a tetrad could jeopardise the functionality of the tetrad completely.

3.7.3 Area F: 3.5 PPA and cultural difference in item perception

Original IRT theory claims that its algorithm is 'sample-free (local independence)' (Wright & Douglas, 1977). It assumes that the descriptors of test items are independent of the sample (Hambleton & Swaminathan, 1985). The item parameters should remain the same despite different sample groups (Hambleton, 1983; Hambleton & Swaminathan, 1985; Yu, 2002).

This study expanded on this postulate and claims that such a relationship could also exist in cross-cultural conditions. However, this study does not follow Wright and Douglas's (1977) claim that IRT is completely 'sample–free'. In this study it is assumed that item bias does occur.

Similar to the Rasch's model, in this study it is assumed that IRT can be used as a golden standard to explore the possible interpretation and functionality within cross cultural applications (as cited by Mellenbergh, 1989). Further, it is assumed that the difference between Part I and Part II can be attributed to the amendments made by the researcher, rather than a difference in samples, due to the reason that Part I and Part II used different sample groups.

CHAPTER 4. METHOD

4.1. Research procedure overview

The study reported in this dissertation comprises two parts. Part I used the old Chinese PPA form to collect 650 samples. All data were collected via standardised procedures. After collection, the data were cleaned and analysed via CTT, IRT and some experimental statistics. The research results of Part I were used to create an amended form. This form was sent to China for Part II of the research and to collect another sample of 307 (see Figure 4.1). It went through the same statistical analysis for comparison with the previous results. The final product of this study is a PPA-IRT research protocol that can be used for future research.

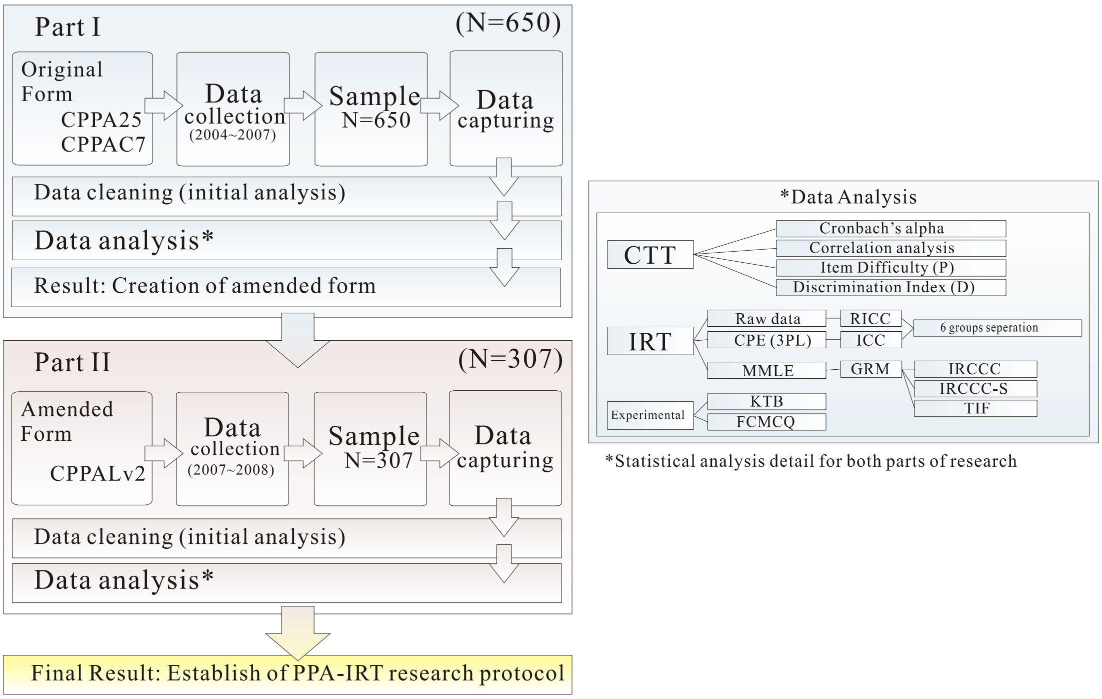


Figure 4.1 Research and statistical process

4.2. Psychometric administration and procedures

All the Thomas International psychometric instruments are administered in a standardised procedure. Clients are given a consent form to sign before testing, requesting the client's agreement that the material will be used for research purposes. All materials are kept confidential.

The client is then given an introduction to PPA. This involves aspects such as that all item sets need to be assigned only one M (most) and one L (least), and that the clients need to select the terms that apply to them the 'most' when they are facing a difficulty of choice. The administrator also explains that there are no right or wrong answers, nor are there time limits for completing the PPA. The written instructions further advise clients to answer the questions as if they are in a working situation. The assessment normally takes 10-15 minutes.

After completion, the administrator rechecks the form for mistakes or missing responses. If any are found, the form is returned to the client for amendment. The result is scored using the Thomas International software package. The system generates the text report in a PDF format. A report is sent to the client, along with verbal feedback from Thomas International psychologists.

4.3. Participants

4.3.1 Part I

The current sample (n=650) was collected from Beijing Martinsen Training & Consulting Co., Ltd., in the Beijing Office, 正东国际大厦 A 座 25I (Form CPPA25 series) and 东城区东湖别墅 C 栋 7 层 (Form CPPAC7 series); the data is dated between 2004-2007. Within this study, the researchers had n=373 CPPA25 series forms, and n=390 CPPAC7 (see Tables 4.1, 4.2, 4.3).

For the purposes of avoiding contamination from external variables, the English Form (a direct back translation from Chinese CPPA25/C7 to English, n=212) was excluded from this study. After excluding problematic forms and duplication, the final sample size was 650. For the purposes of this study, the sample size (n=650) exceeds the rule of thumb (Henry, 1990) at the ratio of 5 participants to one item (96x5=480).

4.3.1.1 Participant description

The majority of the participants were male (49.7%) (see Table 4.1). However, educational details are unclear because the old form did not request participants to specify this information (94.5% unmarked) (see Table 4.2). Most of the candidates are from middle management (35.2%), followed by sales (19.8%), and office related professions (6.5%) (see Table 4.3). The majority are from business (22.6%) and medical related fields (22.5%), mostly from a well-known international medical product sales department (see Table 4.5). Also, 60 participants (9.2%) are from the human resource field. In terms of ethnicity, this is a relatively homogenous group, in which all individuals are Eastern Asian, and belong to the Chinese language group. However, differences between dialects are unspecified via the current biographical form.

4.3.1.2 Sampling frame: original form

Table 4.1 Gender demography of Part I (Original form: A25I & C7)

		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	Missing	60	9.2	9.2	9.2
	Female	267	41.1	41.1	50.3
	Male	323	49.7	49.7	100.0
	Total	650	100.0	100.0	

Table 4.2 Education demography of Part I (Original form: A25I & C7)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Unspecified	614	94.5	94.5	94.5
	Diploma 大專	3	.5	.5	94.9
	Degree 大學本科	27	4.2	4.2	99.1
	Master or above 碩士及以上	6	.9	.9	100.0
	Total	650	100.0	100.0	

Table 4.3 Current/ last position demography of Part I (Original form: A25I & C7)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	**Unspecified**	**221**	**34.0**	**34.0**	**34.0**
	中層主管 Middle management	229	35.2	35.2	69.2
	公司高層決策主管 Executive management	5	.8	.8	70.0
	專業技朮人員 Technical Specialist	19	2.9	2.9	72.9
	業務員 Sales	129	19.8	19.8	92.8
	學生 Student	1	.2	.2	92.9
	辦事員和有關人員 Office related	42	6.5	6.5	99.4
	顧問 Consulting position	4	.6	.6	100.0
	Total	650	100.0	100.0	

Table 4.4 Current/ involvement with human resource field, demography of Part I (original form: A25I & C7)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	590	90.8	90.8	90.8
	Yes	60	9.2	9.2	100.0
	Total	650	100.0	100.0	

Table 4.5 Professional field demography of Part I (Original form: A25I & C7)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Unspecified	157	24.2	24.2	24.2
	公檢法 Legal related	2	.3	.3	24.5
	文化、娛樂與體育業 Culture or entertainment	2	.3	.3	24.8
	水、電、氣供給業 Energy related	4	.6	.6	25.4
	交通、運輸業 Transportation	3	.5	.5	25.8
	金融、保險業、房地產業 Financial, insurance, and Estate	51	7.8	7.8	33.7
	建築業 Architectural	7	1.1	1.1	34.8
	科研、教育事業 Education or research	6	.9	.9	35.7
	商業、貿易 Business and trade	147	22.6	22.6	58.3
	新聞媒體與廣告業 Media and Advertising	1	.2	.2	58.5
	資訊、諮詢服務業 Business Consulting	35	5.4	5.4	63.8
	農林牧漁水利業 Framing	1	.2	.2	64.0
	電腦業與IT行業 Technology, Information technology, and Computer	82	12.6	12.6	76.6

		Frequency	Percent	Valid Percent	Cumulative Percent
	related.				
	製造業 Factory	5	.8	.8	77.4
	醫療 Medical related	146	22.5	22.5	99.8
	黨政管理機關 Governmental	1	.2	.2	100.0
	Total	650	100.0	100.0	

Test administered date: 01 APR 2004 ~ 01 NOV 2007

4.3.2 Part II

The second part of the sample was also collected from the Beijing office of Martinsen Training & Consulting Co., Ltd., from November 2007 to May 2008. This study used the amended form CPPA-LV2. This sample size is 307, which is smaller than the 480 sampling benchmark (Henry, 1990). Part II is used for comparison with the Part I results.

4.3.2.1 Participant description

The sample size of the second part is 307 participants. The relevant research was conducted from 03 November 2007 up to 20 May 2008. The majority of the sample is also male (60.3%) (see Table 4.6). In contrast with the first form, the second form included the education biographical entry. This sample group is composed of higher educational groups, ranging from diploma holders (4.9%), university degree holders (26.1%) to Master's or above (16.3%) (see Table 4.7). The majority of the sample are from middle management (22.5%), followed by office related professions (7.2%), technical specialists (5.9%), and business consultants (4.2%) (see Table 4.8).

Eleven point seven percent (11.7%) of the participants work in human resource related fields (see Table 4.9). In terms of job categories, the majority came from financial, insurance, and real estate fields (29%), followed by technology, information technology, and computer related fields (20.5%), and education or research fields (11.1%) (see Table 4.10). This is also an homogenous group; all of the individuals are Eastern Asian, and belong to the Chinese language group. However, the difference between dialects is unspecified in the current biographical form.

4.3.2.2 *Sampling frame: LV2 form*

Table 4.6 Gender demography of Part II (LV2 form)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Missing	3	1.0	1.0	1.0
	Female	119	38.8	38.8	39.7
	Male	185	60.3	60.3	100.0
	Total	307	100.0	100.0	

Table 4.7 Education demography of Part II (LV2 form)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Unspecified	160	52.1	52.1	52.1
	Junior High 初中	1	.3	.3	52.4
	Senior High 高中	1	.3	.3	52.7
	Diploma 大專	15	4.9	4.9	57.6
	Degree 大學本科	80	26.1	26.1	83.7
	Master or above 碩士及以上	50	16.3	16.3	100.0
	Total	307	100.0	100.0	

Table 4.8 Current/ last position demography of Part II (LV2 form)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Unspecified	170	55.4	55.4	55.4
	中層主管 Middle management	69	22.5	22.5	77.9
	公司高層決策主管 Executive management	4	1.3	1.3	79.2
	服務業人員 Service	4	1.3	1.3	80.5
	基層員工, 生產、運輸設備操作人員及有關人員 Blue-collar (operational, produced, assembling, and logistic)	2	.7	.7	81.1
	專業技朮人員 Technical specialist	18	5.9	5.9	87.0
	業務員 Sales	5	1.6	1.6	88.6
	辦事員和有關人員/ Office related	22	7.2	7.2	95.8
	顧問 Business consulting	13	4.2	4.2	100.0
	Total	307	100.0	100.0	

Table 4.9 Current/ involvement with human resource field, demography of Part II (LV2 form)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	271	88.3	88.3	88.3
	Yes	36	11.7	11.7	100.0
	Total	307	100.0	100.0	

Table 4.10 Professional field demography of Part II (LV2 form)

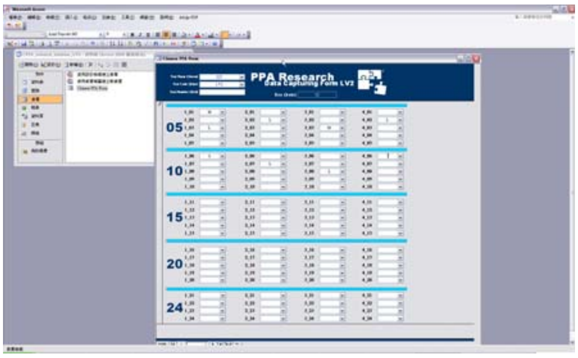
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Unspecified	87	28.3	28.3	28.3
	文化、娛樂與體育業 Culture or entertainment	1	.3	.3	28.7
	交通、運輸業 Transportation	1	.3	.3	29.0
	其他 Other	2	.7	.7	29.6
	金融、保險業、房地產業 Financial, insurance, and estate	89	29.0	29.0	58.6
	科研、教育事業 Education or research	34	11.1	11.1	69.7
	商業、貿易 Business and trade	11	3.6	3.6	73.3
	採掘、礦業 Mining	1	.3	.3	73.6
	新聞媒體與廣告業 Media and Advertising	1	.3	.3	73.9
	資訊、諮詢服務業 Business Consulting	4	1.3	1.3	75.2
	電腦業與IT行業 Technology, Information technology, and Computer related	63	20.5	20.5	95.8
	製造業 Factory	7	2.3	2.3	98.0
	黨政管理機關 Governmental	6	2.0	2.0	100.0
	Total	307	100.0	100.0	

Test administered date. 03 NOV 2007 ~ 20 MAY 2008

4.4. Materials, process, and analysis method in detail

4.4.1 Part I and II: Data capturing

The sample was collected through a standard psychometric administrative method of applying PPA in the Beijing Area and the data were shipped to South Africa in paper form. Microsoft Access was used for data capturing, and the interface is designed by the author for the PPA research specifically (see Figure 4.2, 4.3)



Note: An access data capturing interfaced is designed to standardise procedure

Figure 4.2 Access interface for data capturing

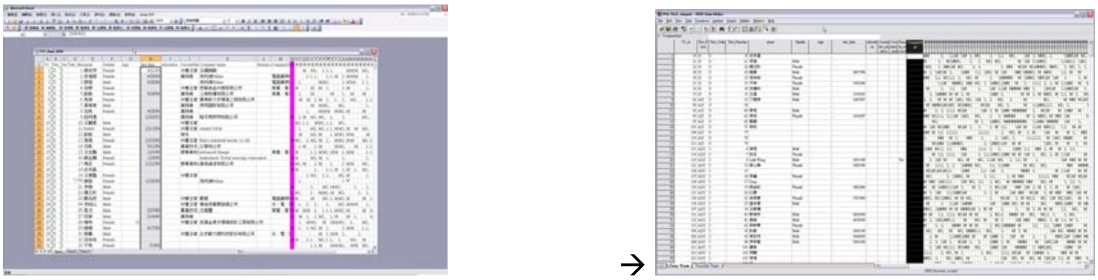
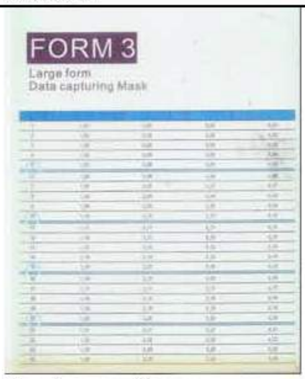


Figure 4.3 Data exported into excel then convert into SPSS

The author also adheres to the double-blind principle by letting two psychology master research assistants (non-Chinese speaking) perform data capturing and data analysis. The 'data capturing masks' (7 types of forms) are created by the author to avoid human error. Also, random checking is applied to assure the accuracy of data.

Table 4.11 Transparency mask to capture data

Mask detail

Form 2	Form 3	Form 4
		
For A251 & C7	For larger form	For thin form

Note: Forms are used to ensure the quality of the data

4.4.2 Double-blind: Capturing and analysis

Two psychology master research assistants (non-Chinese speaking) performed data capturing and data analysis. The ‘data capturing masks’ (7 types of forms, see Tables 4.12, 4.13, 4.14) are created by the author to avoid human error. The standardised procedures are as follows. use the capturing mask (form 2, 3, 4, see Table 4.11 above) to capture Chinese PPA forms (A25I_1, A25I_2, A25I_3, C7_1, C7_2, C7_3, 25, A25, and others) in the Microsoft Access database. Access database then exported into Microsoft Excel spreadsheet and later converted into the SPSS database for analysis (see Figures 4.2, 4.3). In addition, the statistical equivalence of eight different forms has been tested through analysis of variance and various *post hoc* tests. The numbers of data received from each form are reported in Table 4.15.

Table 4.12 Forms used by Chinese office – A25

Detail: data capturing form: Thumbnails for types of form

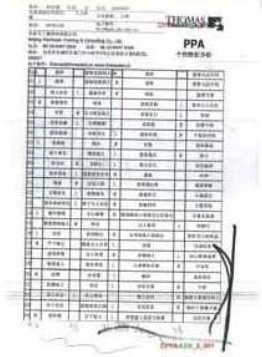
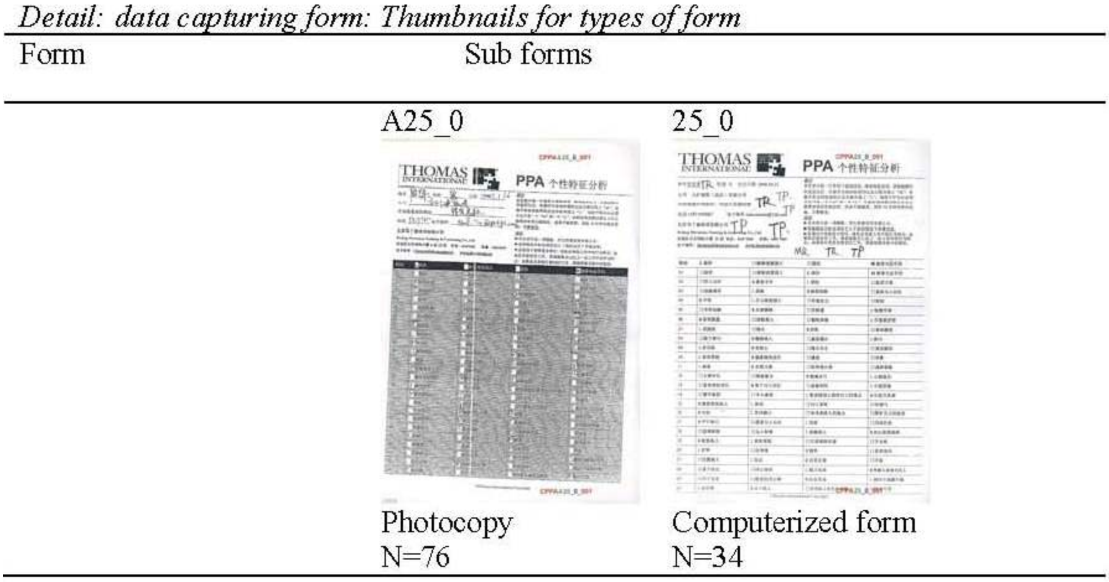
Form	Sub forms		
A251	A251_1	A251_2	A251_3
	 Original form N=121	 Photocopy N= 82	 Computerized form N=41

Table 4.13 Forms used by Chinese office – C7

Detail: data capturing form: Thumbnails for types of form

Form	Sub forms		
C7	C7_1	C7_2	C7_3
	 Original form N=122	 Photocopy n=70	 Computerized form N=90

Table 4.14 Forms used by Chinese office - 25



Detail: data capturing form: Thumbnails for types of form

Form	Sub forms	
	A25_0 Photocopy N=76	25_0 Computerized form N=34

Table 4.15 Test form in this study

		Test form number				Total
		0	1	2	3	
Test Form	Unspecified	0	0	0	11	11
	25	34	0	0	3	37
	A25	76	0	0	0	76
	A25I	0	121	82	41	244
	C7	0	122	70	90	282
Total		110	243	152	145	650

4.4.3 Equivalence of forms

Although all forms are textually equivalent, they are administered under different conditions (paper or Excel form), and a ‘between subjects’ ANOVA with PPA constructs as dependent variable and types of forms as an independent variable was therefore conducted. Hochberg's homogenous sub-test was used due to the unequal sample sizes.

The CI and CII constructs showed minor differences among different forms, CI: *F* (8, 641) = 2.443, *p*=.013; CII: *F*(8, 641) = 2.532, *p*=.010 (see Table 4.16, on the next page).

However, no such differences were shown in the homogeneous sub-test (see Tables 4.17, 4.18). Only the CII constructs showed relatively lower scores for the C7-3 form on C (see Table 4.18). This might be due to the fact that the C7-3 form was mainly administered to respondents from the sales department of a pharmaceutical company. The researcher therefore cannot reject the null hypothesis, which states no one form is different from one another significantly.

Table 4.16 One-way ANOVA comparison of PPA constructs between different forms

		Sum of Squares	df	Mean Square	F	Sig.
DI in percentage	Between Groups	.292	8	.037	1.255	.265
	Within Groups	18.653	641	.029		
	Total	18.945	649			
II in percentage	Between Groups	.618	8	.077	1.870	.062
	Within Groups	26.501	641	.041		
	Total	27.120	649			
SI in percentage	Between Groups	.177	8	.022	.998	.436
	Within Groups	14.248	641	.022		
	Total	14.425	649			
CI in percentage	Between Groups	.664	8	.083	2.443	.013**
	Within Groups	21.783	641	.034		
	Total	22.448	649			
DII in percentage	Between Groups	.196	8	.024	1.371	.206
	Within Groups	11.435	641	.018		
	Total	11.630	649			
III in percentage	Between Groups	.282	8	.035	.866	.545
	Within Groups	26.103	641	.041		
	Total	26.385	649			
SII in percentage	Between Groups	.120	8	.015	.579	.796
	Within Groups	16.667	641	.026		
	Total	16.788	649			
CII in percentage	Between Groups	.402	8	.050	2.532	.010**
	Within Groups	12.731	641	.020		
	Total	13.133	649			
D in percentage	Between Groups	.169	8	.021	1.210	.290
	Within Groups	11.202	641	.017		
	Total	11.371	649			

		Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
I in percentage	Between Groups	.365	8	.046	1.387	.199
	Within Groups	21.073	641	.033		
	Total	21.438	649			
S in percentage	Between Groups	.128	8	.016	.669	.719
	Within Groups	15.298	641	.024		
	Total	15.426	649			
C in percentage	Between Groups	.136	8	.017	.900	.516
	Within Groups	12.120	641	.019		
	Total	12.256	649			

**The ANOVA is significant in <0.05 level

Table 4.17 Homogenous sub test: CI in percentage

	PPA form types	N	Subset for Alpha = .05
			1
Hochberg (a,b)	A25-0	76	.4706
	C7-1	122	.4930
	A25I-2	82	.4953
	A25I-1	121	.5056
	C7-2	70	.5196
	A25I-3	41	.5203
	C7-3	90	.5681
	Others	14	.5712
	25-0	34	.5727
	Sig.		.239

Means for groups in homogeneous subsets are displayed.

a Uses Harmonic Mean Sample Size = 46.767.

b The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Table 4.18 Homogenous sub test: CII in percentage

	PPA form types	N	Subset for Alpha = .05	
			1	2
Hochberg (a,b)	C7-3	90	.4985	
	C7-2	70	.5329	.5329
	C7-1	122	.5485	.5485
	A25I-3	41	.5544	.5544
	25-0	34	.5576	.5576
	A25I-1	121	.5656	.5656
	A25-0	76	.5670	.5670
	A25I-2	82	.5702	.5702
	Others	14		.6128
	Sig.		.399	.203

Means for groups in homogeneous subsets are displayed.

a Uses Harmonic Mean Sample Size = 46.767.

b The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

4.5. Forced choice question item analysis

4.5.1 Internal Consistency, Cronbach's Alpha

Cronbach's Alpha was used to explore the internal consistency of the various forms. The FC result was converted to ordinal scales for reliability of statistics. This analysis was conducted under the preconception that the method may not be easily applicable to the forced choice scale. The aim was to explore the rationale behind the failure of internal consistency in FC, and to use it as a framework to explain the possible threats within the other classical test theory (CTT) methods in forced choice (FC) psychometrics, specifically for the PPA.

Cronbach's Alpha is potentially problematic in forced choice psychometrics (Baron, 1996; Bartram, 2007; Hicks, 1970; Martinussen, Richardsen, & Varum, 2001; McCloy, Heggestad, & Reeve, 2005). Baron (1996) argues that forced choice items could result in extremely high, as well as low, reliability.

The extreme nature of Alpha is due to the unavoidable *odd or artificial correlation* found among FC items. In the Likert form, t four mutually exclusive sub-items would provide three possible options ('Most', 'Least'', and 'Blank'). This setting gives each option a probability of being selected as 'M' or 'L' of exactly 33.3333% (1/3), and if all four items are used, there is a total selection combination of 3^4=81.

In contrast, in the FC format, the probability of such selection would be changed (2 items within the set would be left out, and the total combinations would be ${}_4P_2 = \frac{4!}{(4-2)!} = 12$).

Moreover, there would be an equal probability of 25% (3/12) of being selected as the 'Most' item in the tetrad, as well as a 25% (3/12) probability of being selected as the 'Least' item. This would leave exactly a 50% (6/12) probability of being left out as a blank item. This *structured probability* would lead to inaccurate (or uninterpretable results of Cronbach's Alpha (Brown & Harvey, 2003; Martinussen et al., 2001; Yu, 2008).

The other important reason that FC item sets function differently from the Likert scale is because all sub items within an item group (namely *tetrad*) would compete for selection with one another, resulting in *item dynamics*. Items within a group would compete for their popularity and the winning item would be selected, while the others would be ignored.

In terms of the PPA, this would mean that some items would be more likely to be marked as 'M' (Most), and some items would be relatively more likely to be marked as 'L' (Least). Thus even when four items have equal 'popularity' in a Likert-scale research study, it would be a different story when they are placed within an FC tetrad.

Furthermore, as is the case with most of the psychometric factors, the *item dynamics* are also under the influence of cultural and historical effects.

In short, due to the effect of *artificial correlation* and *item dynamics*, an internal consistency statistic such as Cronbach's Alpha, is not appropriate for FC psychometrics. This point can be illustrated mathematically.

4.5.2 Internal consistency: Mathematical expression.

Cronbach's Alpha in SPSS is a raw form of Cronbach's Alpha which is defined as (UCLA.ATS, 2006; Yu, 2008).

$$\alpha = \left(\frac{N}{N-1}\right)\left(1 - \frac{\Sigma_{i=1}^{N}\sigma_{Y_i}^{2}}{\sigma_{\mathrm{X}}^{2}}\right) \qquad (4\text{-}1)$$

N = Number of components

$\sigma_{Y_i}^{2}$ = Variance of each item

σ_{X}^{2} = Variance of the observed total test scores

In the raw form of Cronbach's Alpha, reliability is measured in terms of the ratio of true score variance (variance of the observed total test score) to observed score variance (variance of the component i) (Yu, 2008). Yu (2008) also explains that raw Cronbach's Alpha is a measurement of item correlation. The stronger the items are interrelated, the more likely the test is to be consistent.

However, this would make FC psychometric unsuitable for Cronbach's Alpha. An FC item's response is influenced by *item dynamics* and *artificial correlation* making the fundamental element of correlation, covariance also *artificial*.

This implies that the covariance results from pre-structured item dynamics and artificial correlations, but not via the constructs. In other words, a FC psychometric does not need a functional construct to create an 'internal consistent' (or reliable) construct, which poses a threat to the use of internal consistency as the evaluating standard in FC psychometrics.

Mathematically, $\sigma_{\mathrm{X}}^2 = \sigma_{\mathrm{Total}}^2$ is not only a function of the item variance. According to (Thompson, 2003), $\sigma_{\mathrm{Total}}^2$ can be expressed as following.

$$\sigma_{\mathrm{Total}}^2 = \sum \sigma_k^2 + \left[\sum COV_{ij} (\text{for } i{<}j) * 2 \right] \qquad (4\text{-}2)$$

$\sigma_{\mathrm{Total}}^2 =$ Sum of all item variance and covariance = variance of total score

Thompson (2003) suggested that the fundamental element of the correlation formula is COV.

$$\because r_{xy} = \frac{COV_{xy}}{\sigma_x \sigma_y} \quad \therefore COV_{xy} = r_{xy}(\sigma_x \sigma_y) \qquad (4\text{-}3)$$

The formulae (4-2, 4-3) indicate that both Cronbach's Alpha and correlation share COV (covariance). Thompson (2003) further suggests that the raw form of Cronbach's Alpha cannot operate with low correlation, and that high item correlation would also lead to high Alpha coefficients.

4.5.3 Construct correlation

Pearson's product moment coefficient ($\rho_{X,Y}$) has been selected for correlation analysis of the PPA sub constructs. The correlation analysis is not done on individual item scores, but on the total scores of each construct. The total score is an 'interval' expression of the FC score. The final product of PPA is 12 constructs.

These are DISC in Working mask, Pressure mask, and Self mask. The 12 constructs are made by different sets of constructs and response types. Some constructs only receive the reaction from the negative response, some only positive and some items are shared.

Pearson's Correlation is used to explore the relationship within the DISC construct within and across three masks. However, these results should be cautiously interpreted in the light of the issues of *item dynamics* and *artificial correlation*.

4.5.4 Item difficulty (P) and item discrimination index (D)

Classical Test Theory item analysis was conducted using (Yu, 2002) software TESTER for Windows 2.0. The summary of the outcome is presented below.

4.5.4.1 Item difficulty index (P) (popularity index)

The difficulty index (P) is defined as the difficulty of the item. The P value ranges from 1 to 0. A large P value means that the test is easy and most respondents would answer it correctly. Alternatively, when the value is approaching 0, it means the item is very difficult (Allen & Yen, 1979; Yu, 2002).

Because the PPA is not a cognitive test, the idea of 'difficulty' is not applicable, and the index is therefore operationalised as 'popularity' in an affective test. A higher value would imply that an item is 'socially popular' and *vice versa for low values*. Past researchers suggest that values of .40-.70 would be acceptable (Ahmanan & Glock, 1981). Some also suggested .40-.80 as standard for multiple choices, and .55-.85 for true-false questions (Chase, 1978). As a rule of thumb, Yu (2008) suggests the benchmark for (P) is 0.5 meaning that at least half of the respondents would answer it correctly.

Due to the reason that the selection probability of the item is 0.25 (3/12) (refer to the Cronbach's Alpha section for detail).

This would make PPA FC more similar to the multiple-choice format than a true and false choice format. Therefore, this study does not use the common 0.5, but uses Chase's (1978) 0.4 as P value's benchmark instead., as shown below.

Mathematical definition of (P) (Yu, 2002).

$P_i = \frac{P_{iH} + P_{iL}}{2}$ Average probability of higher group and lower group

$i = 1, 2, ..., n$ Item

$P_{iH} = \frac{R_{iH}}{N_{iH}}$ Probability of correctness of higher group (first quartile) on item i

$P_{iL} = \frac{R_{iL}}{N_{iL}}$ Probability of correctness of lower group (fourth quartile) on item i

4.5.4.2 Item discrimination index (D)

The item discrimination index (D) is measured by the difference between probability of the high group (first quartile) and the low group (fourth quartile). The larger the gap between the two, the higher the discrimination ability of the item.

Various definitions of 'high' and 'low' have been given by researchers. Some use the top 27% as the high group (Sim & Rasiah, 2006), while others define the first quartile (top 25%) as the high group (Yu, 2002). This study uses Yu's (2008) 25% (first and fourth quartile) model.

The discrimination index (D) ranges from +1.00 to -1.00 (see p. 105). A positive value indicates that the high group achieved higher response rates than the lower group, which is the correct condition. A negative D value means that some items appear to depict incorrect constructs, and that the low group had higher responses rate than the high group.

The negative value is an indication that the item is associated with other constructs r than the intended one (Yu, 2002).

Ebel and Frisbie (1991) suggest a standard for item discrimination index (D) (refer to Table 4.19). Yu (2008) also suggests that the generally acceptable value for (D) is .25. This study uses Yu's (2008) suggestion.

Table 4.19 Interpretation of item discrimination index (D) (Ebel & Frisbie, 1991)

Item Discrimination Index	Interpretation
.40 or above	Very good
.30~.39	Good, but adjustment needed
.20~.29	Acceptable, but adjustment needed
below .19	Poor, delete or adjustment

Mathematical definition of (D) (Yu, 2002).

$$D_i = P_{iH} - P_{iL}$$

$P_{iH} = \frac{R_{iH}}{N_{iH}}$ Indicates probability of correctness of higher group (first quartile) on item i

$P_{iL} = \frac{R_{iL}}{N_{iL}}$ Indicates probability of correctness of lower group (forth quartile) on item i

4.6. PPA DISC construct and scoring

4.6.1 PPA Three Masks

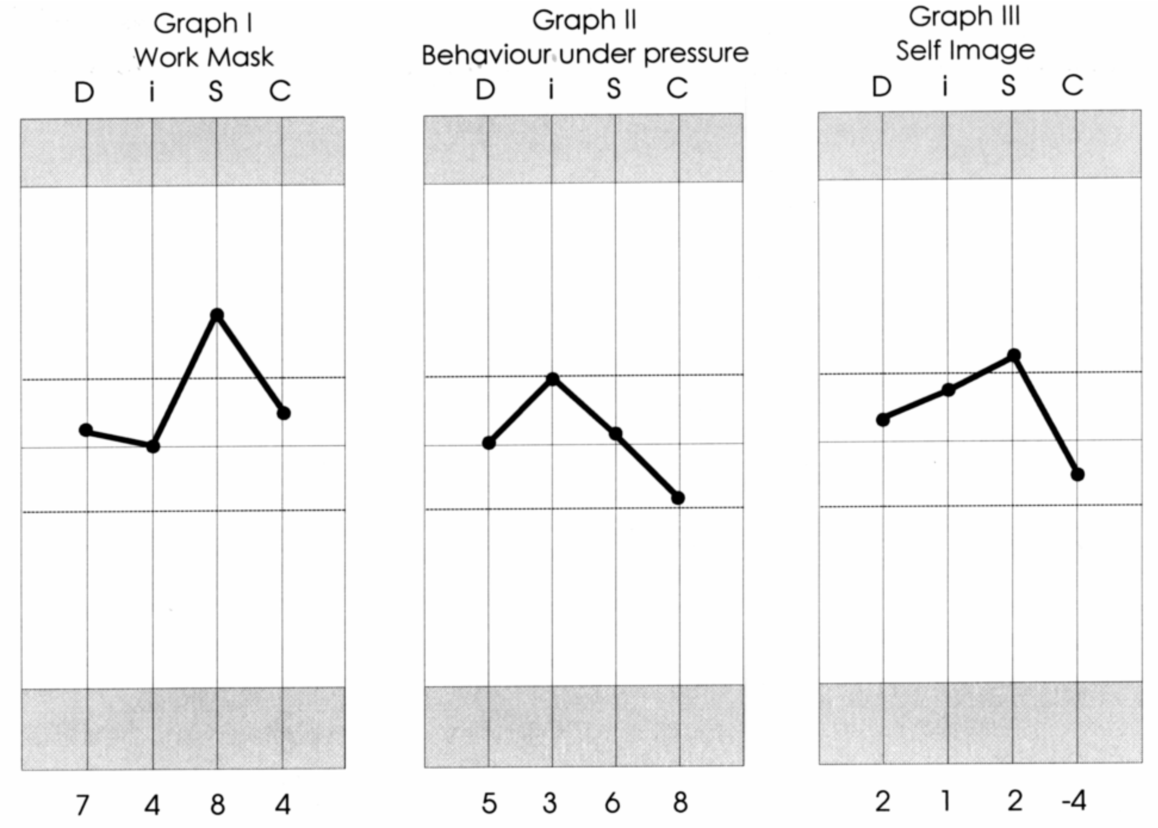


Figure 4.4 PPA Three Masks (Irvine, 2003)

At the end of the PPA report, the DISC profiles are illustrated into three different 'masks'. They are. 'Work Mask' (Graph I work mask, in the left), 'Pressure Mask' (Graph II Behaviour Under pressure, in the middle), and 'Self Mask' (Graph III Self image, in the right) (see Figure 4.4).

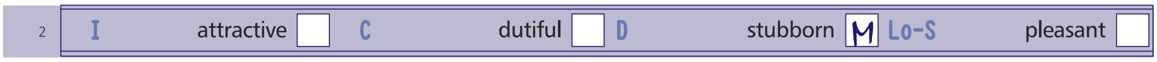
The 'Work Mask' represents how individuals like to mask in order to be successful; this is scored by the 'M' (most) mark. The 'Pressure Mask' represents the characters that would still remain under pressure conditions; this is scored by the 'L' (least) mark. The

‘Self Mask’ represents how individuals see themselves; it is scored by a combination of the work and pressure masks (Hendrickson, Undated/1958).

All three masks illustrate the DISC information into line graphs. Dots that are above the centre line are considered as ‘high’; when they are below the centre line, they are considered as ‘low’. For example, see above in Graph II; this individual’s pressure mask indicated high I, high D and high S, but low C. This represents that this individual is more likely to remain a high I person, would acceptable D and S, but low C during pressure condition. This profile is termed ISD (followed by the strength of the profile) in Thomas system. All three masks contain DISC profiles. However, they should be considered as three different types of DISC constructs. They are represented as. Graph I work mask (DI, II, SI, CI), Graph II Pressure mask (DII, III, SII, CII), and Graph III self image (D, I, S, C). The PPA system contains 12 constructs that were contributed by three masks.

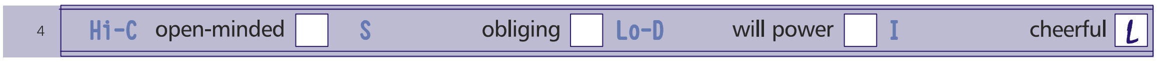
4.6.2 Scoring of DISC constructs with Hi Lo marker

PPA creates its raw score via ‘M’ and ‘L’ marked on each tetrad. If a D item is marked M, one score would add to the D construct, and vice versa with other constructs. For example, if a D item on tetrad 2 (Stubborn) was marked as ‘M’, the *D-Work-mask* construct would add one raw score (see below).



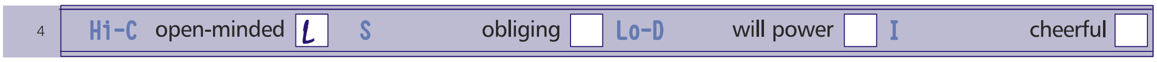
2	I	attractive	□	C	dutiful	□	D	stubborn	M	Lo-S	pleasant	□

Another example; if an I-item in tetrad 4 (Cheerful) was marked as ‘L’, the *I-pressure-mask* profile would add one raw score (see below).



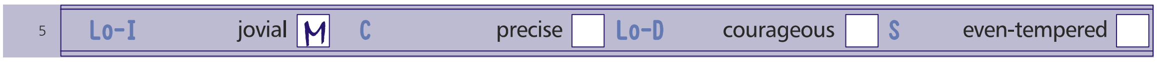
4	Hi-C	open-minded	□	S	obliging	□	Lo-D	will power	□	I	cheerful	L

However, not all marks are scored. Thus, an ‘M’ mark on ‘Lo- terms’ is not scored, and an ‘L’ mark on ‘Hi-terms’ is also not scored. For example, marking ‘L’ on the tetrad 4, item ‘Open minded’ (Hi-C) would not generate any score (see below).



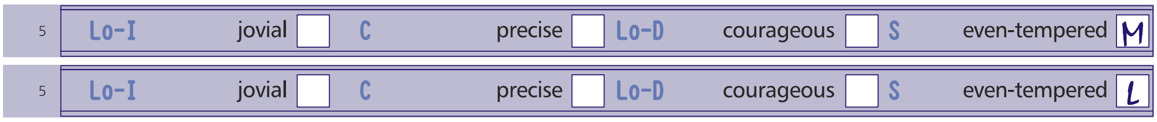
4	Hi-C	open-minded	L	S	obliging	□	Lo-D	will power	□	I	cheerful	□

Also, marking ‘M’ on tetrad 5, item ‘Jovial’ (Lo-I) would not generate any score either (see below).



5	Lo-I	jovial	M	C	precise	□	Lo-D	courageous	□	S	even-tempered	□

In tetrad 5, if an item that has no ‘Hi’ or ‘Lo’ scoring mark, such as ‘even-tempered’ (S) in tetrad 5, would be scored on both ‘M’ and ‘L’. The ‘*M*’ marked items would be scored in the ‘*work mask*’ construct. Alternatively, the ‘*L*’ marked item would be scored in ‘*Pressure mask*’ (see below).



5	Lo-I	jovial	□	C	precise	□	Lo-D	courageous	□	S	even-tempered	M
5	Lo-I	jovial	□	C	precise	□	Lo-D	courageous	□	S	even-tempered	L

The *pressure mask* infers the characteristic that the respondent is willing to sacrifice, or let go, during the pressure of a conflict condition. The raw responses are ranked in percentiles. The final *self mask* is derived from the *most mask* raw score minus the *least mask* raw score. Such raw scores are also ranked in percentile. Therefore, PPA would have four constructs (DISC) in three masks (work, pressure, self), which is 12 constructs in total (see equation 4-6).

$$
\begin{aligned}
&\text{Pressure mask} = \sum\left(\text{Least}_{\text{DISC}}\right) \\
&\text{Work mask} = \sum\left(\text{Most}_{\text{DISC}}\right) \\
&\text{Self mask} = \sum\left(\text{Most}_{\text{DISC}}\right) - \sum\left(\text{Least}_{\text{DISC}}\right)
\end{aligned}
\tag{4-6}
$$

In contrast to the traditional Likert-scale format, PPA measures an individual’s affective traits via 96 items. It measures three types of responses (M, L, and blank*)* and scores them separately, therefore PPA can be considered as three forms in one. The following figure (4.5) summarises the possible information that can be extracted from PPA.

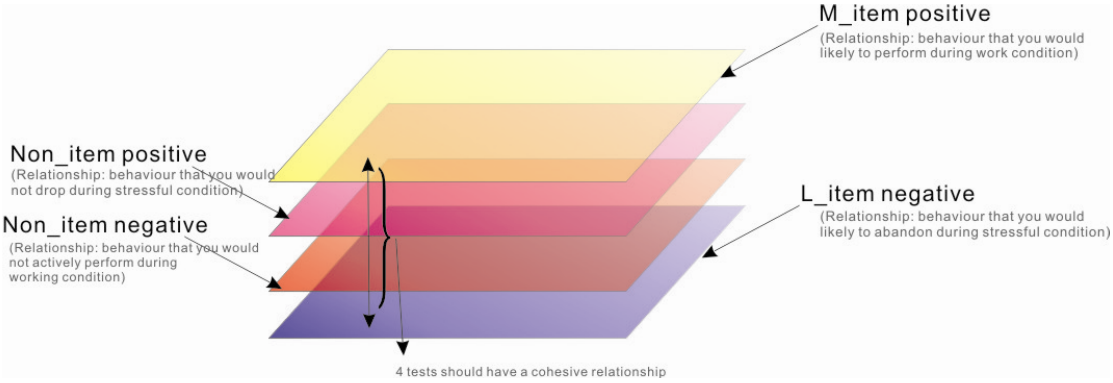

M_item positive
(Relationship: behaviour that you would likely to perform during work condition)
Non_item positive
(Relationship: behaviour that you would not drop during stressful condition)
Non_item negative
(Relationship: behaviour that you would not actively perform during working condition)
L_item negative
(Relationship: behaviour that you would likely to abandon during stressful condition)
4 tests should have a cohesive relationship

Figure 4.5 Possible information that can be extracted from PPA

The M (positive*)* responses can be defined as the image that individuals try to present to themselves. Therefore it is termed as *work mask*. The L (negative) responses can be defined as the image that individuals would prefer to retain within conflict situations, and therefore termed as *pressure mask*. A *blank* result can have both positive and negative implications, and is therefore considered as *undefined*.

4.7. PPA DISC interpretations

The PPA system uses the percentile rank as the aim of interpretation. Individuals who generate a more than a 50 percentile rank would be considered as ‘High’ on this specific construct. For example, high on DI (Dominance and Influence), low on SC (Submission and Compliance). This would categorise as ‘DI’ profile in the Thomas system. The system assigns profile explanations according to these profile categories.

4.8. IRT analysis, from basic to advanced methods

This study had gone through many different types of item response methods to find the most suitable method for PPA force-choice items. One method is the raw item characteristic curve (RICC) suggested by Allen and Yen (1979). RICC is a ‘simple scatter-line graph’ using probability of selection (percentage) as the Y-axis. The 12 PPA construct scores is the X axis.

The mathematical expression of RICC is.

$$X = \frac{cf_l}{N-1} \times 100\% \; ; \; Y = P(R_X) \qquad (4\text{-}7)$$

cf_l =Cumulative frequency of all scores lower than the score of interest

N=Number of total sample size.

R_X =Probability of the target response type of at particular point of percentile rank

The RICC results are later visually categorised in different groups. The following are the examples of RICC.

4.8.1 Positive relationship

When an M item (Yellow Curve) is moving upwards, it is referred to as a 'positive relationship' (see Figure 4.6). A positive relationship can be interpreted as an item that is functioning within the target construct (PAC=PEC). Therefore, it is positively discriminating amongst the respondents. When the respondents have a low percentile rank in the target construct, it would be less likely for them to mark the target item as 'M', and more likely that they would mark such an item as 'L', or 'blank'.

Alternatively when respondents obtain a higher percentile in the target construct, it would be more likely for them to mark this item as 'M', and less likely to mark it as 'L' or 'blank'. In this study it is assumed that if an item generates a positive relationship with the target construct, the item is measuring the target construct, and that it provides an appropriate scale for the target construct.

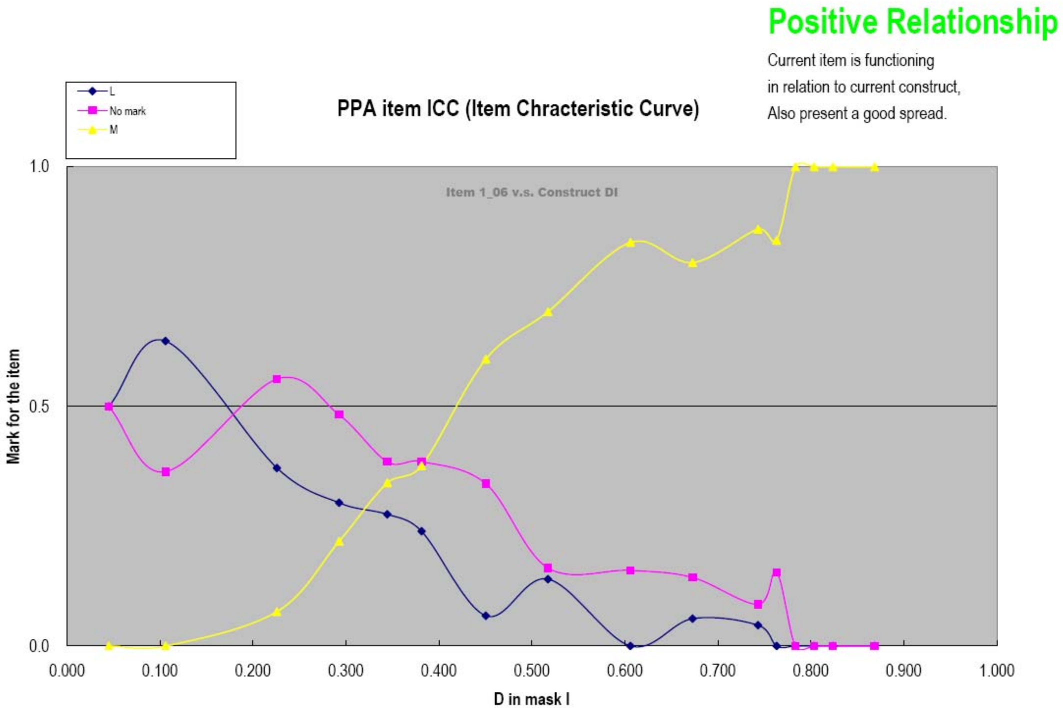
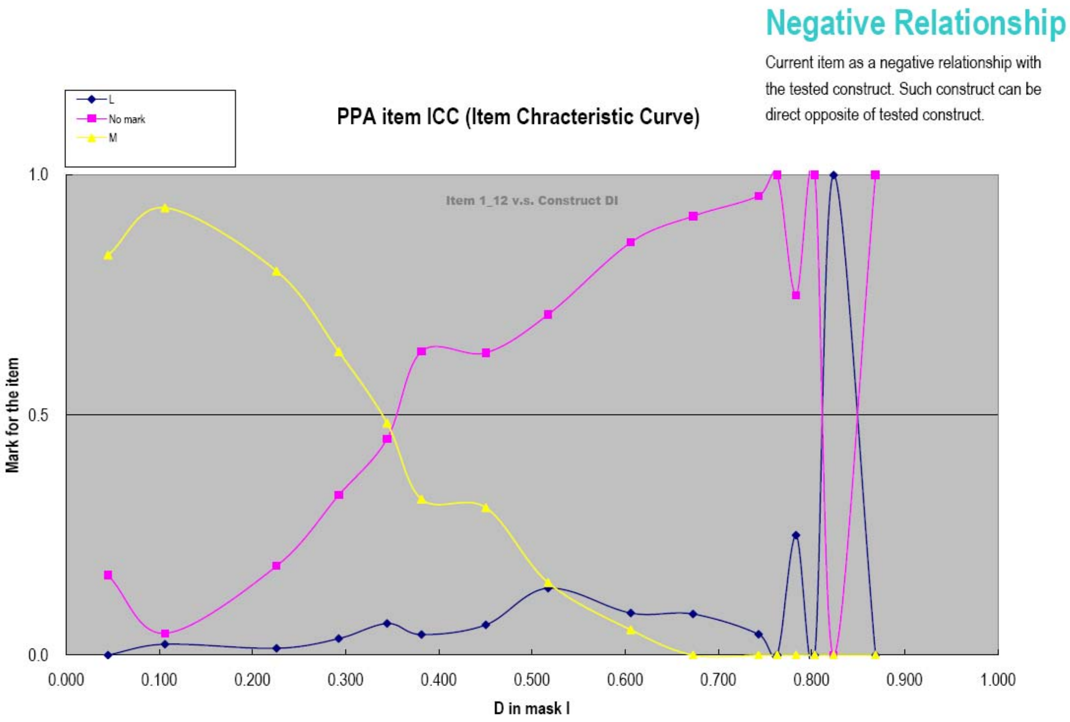


Figure 4.6 Sample Raw Item Characteristic Curve of positive relationship

4.8.2 Negative relationship

When the ‘M’ item is moving downwards it is considered to represent a negative relationship. A negative relationship is defined as an item that is negatively discriminating against the target construct. In such a case, the item does not measure the target construct, but seems to be measuring a different or opposite construct. This case is defined as respondents demonstrating a low percentile rank on the target construct. It is more likely that they would mark the targeting construct as ‘M’.

However, for respondents who are positioned in higher percentile rank of the target construct, the probability of marking the item as ‘M’ decreases, and it would also be more likely for them to mark the item as ‘L’ or ‘blank’ (see Figure 4.7).



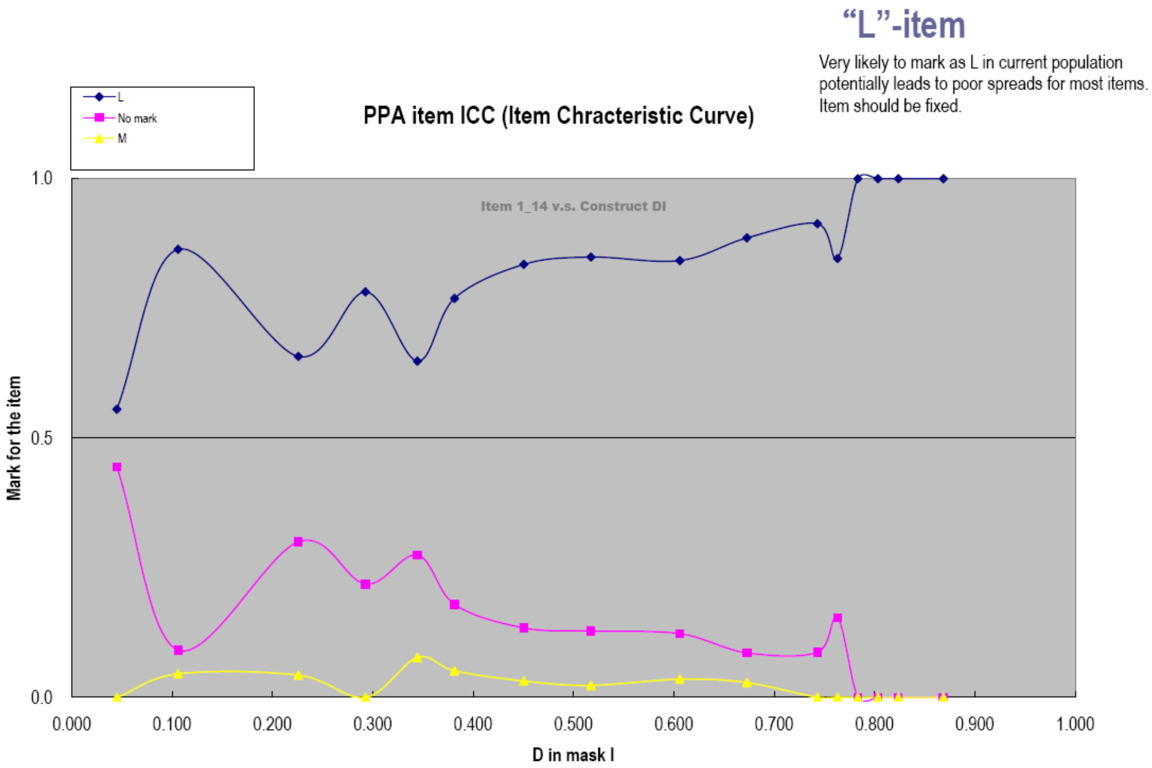
Negative Relationship

Current item as a negative relationship with the tested construct. Such construct can be direct opposite of tested construct.

Figure 4.7 Sample Raw Item Characteristic Curve of negative relationship

4.8.3 L item

When an item demonstrates a high negative response pattern in both high and low percentile ranks, it can be defined as an ‘L-item’. In such cases there is a high negative response pattern overall. This phenomenon could arise for many reasons, such as due to the general unpopularity of the item, unpopularity within the tetrad, or an extreme or a vague item (see Figure 4.8).



“L”-item

Very likely to mark as L in current population potentially leads to poor spreads for most items. Item should be fixed.

Figure 4.8 Sample Raw Item Characteristic Curve of L-item

4.8.3.1 L item as the ‘Hi’ item

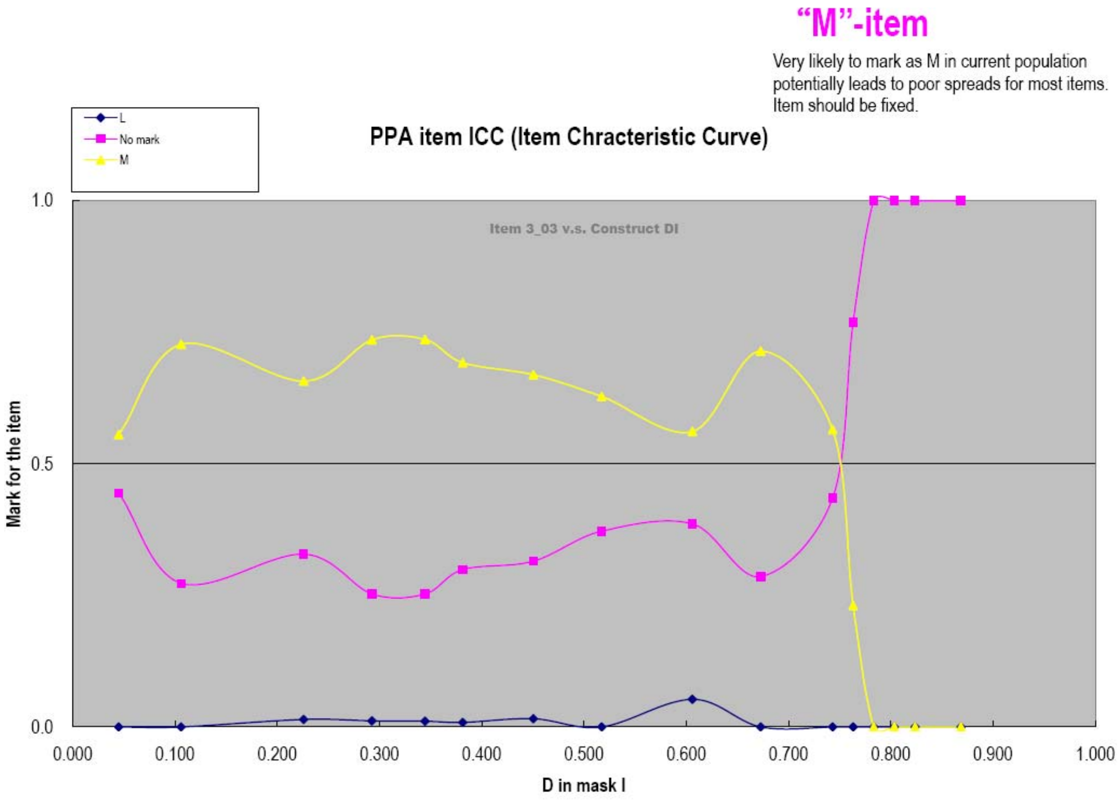
Due to high response in ‘L’, this type of item is very close to the definition of ‘Hi’ item, which means they are only scored when it is marked as ‘M’ (Irvine, 2003). The ‘L’ response would not be scored, due to the high occurrence of them. However, when a ‘Lo’ or normal item appears to be the L-item, some amendments of item or tetrad are required (for more detail on the Hi and Lo marker in PPA scoring system, please see 4.6.2 and Chapter 3 Area A.2).

4.8.4 M-item

When an item is marked as ‘M’ (Most) by respondents in both high and low percentile ranks, it is defined as an ‘M-item’. Such items have a high positive response pattern overall (see Figure 4.9).

This phenomenon could arise for many reasons, such as the general popularity of an item within the culture, popularity within the tetrad, or a common/easy to use item in current culture.

Figure 4.9 Sample Raw Item Characteristic Curve of M-item

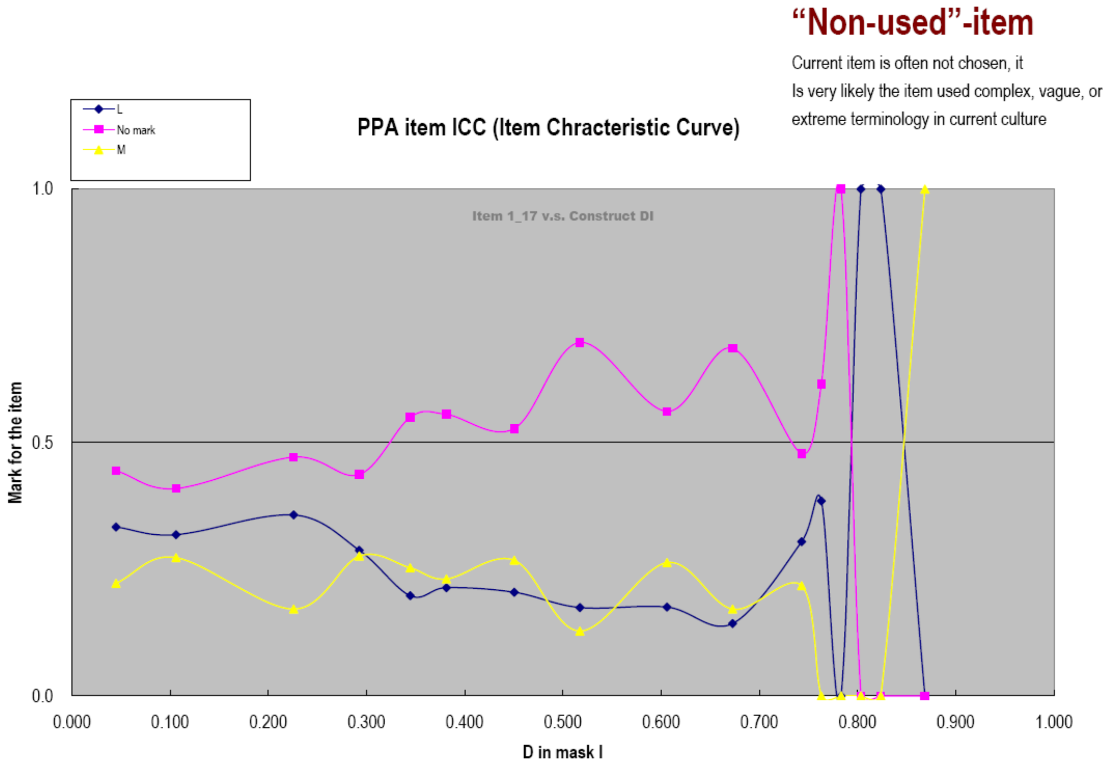


4.8.4.1 M item as the 'Lo' item

This type of item is scored as a 'Lo' item, which means they are only scored when marked as 'L' (Irvine, 2003). 'M' responses are not scored, due to their high occurrence. However, when a 'Hi' or normal item appears to be the M-item, some amendments of the item or tetrad is required (for more detail on the Hi and Lo marker in PPA scoring system, please see 4.6.2 and Chapter 3 Area A.2).

4.8.5 Non-used item

When an item demonstrates high ‘no response’ patterns in both high and low percentile ranks, it can be defined as a ‘non-used item’. In other words, it is allocated a high ‘blank’ response pattern overall (see Figure 4.10).



“Non-used”-item

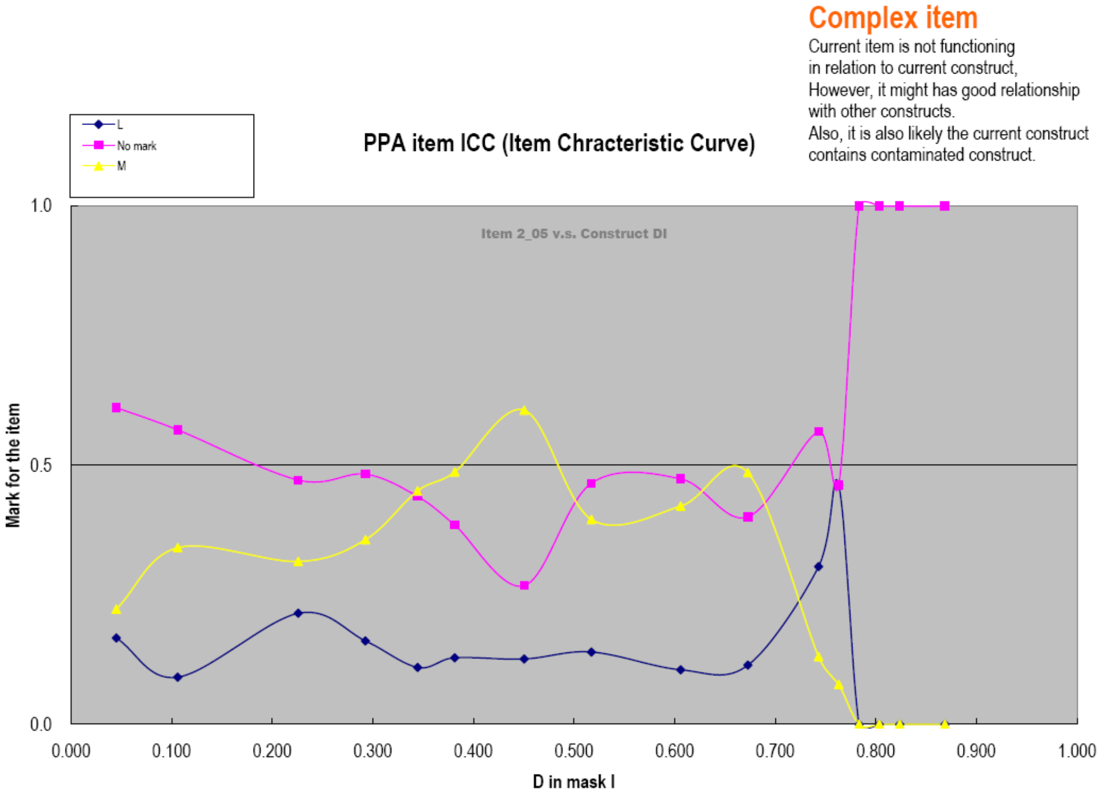
Current item is often not chosen, it
Is very likely the item used complex, vague, or
extreme terminology in current culture

Figure 4.10 Sample Raw Item Characteristic Curve of ‘non used’ item

This phenomenon could arise for several reasons. The population may consider such an item vague, difficult to understand, complex, difficult to respond to, or most candidates may judge another item within the tetrad to be a better choice. Such items (or the particular tetrad) are considered to be problematic and would require an amendment.

4.8.6 *Complex items*

When the ‘M’ items exhibit a normal distribution format, or go into a ‘swing form’ – up and down across all areas, they are defined as ‘complex’ items. The concept of ‘item swing’ is discussed in the next section. A complex item is suspected to contain vague, extreme, or difficult terminology for the target population.



Complex item
Current item is not functioning
in relation to current construct,
However, it might has good relationship
with other constructs.
Also, it is also likely the current construct
contains contaminated construct.

Figure 4.11 Sample Raw Item Characteristic Curve of complex item

A complex item can also be a sign of item contamination. This means that one or more items are not functioning within the tetrad. When an item appears to be a complex item it should be considered as problematic and should be amended.

4.8.7 Raw Item Characteristic Curve (RICC) difficulties

Interpretation difficulties: Item swing

One of interpretation difficulties associated with RICC is a phenomenon commonly observed in research, i.e. the ‘item swing’ effect (see Figure 4.13). However, normality of the data is assumed, as defined by Allen & Yen (1979), then most of the population would be located between 15%-ile to 80%-ile (see Figure 4.12). This range is equivalent to one standard deviation, which theoretically could include 68.26% of population. The upper and lower 15% (0.01%-14.99% and 84.99%-99.99%) would contain relatively smaller proportions of the population. Theoretically, this would be 15.87% of population in each side. The further away from the centre, the less the population.

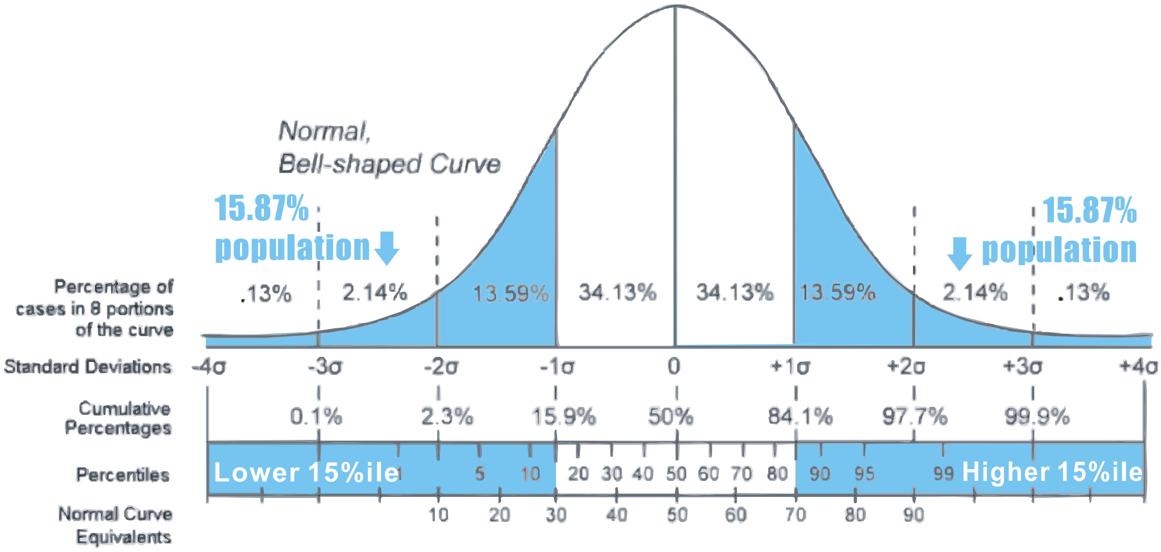
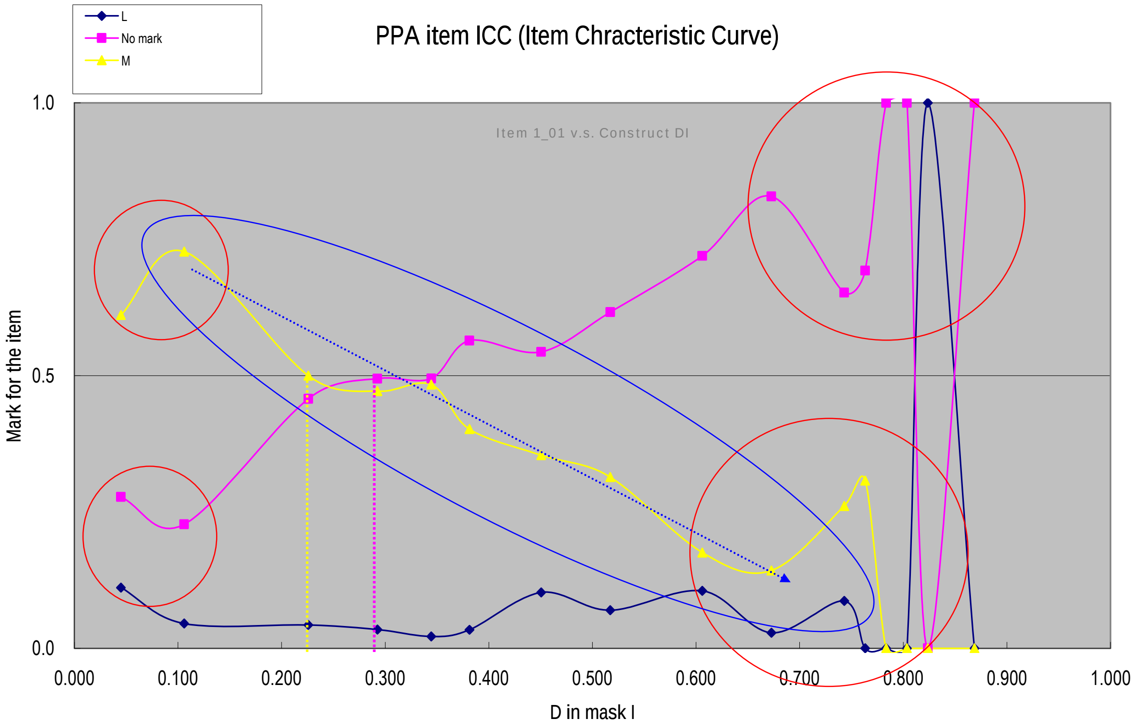


Figure 4.12 Standard normal distribution curve (Jooste, 2003)

When converting this concept to RICC, it means the higher and lower end of the percentile would have a smaller sample size. For example, in a sample size of 100, the sample size from each tail would be less than three individuals.

In such a small sample size, if one participant did not select the target item due to chance, the probability selection would drop from 100% to 66%. When all three of the candidates do not select the item, it would drop to 0%. If all three select the item, one would have 100%. When the data are approaching two tail ends, it is very likely the probability would swing quickly between 100% and 0%. This is due to the small sample size. This study terms it as 'item swing' (see Figure 4.13).

Item swing is a common phenomenon in a small sample size, and could lead to difficulty in interpreting the RICC because these swings probably act as outliers and change how one interprets the RICC.



Note: Within the red circles are the 'item swing's. The blue circle indicates the main direction of the curve

Figure 4.13 Sample Raw Item Characteristic Curve of item swing

4.8.7.1 Time challenge

Conducting RICC is also a highly time-consuming task, and there is no software designed to draw the graphs. In this study the RICCs were created by calculating the percentiles and response patterns per-percentile using SPSS. These results were then exported into Microsoft Excel for plotting. The RICC plots were then analysed manually and assigned to groups of six (negative, positive, L-item, M-item, non-use, and complex). After the graphs had been generated, they were manually captured in a construct summary table for analysis. The entire process took weeks. Therefore, alternative methods should be designed to speed up the process in future applications.

4.9. From RICC to Correlation Parameter Estimation (CPE)

A very simple method is used to generate the parameters. The least square regression analysis of the RICC raw data is used (see Figure 4.14). The slope of this regression line is use as the discrimination parameter A (see Figure 4.15).

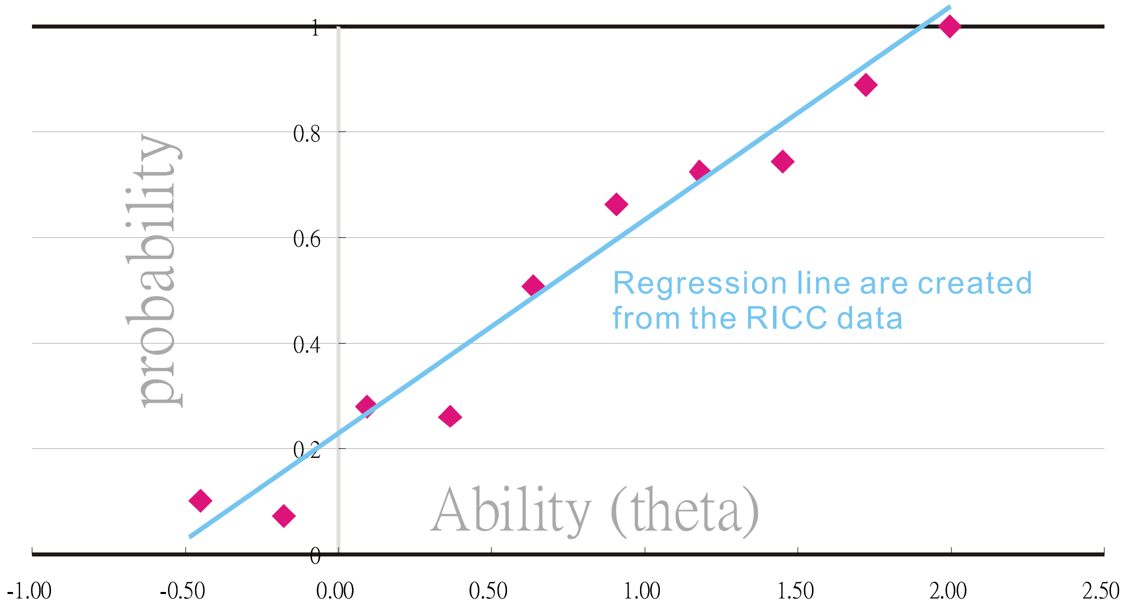


Figure 4.14 Generating regression line from the raw data

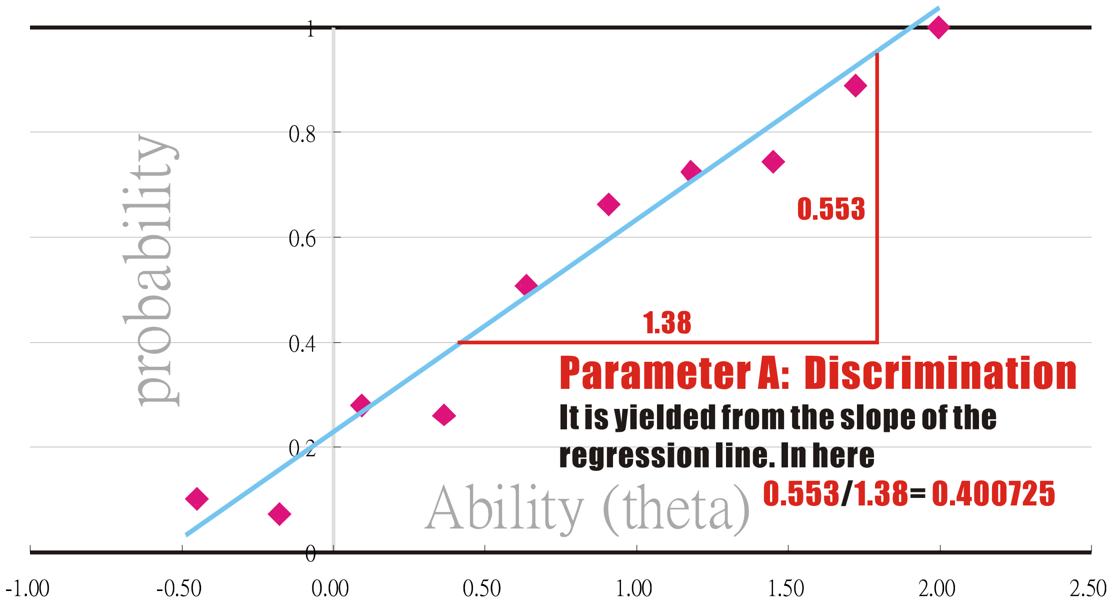


Figure 4.15 Generating the discrimination parameter A

The value of this line across the 0.5 of Y-axis is the difficulty parameter B (see Figure 4.16).

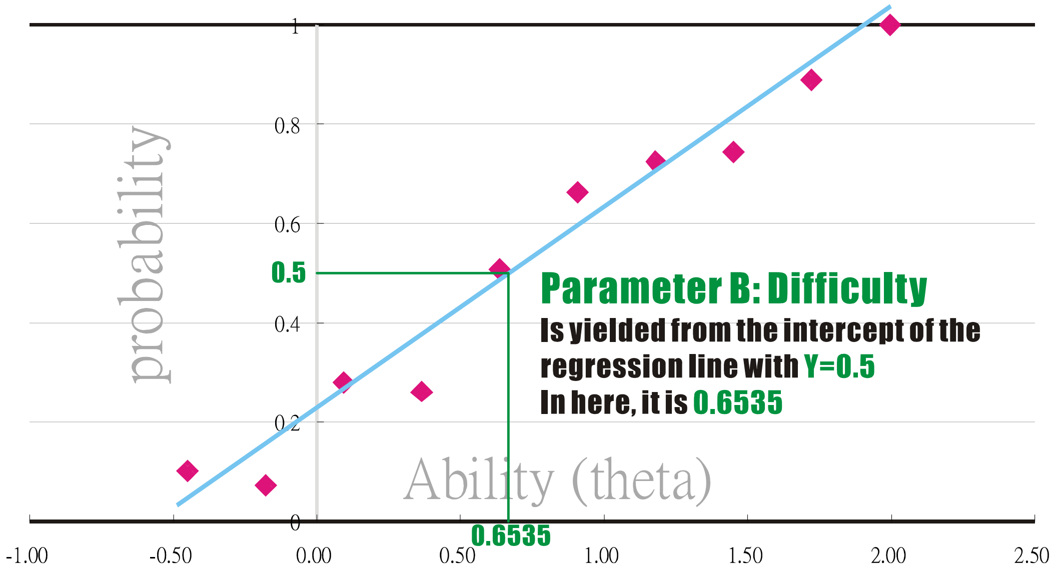


Figure 4.16 Generating the difficulty parameter B

The intercept of the regression line on the Y-axis, is used as the guessing parameter C (or the minimal data) (see Figure 4.17).

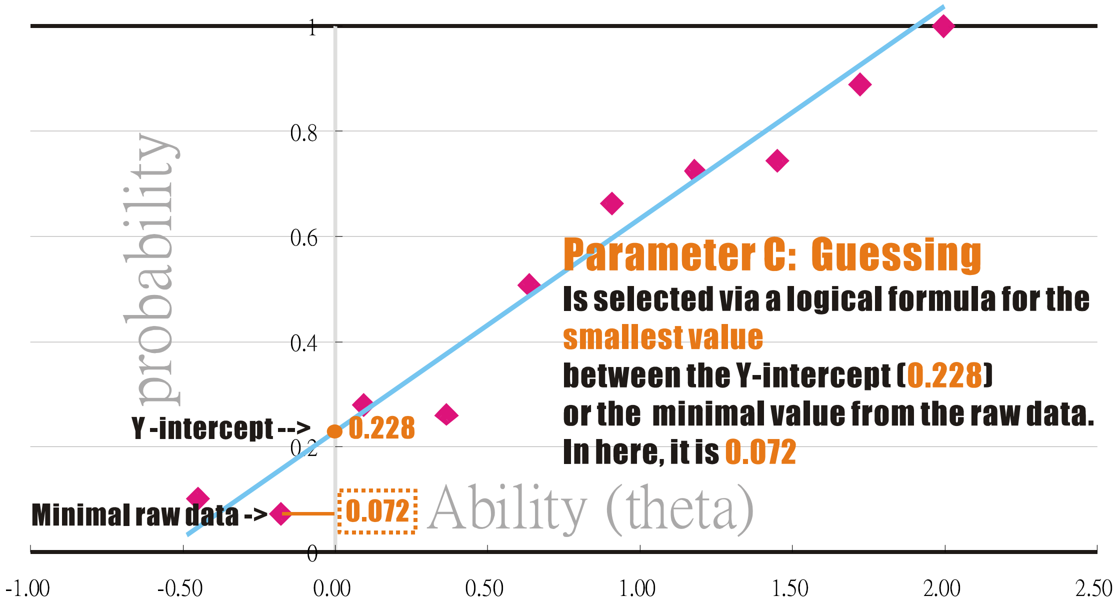


Figure 4.17 Generating guessing parameter C

The three parameters (A, B, and C) are plotted into the three-parameter logistic (3PL) formulae to generating the curve (see Figure 4.18). This is the Correlation Parameter Estimation (CPE) method that was designed to save time.

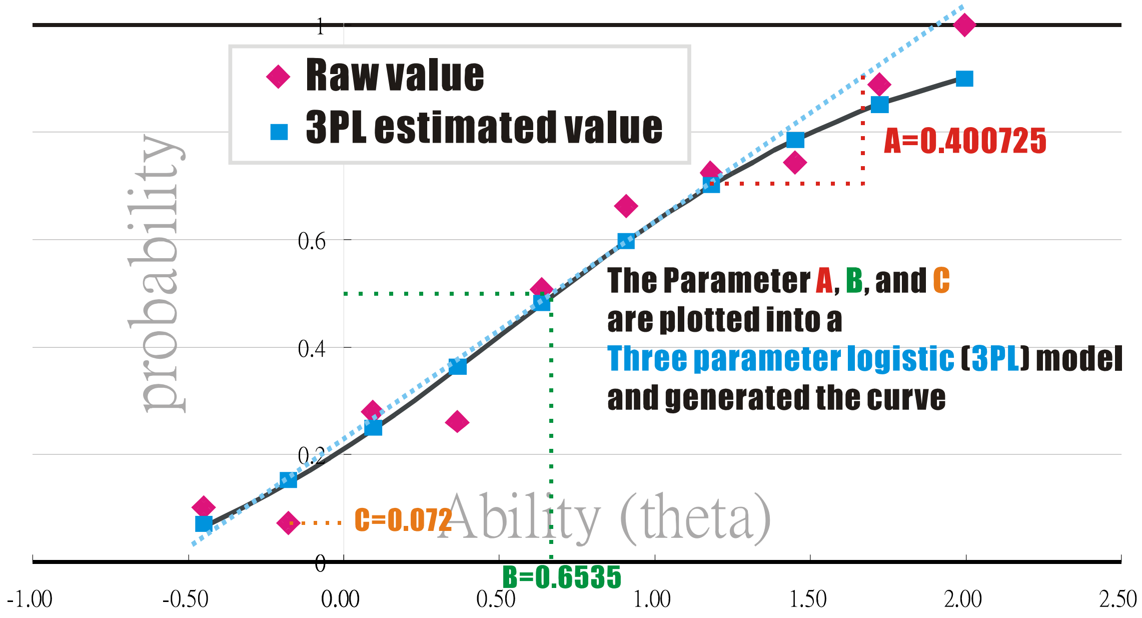


Figure 4.18 Plotting three parameters to generate the curve

Microsoft Excel spreadsheet software was subsequently designed to incorporate the following formulae (in Appendix A), which shortened the analysis process considerably. For the actual outputs of this system please see Figure 4.19.

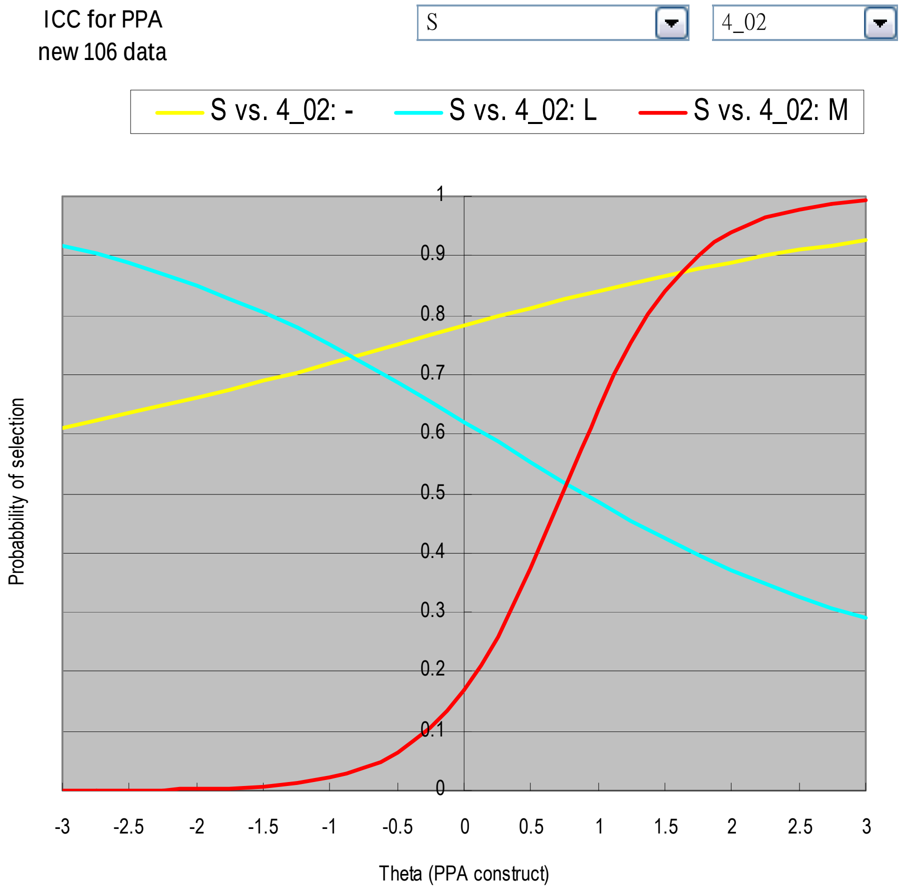


Figure 4.19 Sample IRT comparison of item 4_02 against S construct

The new analysis also includes a section for differential item functioning (DIF) (see Figure 4.20). The Pearson Chi-square goodness of fit method is also used to examine between the fit of the parameter estimation method and the actual data (see Figure 4.21).

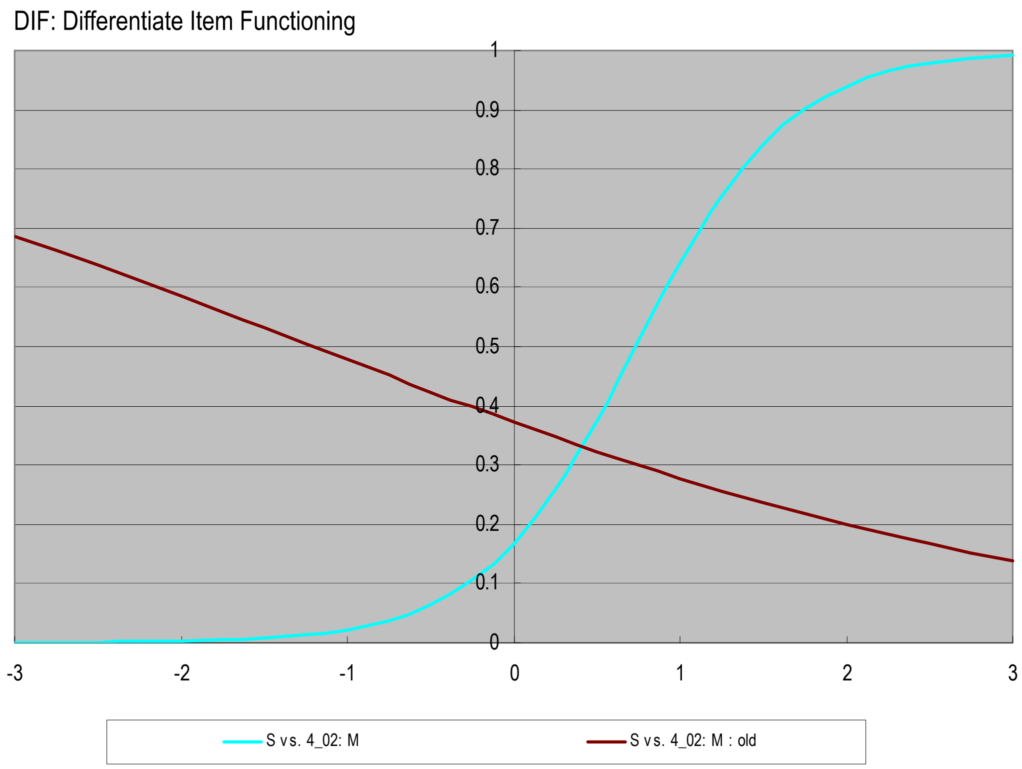


Figure 4.20 DIF analysis between old and new item 4_02, self mask (III)

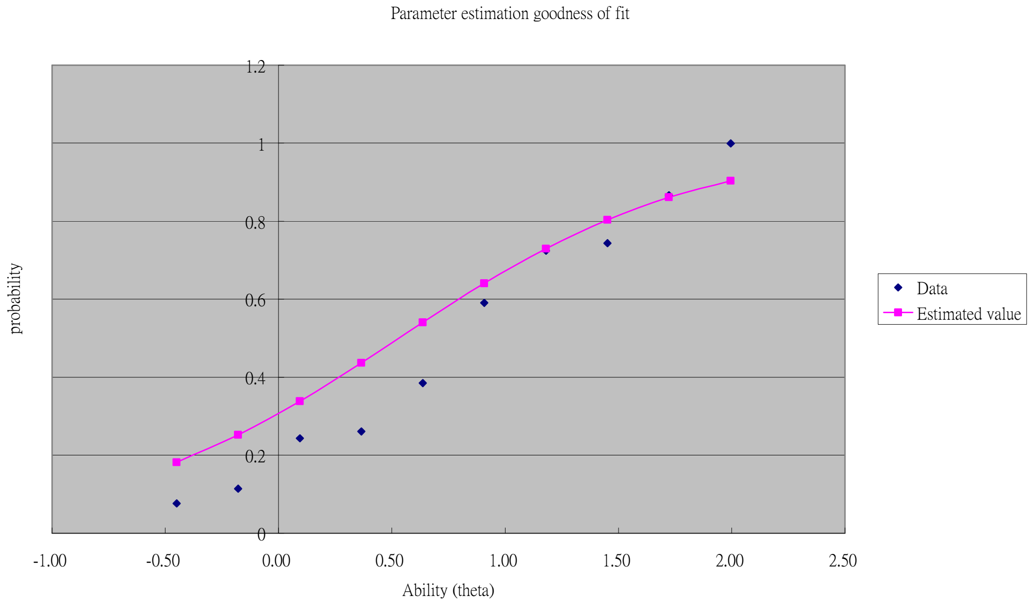


Figure 4.21 Sample Chi-square goodness of fit of SI construct vs. item 1_01

4.10. Graded Response Model (GRM)

The correlation parameter estimation (CPE) method does not fully express the ordinal nature of the PPA. CPE would assume that no sub-options are related to one another. However, when observing the options in the RICC, it is clear that there is an ordinal relationship. For this reason, this study further investigates items using Samejima's (1999) Grade Response Model (GRM), using the R package written by Rizolopulos (2006). (The mathematical details are in Appendix A)

4.10.1 Parameter Estimation: MMLE

Rizolopulos' (2006) Latent Trait model (ltm) model uses marginal maximum likelihood estimation (MMLE) for parameter estimation. MMLE assumes that the respondents represent a random sample from a population and that their ability is distributed according to a normal distribution function. The model parameters are estimated by maximising the observed data's log-likelihood. (This is obtained by integrating out the latent variables, as shown in Appendix A)

4.10.2 Model fit calculation:

The model fit of this research uses two-way and three-way margins. This is an extension of the original goodness of fit method (Rao & Sinharay, 2007).

4.10.3 Item Response Category Characteristic Curve (IRCCC)

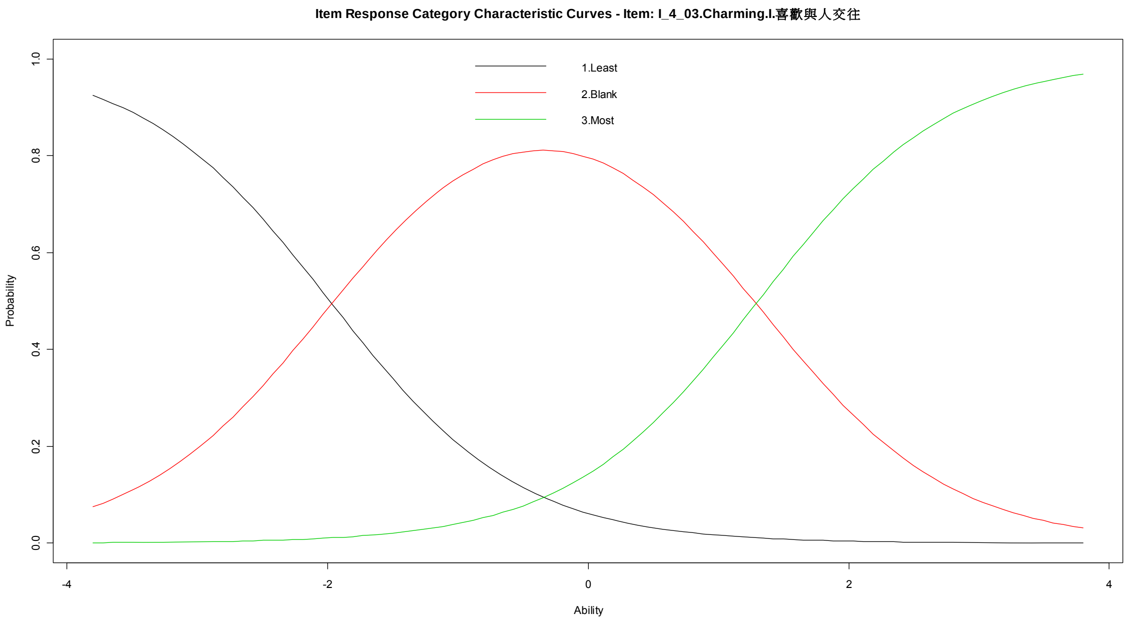


Figure 4.22 Sample Item Response Category Characteristic Curve from PPA

Samejima's Item Response Category Characteristic Curve (IRCCC) is used as the main outcome of the PPA GRM item responses. The above graph (Figure 4.22) is a good example of optimum performance of an IRCCC. The green curve (3.Most) positively discriminates test respondents. The black curve (1.Least) negatively discriminates respondents. The red curve (2.Blank) demonstrates a typical normal distribution curve.

The green and black curves are crossing very close to 0 ability (theta). It is an indication that when an individual's ability (theta), or tendency towards target construct is identified by the current item group, the likelihood of marking the current item as 'M' (Most) increases. Alternatively, when the respondent has a lower tendency towards the target construct, it would be more likely for them to mark the current item as 'L' (Least). Also, it is more likely for an individual to mark the current item as 'Blank' when their tendency towards the target construct is between -2 to 2 (ability) theta.

4.11. Forced choice to multiple choice questions (FCMCQ) flattening method

It has already been shown (see Chapter 3 Area D.1) that within a tetrad, there are

$$ {}_4P_2 = \frac{4!}{(4-2)!} = 12 $$

combinations (choosing 2 words within 4 options, when the order of selection counts as difference) possible options (H(high)D - L(low)C, HDLI, HDLS, HILC, HILD, HILS, HSLC, HSLD, HSLI, HCLD, HCLI, and HCLS). This research has designed a method to count the occurrence of each option (see Figure 4.23). An item set (tetrad) should present equal probability for all 12 options. If over selection of certain option(s) occurs, this could be due to internal and external variables, that is.

- *Internal variables.* Poor terminology for the item (vague or difficult), use of 'over-preferred' or 'under-preferred' terms, and poor combination within a tetrad.
- *External variables.* Different cultural interpretation or preference for an item, tetrad, and construct, current cultural emphasis (or Social Desirable Response), or actual cultural difference.

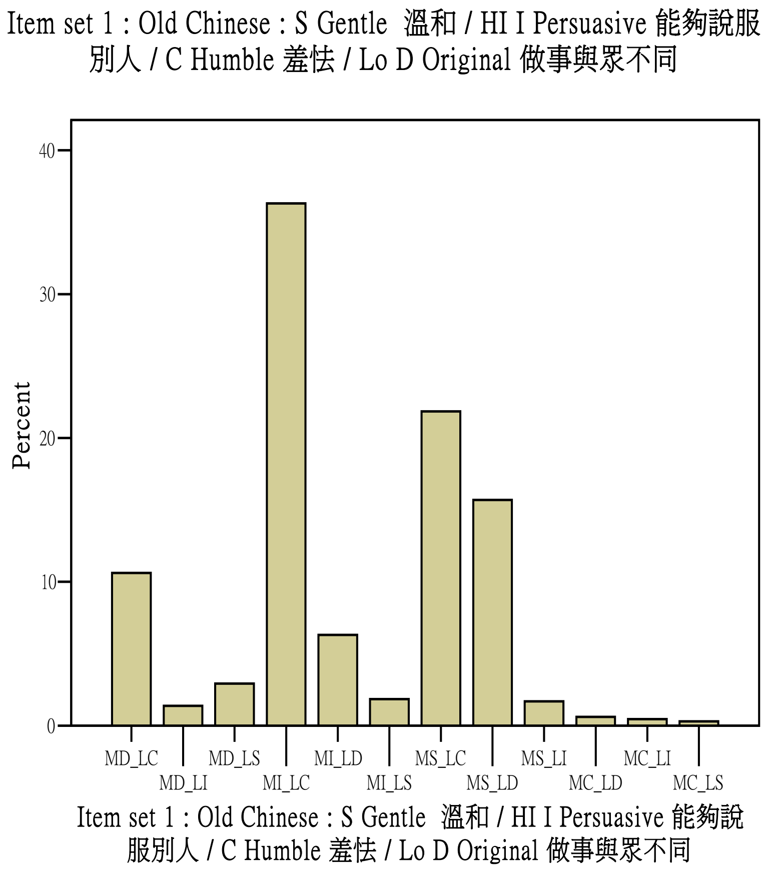
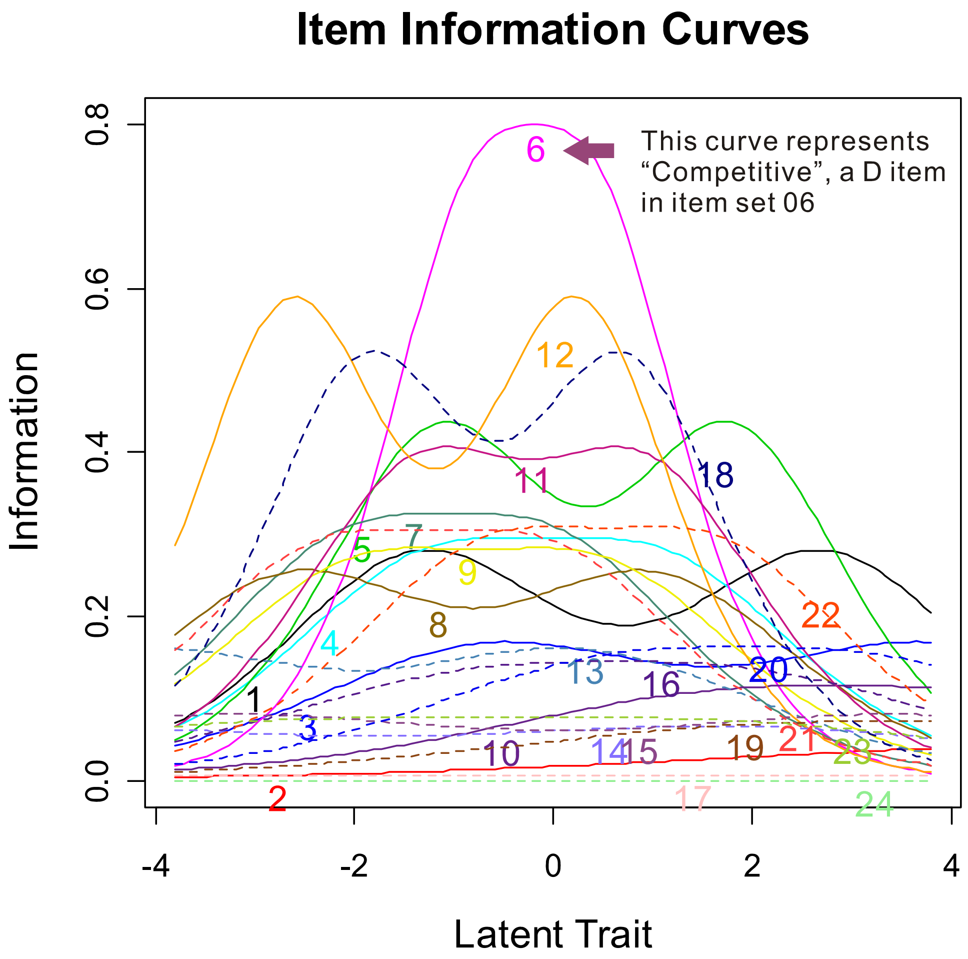


Figure 4.23 Sample forced choice to multiple choice questions (FCMCQ) bar graph

Past research with the Myers-Briggs Type Indicator (MBTI) has indicated that the 16 personality types do not all have equal percentages (Hammer & Mitchell, 1996; Mills & Parker, 1998). It is therefore reasonable to postulate that 12 possible options are also naturally un-equally presented within the population. This could be one of the external variables with unequal representations in the tetrad combination. However, the current research assumes the unequal representations originated from poor item structure.

4.12. Item Information Curve (IIC) and Test Information Function (TIF)



This curve represents "Competitive", a D item in item set 06

Figure 4.24 Sample IIC and TIF, sample item information curve for D construct

Item set 06	D	Competitive	S	Considerate	Lo I	Happy	Lo C	Harmonious

Item Information Curves (IIC) and Test Information Function (TIF) are indications of range and amount of information measured by the current constructs. The IIC illustrates the amount of information and range that is measured by all individual items within the current construct.

For example, as indicated above (item information curve, in Figure 4.24), item 6 illustrated by the magenta line (1_06, Competitive), can generate a large amount of item information between the range -2 to 2 (theta, here as D construct). Item ‘competitive’ is therefore a good item with which to measure the respondent who has -2-2 (theta) or a tendency towards a D construct.

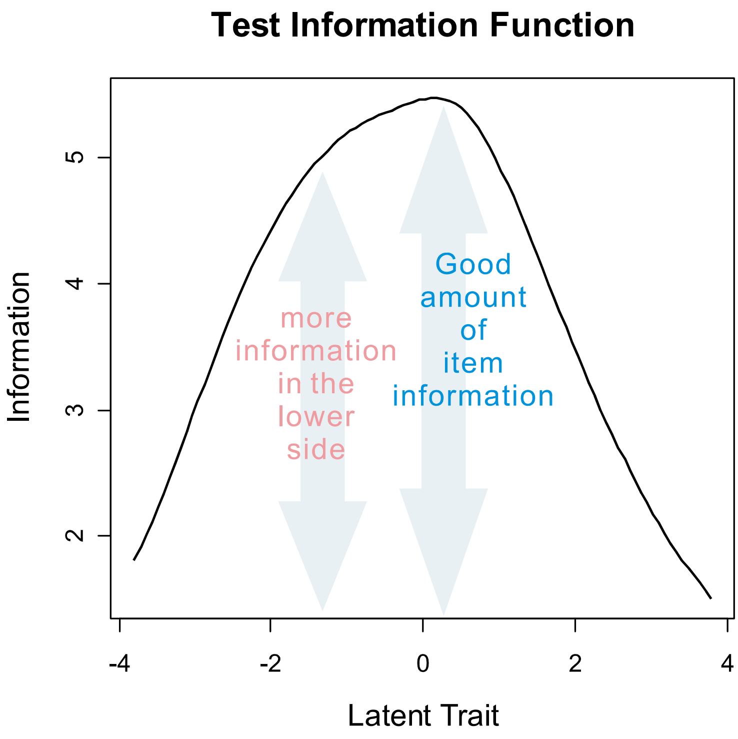


Figure 4.25 Sample TIF

The Test Information Function (TIF) is a curve that summarises all the individual item information curves into one curve (see Figure 4.25). For example, the above test information curve indicates that current items within the D construct could generate a useful amount of information for low and high theta – with bit more information in the lower side of theta. It is a good indication that the current items are functioning appropriately for the construct.

4.13. Item Response Category Characteristic Curves (summary) IRCCC-S

The Item Response Category Characteristic Curves summary graph (IRCCC-S) puts different responses into different graphs contrasting all items within the target construct. The category one (top left in Figure 4.26) indicates the first ordinal response 'L' (1.Least), is generally demonstrating negative discrimination curves (high left with low right). However, a few items are overly negative, which represents a general negative response. The category two (top right in Figure 4.26) indicates a second ordinal response '2, blank.' Most of the curve within the 2nd category would be closer to a bell curve. The category three (Lower left in Figure 4.26) is the 'M' (Most) curve, which is demonstrating the positive curve (low left and high right).

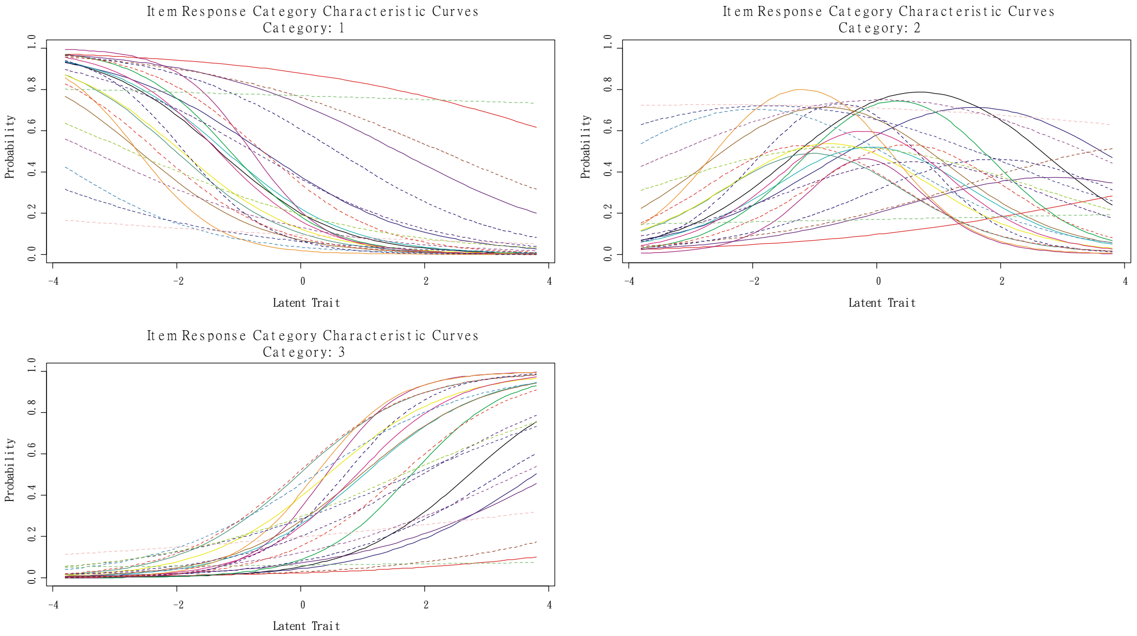


Figure 4.26 Sample RICCC-S

The IRCCC-S is used in this dissertation to examine items within the target constructs. In above example, all the items that have negative discrimination for L curves, normal distribution curve for blank curves, and positive discrimination for M curves are assumed to demonstrate a well functioning construct. In contrast, other trends could indicate problematic formulation or interpretation error.

4.14. Chapter summary and discussion

All the PPA forms were administered according to the standardised procedures. The current study collected three sample groups, Part I of the research collected 650 samples via the old form (CPPA25/C7) in the Beijing area (male=323, Female=267, missing=60). Data for Part I were collected from April 2004 to November 2007. Part II of this research used amended forms in the same area and collected 307 samples (male=185, female=119, missing=3). Data for Part II were collected from November 2007 to May 2008.

The data were captured via Microsoft Access and then analysed in SPSS 12 and Microsoft Excel. This research used a double-blind method of in data collection, analysis, and interpretation. Standardised processes of administering and data capturing were also used. The old forms have various versions (all textually equivalent). The results of ANOVA suggest that these are all equal, apart from some minor differences in the CI and CII constructs (CI. *F* (8, 641) = 2.443, *p* = .013; CII *F* (8, 641) = 2.532, *p* = .010). The Hochberg homogenous sub-test also suggests slight differences in C7-3 (see Table 5.1, p. 136 for a list of abbreviations).

The Classical test analysis of this study utilised Cronbach's Alpha, item difficulty (P), and the item discrimination index. However, in the chapter it is argued that Cronbach's may be unsuitable in the case of FC items, due to its heavy reliance on the correlation technique. Item difficulty is mostly used in ability tests, and in this study (P) was operationalised as a popularity index. The PPA was scored according to its original framework. The details for interpreting constructs, as well as the scoring method were defined.

On the IRT analysis conducted in this study, the RICC definitions of positive, negative, L-item, M-item, non-used item, and complex items were operationalised. The common IRT problem of item swing was also defined.

The current study used correlation parameter estimation (CPE) methods in three-parameter logistic (3PL) methods for the RICC. The goodness of fit was used to validate the result. The General Response Model (GRM) by Samejima (1999) was used in conjunction with the RICC analysis. This study uses Rizopoulos' (2006) 'ltm' pack in R to calculate the GRM model, as well as the Marginal Maximum Likelihood Estimation (MMLE), two-three way margin fit, Item Response Category Characteristic Curve (IRCCC), Item Information Curve (IIC), Test Information Function (TIF), and IRCCC-summary form. In this study a forced choice approach to Multiple Choice Questions (FCMCQ) to explore item dynamics was used.

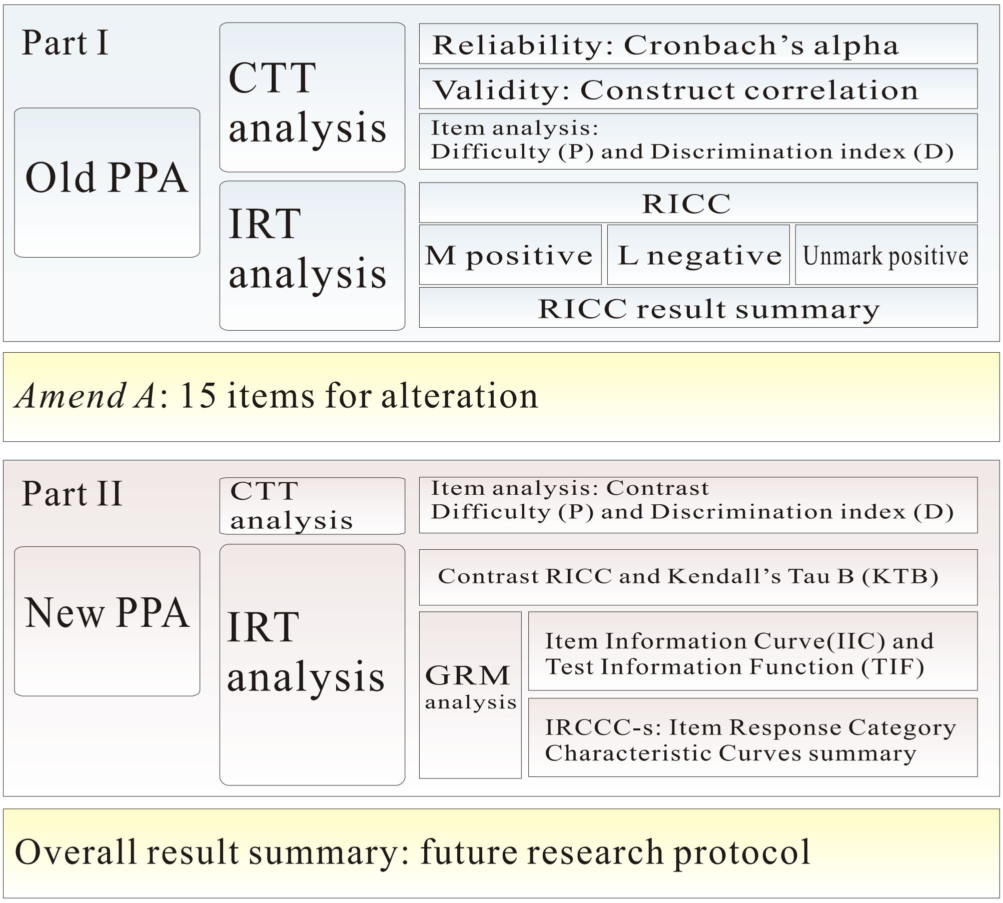
This paper explored many methods with a single database. The current study tried to encompass too many methods which could lead to a superficial research result. This paper might not able to go into sufficient depth of each method and could lead to confusion. Using the current paper and an overview, it is therefore suggested that future research should investigate the methods in-depth for further confirmation.

CHAPTER 5. RESULTS

5.1. Research results

The current study was exploratory in nature, therefore all research results are included to enhance understanding of the field involving the use of item response theory (IRT) in Personal Profile Analysis (PPA) forced choice (FC) instruments. The results are presented in the order of the research process. The findings are reported in the following structure (see Figure 5.1). The results of Part I is presented in section 5.2. The Amend A is presented in section 5.5.

The results of Part II are presented in section 5.6. The final end results of the current study – a research protocol for future PPA-IRT research – are presented in section 5.7. The full details of the original results are documented in Appendix 1 to 4 (on the CD). (For the purpose of convenience, the list of acronyms that been used in this chapter are listed again in Table 5.1)


Part I
Old PPA
CTT analysis
Reliability: Cronbach's alpha
Validity: Construct correlation
Item analysis: Difficulty (P) and Discrimination index (D)
IRT analysis
RICC
M positive
L negative
Unmark positive
RICC result summary
Amend A: 15 items for alteration
Part II
New PPA
CTT analysis
Item analysis: Contrast Difficulty (P) and Discrimination index (D)
IRT analysis
Contrast RICC and Kendall's Tau B (KTB)
GRM analysis
Item Information Curve(IIC) and Test Information Function (TIF)
IRCCC-s: Item Response Category Characteristic Curves summary
Overall result summary: future research protocol

Note: for the details of acronyms please see Table 5.1

Figure 5.1 Structure of research findings

Table 5.1 List of abbreviations

3PL-IRT	Three-parameter Logistic Item Response Theory model
Amend A	Amendment of 16 items after Part I research
C	Compliance construct (or overall score in Self mask)
CI	Compliance construct in Work mask
CII	Compliance construct in Pressure mask
CPE	Correlational Parameter Estimation Method
CTT	Classical Test Theory
D	Dominance construct (or overall score in Self mask)
DI	Dominance construct in Work mask
DII	Dominance construct in Pressure mask
DISC	Marston’s DISC theory (Dominance, Influence, Submission, and Conformity)
ERB	Extreme Response Bias
FC	Forced choice
FCMCQ	Forced choice to Multiple Choice Question (modified model)
GRM	Samejima’s General (graded) Response Model
Hi	High scoring. PPA scoring method - items only scored when marked ‘high’
I	Influence construct (or overall score in Self mask)
ICC	Item Characteristic Curve
II	Interactive construct in Work mask
IIC	Item Information Curve graph
III	Interactive construct in Pressure mask
Index (D)	Discrimination index
Index (P)	Difficulty index
Index (PI)	Preference index
IRCCC	Item Response Categories Characteristic Curve
IRCCC-s	Item Response Categories Characteristic Curve summary graph
IRT	Item Response Theory
IRTCI	IRT Construct interpretation method
ITCC	Item Total Correlation Co-efficient (construct interpretation method)
KTB	Kendal’s Tau B Ordinal Correlation analysis (modified)
Lo	Low scoring. PPA scoring method - items are only scored when marked ‘least’
Part I	Research Part I (Beijing sample, n=650) are collected via old Chinese PPA form
Part II	Research Part II (Beijing sample, n=307) are collected via New

	Chinese PPA form with Amend A
PPA	Personal Profile Analysis
RICC	Raw Item Characteristic Curve
S	Submission construct (or overall score in Self mask)
SDR	Social Desirable Response
SI	Submission construct in Work mask
SII	Submission construct in Pressure mask
TIF	Test Information Function graph

5.2. Old PPA (Form A251, C7, and others)

5.2.1 Classical Test Theory (CTT): old form

5.2.1.1 Reliability: Cronbach's Alpha result

Cronbach's Alpha was conducted to explore the internal consistency of the form (see Table 5.1). The FC format was converted to ordinal form for reliability of statistics because reliability of FC psychometrics is potentially problematic due to its format (Baron, 1996; Bartram, 2007; Hicks, 1970; Martinussen, et al., 2001; McCloy, et al., 2005). This could be the reason that all PPA constructs yield low Cronbach's Alpha coefficients. D constructs (D in work mask, pressure mask, and self mask) seem to have better reliability compared to the others (α=0.665(DI), α=0.724(DII), α=0.722(D)). The construct said to be most problematic would be the C construct (α=0.211(CI), α=0.435(CII), α=0.434 (C)). The reason for low the Alpha was discussed in section 4.5.

Table 5.2 Old form (A251, C7) reliability statistic (n=650)

Construct	Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
D work mask	0.665	0.649	20
I work mask	0.563	0.555	17
S work mask	0.525	0.513	19
C work mask	0.227	0.211	15
D Pressure mask	0.725	0.718	21
I Pressure mask	0.559	0.548	19
S Pressure mask	0.477	0.458	19
C Pressure mask	0.437	0.454	16
D Self Mask	0.722	0.710	24
I Self Mask	0.558	0.544	23
S Self Mask	0.569	0.560	25
C Self Mask	0.434	0.448	24

5.2.1.2 Construct correlation

In the current study, the item-construct-function was similar to the UK PPA form research (see Table 5.6). This particular relationship has been widely observed by international PPA researchers (Irvine, 2003). The original correlation relationships showed that the D construct has little to no relationship with the I construct (r= -.127~.11). However, the D construct seems to be significantly different from the S construct (r=-.78~-.60), and from the C construct (r=-.60~-.45). Also, there was a moderate relationship between the S and the C constructs (r=.12~.47).

Correlation within mask

In the current study (see Tables 5.3~5.5), a strong opposition (negative correlation) between D and S existed within the self mask (*r*=-.61, *p*<.01), work mask (*r*=-.59, *p*<.01), and pressure mask (*r*=-0.5, *p*<.01), followed by moderate opposition (negative correlation) between D and C within the self mask (*r*=-.45, *p*<.01), work mask (*r*=-.20, *p*<.01), and pressure mask (*r*=-.40, *p*<.01).

Finally, the low correlation was also evident in the current study between S and C in the ‘self mask’ (r=.12, p<.01) and pressure mask (r=.14, p<.01).

Table 5.3 Self mask correlations for simplified Chinese translation (n=650)

Variables	1	2	3	4
1. D in percentage	-			
2. I in percentage	-.127**	-		
3. S in percentage	-.606**	-.398**	-	
4. C in percentage	-.450**	-.368**	.121**	-

** Correlation is significant at the 0.01 level (2-tailed).

Table 5.4 Work mask correlation for simplified Chinese translation (n=650)

	5	**6**	**7**	**8**
5. DI in percentage	-			
6. II in percentage	-.087*	-	-	
7. SI in percentage	-.585**	-.383**	1	
8.CI in percentage	-.200**	-.296**	-.041	-

* Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed).

Table 5.5 Pressure mask correlation for Simplified Chinese translation (n=650)

	9	**10**	**11**	**12**
9. DII in percentage	-			
10. III in percentage	-.228**	-		
11. SII in percentage	-.500**	-.288**	-	
12. CII in percentage	-.403**	-.279**	.141**	-

** Correlation is significant at the 0.01 level (2-tailed). I=work mask, II=pressure mask, and III=self mask

Table 5.6 Correlation of Graph III with international PPA research, (Irvine, 2003)

Country	**Constructs (Correlation)**			
Russia n=600	D	I	S	C
Dominance		0.03	-0.65	-0.51
Influence			-0.50	-0.58
Steadiness				0.47
Compliance				
Holland n=127	D	I	S	C
Dominance		0.11	-0.70	-0.60
Influence			-0.31	-0.57
Steadiness				0.38
Compliance				
Turkey n=214	D	I	S	C
Dominance		0.04	-0.72	-0.47
Influence			-0.32	-0.52
Steadiness				0.28
Compliance				
Denmark n=539	D	I	S	C
Dominance		-0.09	-0.73	-0.50
Influence			-0.31	-0.44
Steadiness				0.32
Compliance				
USA n=1512	D	I	S	C
Dominance		0.05	-0.78	-0.54
Influence			-0.45	-0.61
Steadiness				0.46
Compliance				
UK n= 4083	D	I	S	C
Dominance		-0.15	-0.75	-0.49
Influence			-0.21	-0.40
Steadiness				0.24
Compliance				
China n=650	D	I	S	C
Dominance		-0.127	-0.606	-0.450
Influence			-0.398	-0.368
Steadiness				0.121
Compliance				

Correlation among masks

Note that although the three personal profile masks share the same DISC construct names and items, they are not measuring the same constructs. The work mask is a composite of positive responses (M= most like me), the pressure mask is a composite of negative responses (L= least like me), and the self mask is a composite of positive response minus negative responses. In the current study, the D, I, and S constructs generally retained a good moderate correlation between work mask (I) and pressure mask (II). The C construct had a low correlation between work mask (I) and pressure mask (II) (see Tables 5.7 – 5.10).

Table 5.7 D Correlations across three masks (n=650)

		DI in percentage	DII in percentage	D in percentage
DI in percentage	Pearson Correlation	-		
DII in percentage	Pearson Correlation	.577**	-	
D in percentage	Pearson Correlation	.885**	.853**	-

** Correlation is significant at the 0.01 level (2-tailed). I=work mask, II=pressure mask, and III=self mask

Table 5.8 I Correlations across three masks (n=650)

		II in percentage	III in percentage	I in percentage
II in percentage	Pearson Correlation	-		
III in percentage	Pearson Correlation	.482**	-	
I in percentage	Pearson Correlation	.865**	.838**	-

** Correlation is significant at the 0.01 level (2-tailed). I=work mask, II=pressure mask, and III=self mask

Table 5.9 S Correlations across three masks (n=650)

		SI in percentage	**SII in percentage**	**S in percentage**
SI in percentage	Pearson Correlation	-		
SII in percentage	Pearson Correlation	.459**	-	
S in percentage	Pearson Correlation	.886**	.804**	-

** Correlation is significant at the 0.01 level (2-tailed). I=work mask, II=pressure mask, and III=self mask

Table 5.10 C Correlations across three masks (n=650)

		CI in percentage	**CII in percentage**	**C in percentage**
CI in percentage	Pearson Correlation	-		
CII in percentage	Pearson Correlation	.118**	-	
C in percentage	Pearson Correlation	.702**	.769**	-

** Correlation is significant at the 0.01 level (2-tailed). I=work mask, II=pressure mask, and III=self mask

5.2.1.3 Item reliability statistics: Old form

The current study generated Cronbach's Alpha coefficients for exploratory purposes. The D reliability ranged from .692 to .730. The I reliability ranged from .580 to .506. The S reliability ranged from .588 to .521, and the C reliability ranged from .464 to .381. In general, D and I constructs had better correlation among items. This was operationalised by low item-total correlation (r_{it}).

There are items reported to have low item total correlation in each construct. There are four items in Dominance (D) construct ($r_{it=}$ .06 ~ -.007); six items in Influence (I) construct (r_{it}= .095 ~ -.073); nine items in Submission (S) construct (r_{it}= .09 ~ -.07); and eleven items in Compliance (C) construct (r_{it}= .093 ~ -.117). In terms of Cronbach's Alpha, the D, I and S construct are more reliable than the C construct. The C constructs contains too many unreliable items and generate the lowest Cronbach's Alpha.

Table 5.11 Cronbach's Alpha: list of problematic items (n=650)

Item	Original English Item	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Problematic items in construct Dominance (D)					
2_14	Daring	0.1923	29.699	-0.007	0.73
4_16	Assertive	0.9677	29.221	0.06	0.728
4_17	Persistent	0.1985	29.478	0.018	0.73
1_24	Faithful	1.1031	29.593	0.017	0.728
Problematic items in construct Influence (I)					
1_02	Persuasive	1.1169	17.946	-0.024	0.57
2_08	Inspiring	1.1831	17.392	0.08	0.559
3_10	Polite	1.4892	18.096	-0.068	0.58
1_12	Polished	1.2831	18.025	-0.051	0.577
1_16	Confident	1.2554	17.226	0.095	0.558
1_21	Trusting	1.7031	18.147	-0.073	0.578
Problematic items in construct Submission (S)					
4_02	Pleasant	1.0369	18.652	0.066	0.57
2_04	Obliging	1.4692	18.838	0.013	0.577
4_10	Moderate	0.7246	18.878	-0.018	0.585
3_13	Soft-touch	1.3754	18.987	-0.025	0.583
2_17	Generous	0.3692	18.357	0.09	0.569
2_21	Contented	1.1892	18.699	0.007	0.583
4_23	Restrained	1.3677	18.677	0.019	0.58
2_24	Neighbourly	0.5354	18.939	0.006	0.576
1_21	Trusting	0.8508	19.224	-0.07	0.588
Problematic items in construct Compliance (C)					
2_02	Dutiful	-2.5938	14.297	-0.117	0.464
1_04	Open minded	-2.4938	13.696	0.035	0.438
1_07	Fussy	-1.8015	13.697	-0.005	0.45
2_10	Receptive	-2.5385	13.617	0.038	0.439
3_12	Diplomatic	-1.9923	13.465	0.025	0.446
3_15	Agreeable	-2.0108	14.026	-0.052	0.454
1_17	Well-disciplined	-2.1554	13.324	0.065	0.436
2_20	Adaptable	-2.4354	14.009	-0.063	0.46
2_22	Cultured	-2.1215	13.417	0.089	0.429

Item	Original English Item	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
2_23	Accurate	-2.5000	13.437	0.093	0.428
4_24	Faithful	-2.6046	13.605	0.054	0.435

5.2.1.4 Item analysis: Difficulty analysis (P) and Discrimination index (D)

D construct

The difficulty index (P) revealed that 12 D construct items were considered to be difficult for individuals who were considered as high D as well as Low D. The result indicated that the Chinese translation for the terms ‘Assertive’ (D_16, p=.025), ‘Restless’ (D_24, p=.037), ‘Stubborn’ (D_02, p=.065), ‘Pioneering’ (D_19, p=.083), ‘Bold’ (D_03, p=.086), ‘Argumentative’ (D_20, p=.136), ‘Adventurous’ (D_10, p=.142), ‘Original’ (D_01, p=.179), ‘Courageous’ (D_05, p=.213), ‘Determined’ (D_14, p=.232), ‘Vigorous’ (D_22, p=.238), and ‘Persistent’ (D_17, p=.244) were considered too ‘difficult (unpopular)’ for the current sample.

The result suggested that cognitive and affective tests should set different standards for the P index. The results also suggested that the difficulty in affective self-report tests could be defined as popularity. The current results further suggested that all the above items were unpopular for use as a self-descriptive term in the current Chinese population. The 9 out of 12 items scoring low in P value also scored low in D value (discrimination index). The results of this study therefore suggest that when items are too unpopular, the discrimination index value also decreases.

I construct

The difficulty index (P) revealed that 11 D construct items were found to be difficult for individuals who were considered ‘High I’ as well as ‘Low I’. The results indicated that the Chinese translation for the terms ‘Admirable’ (I_18 , p=.093), ‘Dramatic’ (I_17 , p=.099), ‘Light-hearted’ (I_20 , p=.099), ‘Trusting’ (I_21 , p=.108), ‘Happy’ (I_06 , p=.111), ‘Playful’ (I_07 , p=.114), ‘Cheerful’ (I_04 , p=.154), ‘Good-mixer’ (I_22 , p=.185), ‘Jovial’ (I_05 , p=.198), ‘Talkative’ (I_11 , p=.219), and ‘Cordial’ (I_10 , p=.244) were considered too ‘difficult (unpopular)’ for the sample.

It was found that 8 out of 11 items scoring low in the P value also scored low in the D value (discrimination index).

S construct

The difficulty index (P) revealed that 14 S construct items were considered difficult for individuals who were seen as High S as well as Low S. The result showed that the Chinese translation for terms ‘Pleasant’ (S_02, P=.015), ‘Obliging’ (S_04, P=.049), ‘Satisfied’ (S_12, P=.053), ‘Soft-touch’ (S_13, P=.053), ‘Neighbourly’ (S_24, P=.090), ‘Lenient’ (S_22, P=.096), ‘Submissive’ (S_08, P=.102), ‘Contented’ (S_21, P=.117), ‘Sympathetic’ (S_16, P=.176), ‘Obedient’ (S_07, P=.179), ‘Accommodating’ (S_19, P=.198), ‘Moderate’ (S_10, P=.216), ‘Trusting’ (S_21, P=.235), ‘Patient’ (S_09, P=.238) were too ‘difficult (unpopular)’ for the sample. The results indicated that 11 out of 14 items that were scored low in P value were also scored low in D value (discrimination index).

C construct

The difficulty index (P) revealed that 15 C construct items can be regarded as difficult for individuals who were considered as High C as well as Low C. The result showed that the Chinese translations of the terms ‘Humble’ (C_01, P=.015), ‘Timid’ (C_08, P=.034), ‘Resigned’ (C_18, P=.037), ‘Easily Led’ (C_03, P=.053), ‘Cautious’ (C_14, P=.056), ‘Fearful’ (C_13, P=.071), ‘Agreeable’ (C_15, P=.077), ‘Fussy’ (C_07, P=.099), ‘Soft-Spoken’ (C_09, P=.105), ‘Conventional’ (C_11, P=.154), ‘Cultured’ (C_22, P=.157), ‘Diplomatic’ (C_12, P=.207), ‘Harmonious’ (C_06, P=.216), ‘Peaceful’ (C_21, P=.216), ‘Well-disciplined’ (C_17, P=.228) were considered too ‘difficult (unpopular)’ for the respondents in the sample. The results indicated that 10 out of 15 items that were scored low in P value, were also scored low in D value (discrimination index).

5.3. Item Response Theory (IRT): old form

5.3.1 Raw Item Characteristic Curve (RICC) model:

In this study RICC positive relationship was operationalised in two formats. One was the positive relationship with 'Most (M)' curve and the other was a negative relationship with 'Least (L)' curve (refer to Chapter 4, section 4.8 for operationalisation detail).

5.3.1.1 M positive

The RICC-M-positive curve RICC analysis showed that 25 out of 96 items were 'contaminated' (see the definition of 'contamination' on Chapter 3 Area E.3~5, Chapter 4, section 4.8). These items are given such rating due to the reason that they are expressing two constructs (please see Table 5.9, they are marked as ***). The 25 contaminated items are. 1_01, 1_02, 2_02, 3_03, 1_04, 2_07, 1_08, 2_09, 4_09, 3_10, 2_11, 1_12, 4_14, 1_16, 2_16, 3_16, 2_17, 4_17, 2_18, 4_18, 2_20, 3_20, 3_21, 4_21, and 4_22.

The analysis suggested that 30 out of 96 items can be classified as weak because they did not show any specific item expressed construct (IEC) in the Most (M) positive curve (see Table 5.9, they are the items the marked *).

The 30 weak items are. 3_01, 3_02, 4_02, 1_03, 2_03, 2_04, 3_06, 1_07, 4_07, 4_08, 4_10, 3_12, 4_12, 3_13, 1_14, 3_14, 2_15, 3_15, 4_16, 1_17, 1_18, 3_18, 3_19, 1_21, 2_21, 1_22, 2_22, 4_23, 1_24, and 3_24. Item 4_24 can be regarded as a highly contaminated item because the IEC analysis indicates it is expressing three constructs, (all CDS). Please see Table 5.12, this type of item is marked with ****. This rating is given only to the items that express three different constructs across DISC. For example, an item express traits for D, I, and C would be considered as highly contaminated. However, if it only expresses sub constructs, it would not be considered as contamination.

For example item expressing D, DI, and DII are all the sub constructs within the D construct; therefore not contamination (refer to Table 5.9 for full RICC of the old PPA).

5.3.1.2 L Negative

The L-negative showed similar results. The 21 out of 96 items are considered as contaminated due to the reason that they are showing more than one constructs. Please see Table 5.13, they are the items that marked with *** (for further details please see Chapter 4, section 4.8.2). The 21 contaminated items are. 3_01, 1_03, 2_04, 2_05, 4_05, 1_07, 2_07, 2_09, 4_10, 2_11, 3_11, 4_12, 2_15, 3_15, 2_16, 4_20, 2_21, 4_21, 4_23, 1_24, and 4_24.

Also, five item identified by M-Positive are also showing up again in L negative contamination list (2_07, 2_09, 2_11, 2_16, and 4_21). There are 27 out of 96 items are rated as weak items due to lack of expressing any construct (please see Table 5.13, they are marked with *). These 27 items are. 2_01, 2_02, 3_03, 1_04, 3_04, 3_07, 1_08, 3_08, 3_12, 1_13, 3_13, 2_14, 4_14, 1_15, 4_15, 1_16, 3_16, 4_17, 1_18, 2_18, 1_19, 3_19, 3_20, 3_21, 2_23, 3_23, and 2_24. There are four items in the M-positive weak list also appearing in the L-Negative List (3_12, 3_13, 1_18, and 3_19).

5.3.1.3 Not marked item-positive result

Not Marked items cannot be interpreted. However, these items were still processed to achieve a better understanding of the RICC. Items I_01, 2_01, 3_01, 1_02, 2_02, 4_02, 1_03, 2_04, 1_05, 1_07, 3_07, 1_08, 2_09, 2_10, 1_12, 2_12, 4_12, 2_13, 4_13, 1_16, 2_16, 4_17, 3_19, and 3_21 were found d to be contaminated.

Items 3_03, 4_03, 3_04, 4_04, 2_06, 4_06, 3_08, 3_11, 3_12, 3_14, 3_15, 1_17, 3_17, 1_19, 4_21, and 4_22 were found to be weak. Items 4_02 and 4_20 were heavily contaminated items (refer to Table 5.14 for full RICC of the old PPA).

The following tables are the results yielded from raw item characteristic curve (RICCC) (please see reading key below for interpretation instructions). It is defined as ‘expressing trait’ when a M response is demonstrating a positive RICC curve, and when a L response is demonstrating a negative RICC curve (see section 4.8 for more details). The results are summarised into table form (Table 5.12~5.14).

Reading key for Table 5.12~5.14:

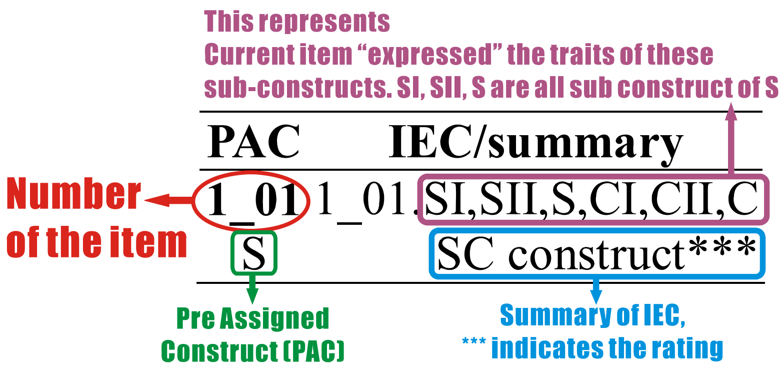


Table 5.12 Result from RICC, IEC from item marked 'Most'-positive relationship (old form n=650)

PAC[I]	IEC/summary	PAC[I]	IEC/summary	PAC[I]	IEC/Summary	PAC[I]	IEC/Summary
1_01	1_01.SI,SII,S,CI,CII,C	2_01	2_01,DI,I,II,III,	3_01	3_01,	4_01	4_01,D,
S	SC construct***	Hi I	I construct (with some D)	C	*	loD	D construct
1_02	1_02,SI?,CII?,	2_02	2_02,DI,CI,	3_02	3_02,	4_02	4_02,
I	SC construct***	C	DC***	D	*	loS	*
1_03	1_03,	2_03	2_03,	3_03	3_03,SI,SII,CI?,	4_03	4_03,I,II,III,
Low C	*	D	*	HiS	SC***	I	I construct
1_04	1_04,SI,S,C	2_04	2_04,	3_04	3_04,D,DI,	4_04	4_04,CI,
HiC	SC construct***	S	*	LoD	D	I	C
1_05	1_05,II,III,	2_05	2_05,CI,C	3_05	3_05,D,DI,	4_05	4_05,SI,SII,S,
LoI	I construct	C	C construct	LoD	D construct	S	S construct
1_06	1_06,D,DI,DII,	2_06	2_06,SI,SII?,S,	3_06	3_06,	4_06	4_06,CII,C
D	D construct	S	S construct	LoI	*	loC	C construct
1_07	1_07,	2_07	2_07,SI,SII,S,CI?,CII,C	3_07	3_07,D,DI,DII,	4_07	4_07,
loC	*	HiS	SC construct***	D	D construct	I	*
1_08	1_08,D,DI,DII,III,	2_08	2_08,I,II,	3_08	3_08,CI,CII,C	4_08	4_08,
HiD	DI construct***	HiI	I construct	LoS	C construct	LoC	*
1_09	1_09,I,II,	2_09	2_09,SI,S,CII,	3_09	3_09,D,DI,DII,	4_09	4_09,C
I	I construct	S	SC construct***	D	D construct	HiC	C construct (but weak)***
1_10	1_10,D,DII,	2_10	2_10,CI,C	3_10	3_10,II,SI,S,CII,	4_10	4_10,
D	D construct	HiC	C construct	LoI	SI construct***	S	*
1_11	1_11,I,II,III,	2_11	2_11,SI,SII,S,CI?,	3_11	3_11,CII,C	4_11	4_11,D,DI,
I	I construct	S	SC construct***	loC	C construct	D	D construct
1_12	1_12,SI?,S,CII,C	2_12	2_12,DI,	3_12	3_12,	4_12	4_12,
LoI	SC construct***	D	D construct	HiC	*	S	*
1_13	1_13,DI,SI?,S,CI,	2_13	2_13,I,II,III,	3_13	3_13,	4_13	4_13,C?
HiD	SDC construct****	I	I construct	S	*	loC	C construct
1_14	1_14,	2_14	2_14,DI,	3_14	3_14,	4_14	4_14,III,SI,S,
C	*	HiD	D construct	I	*	HiS	IS construct***
1_15	1_15,SI,S,	2_15	2_15,	3_15	3_15,	4_15	4_15,D,DII,
HiS	S construct	--	*	C	*	LoD	D construct
1_16	1_16,DI,DII,I,II,	2_16	2_16,S,CII,C	3_16	3_16,SII,S,CI,C	4_16	4_16,
HiI	DI construct***	LoS	SC construct***	LoC	SC construct***	D	*
1_17	1_17,	2_17	2_17,II,SI,S,	3_17	3_17,III,	4_17	4_17,D,CII,
HiC	*	S	SI construct***	LoI	I construct	D	DC construct***
1_18	1_18,	2_18	2_18,SI,SII,S,CII?,C	3_18	3_18,	4_18	4_18,D,DI,DII,
HiI	*	HiS	SC construct***	loC	*	D	DI construct***

PAC^{I}	**IEC/summary**	**PAC^{I}**	**IEC/summary**	**PAC^{I}**	**IEC/Summary**	**PAC^{I}**	**IEC/Summary**
1_19	1_19,CI,CII?,C	2_19	2_19,DI,	3_19	3_19,	4_19	4_19,SI,
HiC	C construct	D	D	I	*	S	S construct
1_20	1_20,D,DI,DI,	2_20	2_20,DI,CI?,C?	3_20	3_20,SI,SII,S,CII,	4_20	4_20,III,
D	D construct	HiC	DC construct***	LoS	SC construct***	I	I construct
1_21	1_21,	2_21	2_21,	3_21	3_21,D,DI,II,	4_21	4_21,S,CII,C
HiS/LoI	*	LoS	*	D	DI construct***	C	SC construct***
1_22	1_22,	2_22	2_22,	3_22	3_22,D,DI,	4_22	4_22,SI,SII,S,CI,C
I	*	loC	*	D	D construct	S	SC construct***
1_23	1_23,II,III,	2_23	2_23,CI,CII?,C	3_23	3_23,D,DI,	4_23	4_23,
I	I construct	HiC	C construct	D	D construct	LoS	*
1_24	1_24,	2_24	2_24,SI,	3_24	3_24,	4_24	4_24,DI,SII,CI,C
D	*	S	S construct	I	*	C	CDS construct****

NOTE: PAC^{I}=Pre-assigned construct with item number, IEC= Item Expressed Construct. D,I,S,C= DISC in Self Mask; DI,II,SI,CI= DISC in Work Mask (most); DII,III,SII,CII= DISC in Pressure Mask (lest). *=no dominant IEC under RICC. ***=Item contamination with 2 constructs. ****=item contamination with 3 constructs. ?=Item swing leads to uncertainty of IEC. A question mark represents uncertainty.

Interpretation note: For **Hi-D, Hi-I, Hi-S, Hi-C** it does not imply 'High' - that denoted the scoring method used with the PPA. This type of item is only scored when item marked 'M'. Also **lo-D, I, S, & C** items do not imply as 'low' - that denoted the scoring method used with the PPA and only scored when marked with 'L'. Most items in PPA are assigned with one PAC only. However item 1_21 has assigned for IS construct. Lo- and Hi- should read as following: Lo-items are commonly selected, therefore, only when marked as 'L', do they mean something, Lo-items are normally 'popular,' or 'preferable' terminologies. When marked as 'M' they do not count. Hi-items are normally 'extreme' or 'unpopular' terminologies. Most people would mark them 'L'. Therefore, such scores are only considered when marked with 'M'. 'L' does not count.

Table 5.13 Result from RICC, IEC from item marked 'Least'-Negative relationship (old form n=650).

PAC[I]	IEC/summary	PAC[I]	IEC/summary	PAC[I]	IEC/Summary	PAC[I]	IEC/Summary
1_01	1_01. SI, SII, S?	2_01	2_01.	3_01	3_01. S?,CI,CII,C	4_01	4_01. D,DII,III
S	S construct	Hi I	*	C	SC construct***	loD	D construct
1_02	1_02. SII	2_02	2_02.	3_02	3_02. D, DII	4_02	4_02. III,SI?,SII,S?,CI
I	S construct	C	*	D	D construct	loS	ISC construct****
1_03	1_03. SI,S,CI CII C	2_03	2_03. D, DII	3_03	3_03.	4_03	4_03. I,II,III
low C	SC construct***	D	D construct	HiS	*	I	I construct
1_04	1_04.	2_04	2_04. III, CII?	3_04	3_04.	4_04	4_04.
HiC	*	S	IC construct***	LoD	*	I	*
1_05	1_05. I,II,III	2_05	2_05. SII, CI, CII, C	3_05	3_05. D, DII	4_05	4_05. SI, SII, S, CII, C
LoI	I construct	C	SC construct***	LoD	D construct	S	SC construct
1_06	1_06. D, DI, DII	2_06	2_06. SI, SII, S	3_06	3_06. CII, C	4_06	4_06. CII, C
D	D construct	S	S construct	LoI	C construct	loC	C construct
1_07	1_07. DII, CII, C	2_07	2_07. SI, SII, S, CI	3_07	3_07.	4_07	4_07. I, II, III
loC	DC construct***	HiS	SC construct***	D	*	I	I construct
1_08	1_08.	2_08	2_08. I, III	3_08	3_08.	4_08	4_08. CII, C
HiD	*	HiI	I construct	LoS	*	LoC	C construct
1_09	1_09. I, II, III	2_09	2_09. SI, SII, S, CII, C	3_09	3_09.	4_09	4_09. CI, CII, C
I	I construct	S	SC construct***	D	*	HiC	C construct
1_10	1_10. D, DI, DII	2_10	2_10. CII?, C	3_10	3_10. SI?, SII?, S?	4_10	4_10. SII, S, CI?, CII?, C
D	D construct	HiC	C construct	LoI	S construct	S	SC construct***
1_11	1_11. I, II, III	2_11	2_11. SI, SII, S, CI?	3_11	3_11. SI,SII?,S,CII,C	4_11	4_11. D,DII
I	I construct	S	SC construct***	loC	SC construct***	D	D construct
1_12	1_12. III	2_12	2_12. D, DII	3_12	3_12.	4_12	4_12. SI, SII, S, CII
LoI	I construct	D	D construct	HiC	*	S	SC construct***
1_13	1_13.	2_13	2_13. I, III	3_13	3_13.	4_13	4_13. CII, C
HiD	*	I	I construct	S	*	loC	C construct
1_14	1_14. CII, C	2_14	2_14.	3_14	3_14. DII	4_14	4_14.
C	C construct	HiD	*	I	D construct	HiS	*
1_15	1_15.	2_15	2_15. DI, DII, CII	3_15	3_15. II, SI	4_15	4_15.
HiS	*	--	DC construct***	C	IS construct***	LoD	*
1_16	1_16.	2_16	2_16. SI, SII, S, CII	3_16	3_16.	4_16	4_16. D, DII
HiI	*	LoS	SC construct***	LoC	*	D	D construct
1_17	1_17. CI, CII, C	2_17	2_17. SI SII? S	3_17	3_17. I, III	4_17	4_17.
HiC	C construct	S	S construct	LoI	I construct	D	*
1_18	1_18.	2_18	2_18.	3_18	3_18. CII,C	4_18	4_18. DII
HiI	*	HiS	*	loC	C construct	D	D construct

PAC[I]	IEC/summary	PAC[I]	IEC/summary	PAC[I]	IEC/Summary	PAC[I]	IEC/Summary
1_19	1_19.	2_19	2_19. D, DI, DII	3_19	3_19.	4_19	4_19. SII
HiC	*	D	D construct	I	*	S	S construct
1_20	1_20. D, DI, DII	2_20	2_20. CI?	3_20	3_20.	4_20	4_20. I, II, III, S?
D	D construct	HiC	C construct	LoS	*	I	IS construct***
1_21	1_21. DII, III	2_21	2_21. SII, S?, CII	3_21	3_21.	4_21	4_21. SI, CII, C
HiS/LoI	D construct	LoS	SC construct***	D	*	C	SC construct***
1_22	1_22. I, III	2_22	2_22. CII, C	3_22	3_22. D, DI, DII	4_22	4_22. SI, SII, S
I	I construct	loC	C construct	D	D construct	S	S construct
1_23	1_23. III	2_23	2_23.	3_23	3_23.	4_23	4_23. SII, S, CII
I	I construct	HiC	*	D	*	LoS	SC construct***
1_24	1_24. D?, DII, CII?	2_24	2_24.	3_24	3_24. III	4_24	4_24.
D	DC construct***	S	*	I	I construct	C	*

NOTE: PAC[I]=Pre-assigned construct with item number, IEC= Item Expressed Construct. D,I,S,C= DISC in Self Mask; DI,II,SI,CI= DISC in Work Mask (most); DII,III,SII,CII= DISC in Pressure Mask (lest). *=no dominant IEC under RICC. ***=Item contamination with 2 constructs. ****=item contamination with 3 constructs. ?=Item swing leads to uncertainty of IEC. A question mark represents uncertainty.

Interpretation note: for **Hi-D, Hi-I, Hi-S, Hi-C**, it does not imply 'High', - that denoted the scoring method used with the PPA. This type of item only scored when item marked 'M'. Also **lo-D, I, S, & C** items, does not imply as 'low' - that denoted the scoring method used with the PPA and only scored when marked with 'L'. Most items in PPA are assigned with one PAC only. However item 1_21 has assigned for IS construct. Lo- and Hi- should read as follows: Lo-items are commonly selected, therefore, only when marked as 'L', they mean something, Lo-items are normally 'popular,' or 'preferable' terminologies. When they are marked as 'M' it does not count. Hi-items are normally 'extreme' or 'unpopular' terminologies. Most people would mark them 'L'. Therefore, such scores are only considered when marked with 'M.' 'L' does not count.

Table 5.14 Result from RICC, IEC form item marked 'not marked'-positive relationship (old form n=650)

PAC^{I}	IEC/summary	PAC^{I}	IEC/summary	PAC^{I}	IEC/Summary	PAC^{I}	IEC/Summary
1_01	1_01. , DI, DII, I, II,	2_01	2_01. , S, SI, SII, C, CII	3_01	3_01. , S, C, CI, CII	4_01	4_01. , I, II, III,
S	DI cons***	Hi I	SC construct***	C	SC construct***	loD	I construct
1_02	1_02. D, DI, CI,	2_02	2_02. , II, CII	3_02	3_02. D, DII,	4_02	4_02. , I, III, SII, CI?
I	DC ***	C	IC***	D	D	loS	ISC***8
1_03	1_03. S, SI, SII?, C, CI, CII	2_03	2_03. D, DII,	3_03	3_03. ,	4_03	4_03. ,
low C	SC***	D	D	HiS	*	I	*
1_04	1_04. D, DI,	2_04	2_04. , SII, CII?	3_04	3_04. ,	4_04	4_04. ,
HiC	D	S	SD***	LoD	*	I	*
1_05	1_05. , III, SI,	2_05	2_05. , S, SII,	3_05	3_05. , CI?,	4_05	4_05. , CI,
LoI	IS***	C	S	LoD	C	S	C
1_06	1_06. , S, SI, SII,	2_06	2_06. ,	3_06	3_06. , CII	4_06	4_06. ,
D	S	S	*	LoI	C	loC	*
1_07	1_07. D, C, CII	2_07	2_07. ,	3_07	3_07. , SII, C, CI, CII	4_07	4_07. , II,
loC	DC ***	HiS	*	D	SC***	I	I
1_08	1_08. , II, CI,	2_08	2_08. D?,	3_08	3_08. ,	4_08	4_08. , C, CII
HiD	IC***	HiI	D	LoS		LoC	C
1_09	1_09. , III,	2_09	2_09. , SII, C,	3_09	3_09. , SI,	4_09	4_09. , SI,
I	I	S	SC***	D	S	HiC	S
1_10	1_10. D, DII,	2_10	2_10. , II, CII	3_10	3_10. , CI,	4_10	4_10. , C, CI,
D	D	HiC	IC***	LoI	C	S	C
1_11	1_11. D,	2_11	2_11. D, DI,	3_11	3_11. ,	4_11	4_11. , S, SI,
I	D	S	D	loC	*	D	S
1_12	1_12. D, DI, III,	2_12	2_12. , DII, C?, CI,	3_12	3_12. ,	4_12	4_12. , S, SII, C, CII
LoI	DI cons***	D	DC***	HiC	*	S	SC***
1_13	1_13. , I, II,	2_13	2_13. D, DI, CI,	3_13	3_13. D,	4_13	4_13. , II, CII
HiD	I	I	DC***	S	D	loC	IC***
1_14	1_14. , C, CII	2_14	2_14. , I, III,	3_14	3_14. ,	4_14	4_14. D, DI,
C	C	HiD	I	I	*	HiS	D
1_15	1_15. D,	2_15	2_15. , CII	3_15	3_15. ,	4_15	4_15. , S, SI,
HiS	D	--	C	C	*	LoD	S
1_16	1_16. , S, SI, SII, C, CI?,	2_16	2_16. D, S, SII?,	3_16	3_16. , II,	4_16	4_16. , DII,
HiI	SC***	LoS	DS***	LoC	I	D	D
1_17	1_17. ,	2_17	2_17. , CI, CII	3_17	3_17. ,	4_17	4_17. , I, III, S, SI,
HiC	*	S	C	LoI	*	D	IS***
1_18	1_18. ,	2_18	2_18. , DI, DII,	3_18	3_18. , C, CII	4_18	4_18. , S, SI,
HiI	*	HiS	D	loC	C	D	S

PAC^{I}	**IEC/summary**	**PAC^{I}**	**IEC/summary**	**PAC^{I}**	**IEC/Summary**	**PAC^{I}**	**IEC/Summary**
1_19	1_19. ,	2_19	2_19. D, DII,	3_19	3_19. , S, SI, C,	4_19	4_19. , CI,
HiC	*	D	D	I	SC***	S	C
1_20	1_20. D,	2_20	2_20. , SI, SII,	3_20	3_20. D, DI,	4_20	4_20. , I, III, S, C, CI,
D	D	HiC	S	LoS	D	I	ISC****
1_21	1_21. , III,	2_21	2_21. , S, SII?,	3_21	3_21. , SI?, C, CI, CII	4_21	4_21. ,
HiS/LoI	I	LoS	S	D	SC***	C	*
1_22	1_22. , III,	2_22	2_22. , C, CII	3_22	3_22. , S, SI,	4_22	4_22. ,
I	I	loC	C	D	S	S	*
1_23	1_23. , CI,	2_23	2_23. , II,	3_23	3_23. , C, CI?,	4_23	4_23. , SII,
I	C	HiC	I	D	C	LoS	S
1_24	1_24. , DII,	2_24	2_24. , C,	3_24	3_24. , CI,	4_24	4_24. , I, III,
D	D	S	C	I	C	C	I

NOTE: PAC^{I}=Pre-assigned construct with item number, IEC= Item Expressed Construct. D,I,S,C= DISC in Self Mask; DI,II,SI,CI= DISC in Work Mask (most); DII,III,SII,CII= DISC in Pressure Mask (lest). *=no dominant IEC under RICC. ***=Item contamination with 2 constructs. ****=item contamination with 3 constructs. ?=Item swing leads to uncertainty of IEC. A question mark represents uncertainty.

Interpretation note: for **Hi-D, Hi-I, Hi-S, Hi-C**, it does not imply 'High - that denoted the scoring method used with the PPA. This type of item only scored when item marked 'M'. Also **lo-D, I, S, & C** items do not imply 'low' - that denoted the scoring method used with the PPA and only scored when marked with 'L'. Most items in PPA are assigned with one PAC only. However item 1_21 has assigned for IS construct. Lo- and Hi- should read as follows: Lo-items are commonly selected, therefore, only when marked as 'L', they mean something, Lo-items are normally 'popular,' or 'preferable' terminologies. When marked as 'M' they do not count. Hi-items are normally 'extreme' or 'unpopular' terminologies. Most people would mark them 'L'. Therefore, such scores are only considered when marked with 'M.' 'L' does not count.

5.4. RICC result summary

5.4.1 Item contamination

M-positive (Most-positive) reported 25 contaminated items, *L-negative* (*Least-negative)* reported 21 contaminated items, and *Non-positive* reported 24 contaminated items. Within all these contaminated items, items 2_09 and 2_16 were found to be contaminated within all three analyses.

Items 1_01, 1_02, 1_08, 1_12, 1_16, 2_02, 2_09, 2_16, 3_21, and 4_17 were found to be contaminated in both *Non-positive* and *Most-positive* RICC. Items 4_21, 2_16, 2_11, 2_09, and 2_07 were found to be contaminated in both *Least-negative* and *Most-positive* RICC.

Items 1_03, 1_07, 2_04, 2_09, 2_16, 3_01, and 4_12 were found to be contaminated in *Least-negative* and *non-positive*. It was interesting to observe that *Non-positive* actually shared 10 similarities with *Most-positive*, and 7 similarities with *Least-negative*. In contrast, according to the theoretical postulates of this study, *Least-negative* and *Most-positive* were expected to be similar, yet only exhibited 5 similarities.

5.4.2 Weak items

With regard to the weak items, the analyses found that they had less in common. There were only 4 items shared between *Least-Negative* and *Most-Positive* (1_18, 3_12, 2_13, 3_19). There were 5 items shared between *Least-negative* and *Non-positive* (1_19, 3_03, 3_04, 3_08, 3_12). There were 4 items shared between *Most-positive* and *Non-positive* (1_17, 3_12, 3_14, 3_15). Item 3_12 was the only item that showed weak in all three analyses.

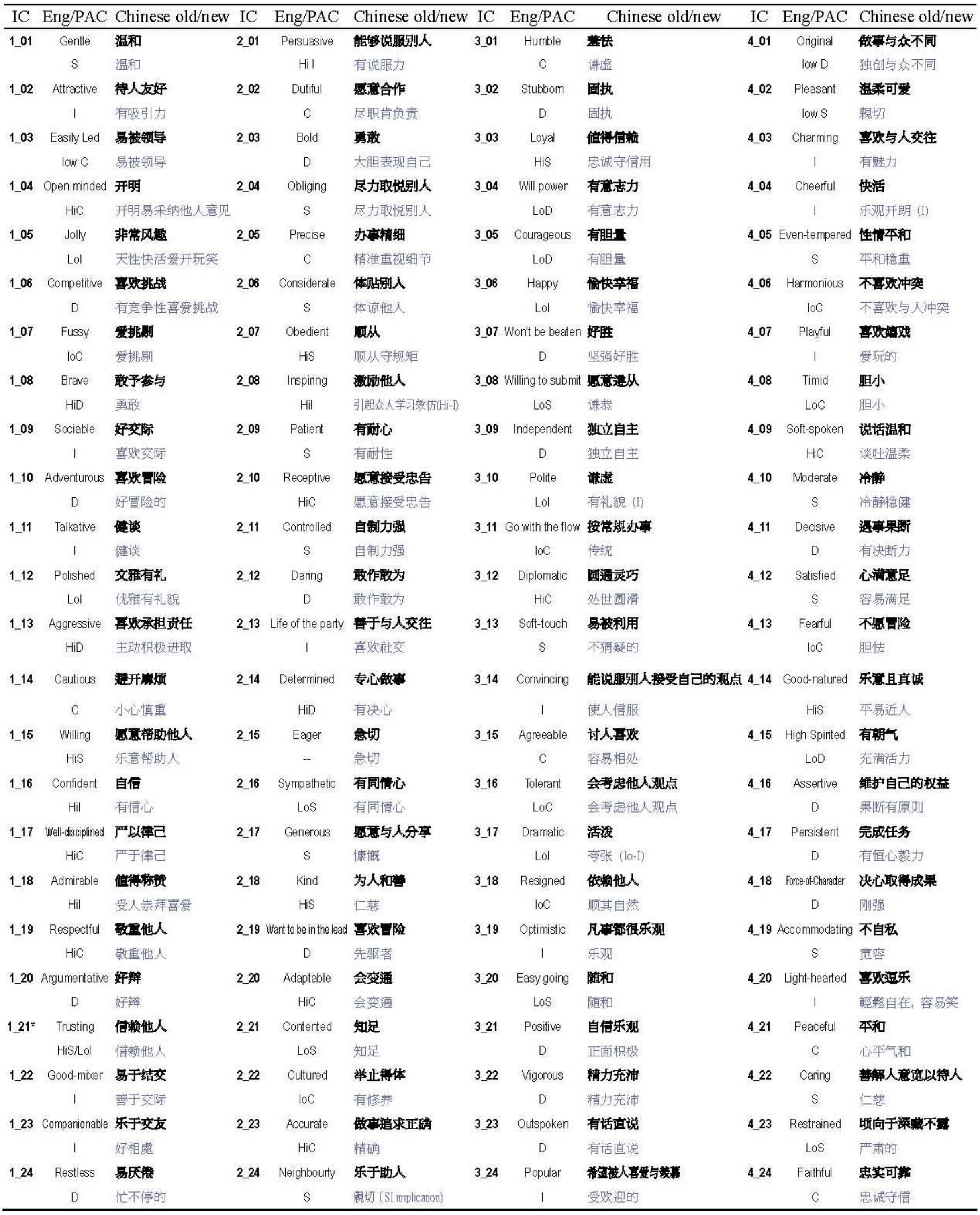
5.5. Amend A: Identifying 15 problematic items for amendment

Result gathered from RICC, as well as the textual interpretation of old Chinese form (in discussions by the researcher and local Chinese researcher), this research suggested various items should be amended for better performance. Two bilingual researchers' opinions were canvassed for possible alterations. The two researchers' language backgrounds were. 1. Chinese researcher: (China) Simplified Chinese and English. 2. South African researcher: (Taiwan) Traditional Chinese and English.

Both researchers had the Thomas standardised psychometric training and more than one year's practical experience with the instrument.

The amendment (Amend A) suggestion was decided via consent from both researchers and the statistical results. The final decision was only made when all the statistical analyses showed congruent results. Therefore the 15 items were recommended (Amend A). They are: 4_02, 2_04, 4_04, 3_06, 3_08, 3_10, 3_12, 1_13, 3_13, 3_14, 3_15, 4_16, 1_18, 1_24, and 4_24. However, items such as 3_01, 2_02, 2_03, 3_04, 1_07, 3_07, 4_07, 3_09, 4_14, 1_16, 2_17, 4_17, 1_19, 3_19, 2_20, 4_20, 1_21, 3_21, 1_22, and 2_22 might also need amendment to retain CPPA's construct integrity. The *Amend A* was therefore assigned for alteration (Table 5.15).

Table 5.15 PAC for CPPA, original items, and alteration suggestion (in grey)(Amend A)



IC	Eng/PAC	Chinese old/new	IC	Eng/PAC	Chinese old/new	IC	Eng/PAC	Chinese old/new	IC	Eng/PAC	Chinese old/new
1_01	Gentle	温和	2_01	Persuasive	能够说服别人	3_01	Humble	羞怯	4_01	Original	做事与众不同
	S	温和		Hi I	有说服力		C	谦虚		low D	独创与众不同
1_02	Attractive	待人友好	2_02	Dutiful	愿意合作	3_02	Stubborn	固执	4_02	Pleasant	温柔可爱
	I	有吸引力		C	尽职肯负责		D	固执		low S	親切
1_03	Easily Led	易被领导	2_03	Bold	勇敢	3_03	Loyal	值得信赖	4_03	Charming	喜欢与人交往
	low C	易被领导		D	大胆表现自己		HiS	忠诚守信用		I	有魅力
1_04	Open minded	开明	2_04	Obliging	尽力取悦别人	3_04	Will power	有意志力	4_04	Cheerful	快活
	HiC	开明易采纳他人意见		S	尽力取悦别人		LoD	有意志力		I	乐观开朗 (I)
1_05	Jolly	非常风趣	2_05	Precise	办事精细	3_05	Courageous	有胆量	4_05	Even-tempered	性情平和
	LoI	天性快活爱开玩笑		C	精准重视细节		LoD	有胆量		S	平和稳重
1_06	Competitive	喜欢挑战	2_06	Considerate	体贴别人	3_06	Happy	愉快幸福	4_06	Harmonious	不喜欢冲突
	D	有竞争性喜爱挑战		S	体谅他人		LoI	愉快幸福		loC	不喜欢与人冲突
1_07	Fussy	爱挑剔	2_07	Obedient	顺从	3_07	Won't be beaten	好胜	4_07	Playful	喜欢嬉戏
	loC	爱挑剔		HiS	顺从守规矩		D	坚强好胜		I	爱玩的
1_08	Brave	敢予参与	2_08	Inspiring	激励他人	3_08	Willing to submit	愿意遵从	4_08	Timid	胆小
	HiD	勇敢		HiI	引起众人学习效仿(Hi-I)		LoS	谦恭		LoC	胆小
1_09	Sociable	好交际	2_09	Patient	有耐心	3_09	Independent	独立自主	4_09	Soft-spoken	说话温和
	I	喜欢交际		S	有耐性		D	独立自主		HiC	谈吐温柔
1_10	Adventurous	喜欢冒险	2_10	Receptive	愿意接受忠告	3_10	Polite	谦虚	4_10	Moderate	冷静
	D	好冒险的		HiC	愿意接受忠告		LoI	有礼貌 (I)		S	冷静稳健
1_11	Talkative	健谈	2_11	Controlled	自制力强	3_11	Go with the flow	按常规办事	4_11	Decisive	遇事果断
	I	健谈		S	自制力强		loC	传统		D	有决断力
1_12	Polished	文雅有礼	2_12	Daring	敢作敢为	3_12	Diplomatic	圆通灵巧	4_12	Satisfied	心满意足
	LoI	优雅有礼貌		D	敢作敢为		HiC	处世圆滑		S	容易满足
1_13	Aggressive	喜欢承担责任	2_13	Life of the party	善于与人交往	3_13	Soft-touch	易被利用	4_13	Fearful	不愿冒险
	HiD	主动积极进取		I	喜欢社交		S	不猜疑的		loC	胆怯
1_14	Cautious	避开麻烦	2_14	Determined	专心做事	3_14	Convincing	能说服别人接受自己的观点	4_14	Good-natured	乐意且真诚
	C	小心慎重		HiD	有决心		I	使人信服		HiS	平易近人
1_15	Willing	愿意帮助他人	2_15	Eager	急切	3_15	Agreeable	讨人喜欢	4_15	High Spirited	有朝气
	HiS	乐意帮助人		--	急切		C	容易相处		LoD	充满活力
1_16	Confident	自信	2_16	Sympathetic	有同情心	3_16	Tolerant	会考虑他人观点	4_16	Assertive	维护自己的权益
	HiI	有信心		LoS	有同情心		LoC	会考虑他人观点		D	果断有原则
1_17	Well-disciplined	严以律己	2_17	Generous	愿意与人分享	3_17	Dramatic	活泼	4_17	Persistent	完成任务
	HiC	严于律己		S	慷慨		LoI	夸张（lo-I)		D	有恒心毅力
1_18	Admirable	值得称赞	2_18	Kind	为人和善	3_18	Resigned	依赖他人	4_18	Force-of-Character	决心取得成果
	HiI	受人崇拜喜爱		HiS	仁慈		loC	顺其自然		D	刚强
1_19	Respectful	敬重他人	2_19	Want to be in the lead	喜欢冒险	3_19	Optimistic	凡事都很乐观	4_19	Accommodating	不自私
	HiC	敬重他人		D	先驱者		I	乐观		S	宽容
1_20	Argumentative	好辩	2_20	Adaptable	会变通	3_20	Easy going	随和	4_20	Light-hearted	喜欢逗乐
	D	好辩		HiC	会变通		LoS	随和		I	輕鬆自在，容易笑
1_21*	Trusting	信赖他人	2_21	Contented	知足	3_21	Positive	自信乐观	4_21	Peaceful	平和
	HiS/LoI	信赖他人		LoS	知足		D	正面积极		C	心平气和
1_22	Good-mixer	易于结交	2_22	Cultured	举止得体	3_22	Vigorous	精力充沛	4_22	Caring	善解人意宽以待人
	I	善于交际		loC	有修养		D	精力充沛		S	仁慈
1_23	Companionable	乐于交友	2_23	Accurate	做事追求正确	3_23	Outspoken	有话直说	4_23	Restrained	倾向于深藏不露
	I	好相處		HiC	精确		D	有话直说		LoS	严肃的
1_24	Restless	易厌倦	2_24	Neighbourly	乐于助人	3_24	Popular	希望被人喜爱与羡慕	4_24	Faithful	忠实可靠
	D	忙不停的		S	親切（SI implication)		I	受欢迎的		C	忠诚守信

IC=Item code, PAC=Reassigned Construct *1_21 is the only item that is shared between S and I construct.

5.6. New PPA: Comparison of the differences

5.6.1 Classical Test Theory (CTT): New form

5.6.1.1 Item difficulty (P) and item discrimination index (D)

When compared with the old item, the amendment successfully increased the item difficulty (P) and item discrimination index (D). Out of 15 amended items, 12 items showed improvement on difficulty (P), and 12 items also showed improvement on item discrimination index (D).

However, as an affective test, FC items do not show good item difficulty in general. In the current study both the before (old form) and after (amended new form) show poor average difficulty (P) (< .25). In the current study, the discrimination index (D) is considered as a better measurement benchmark for affective psychometrics (Table 5.16). This study further indicates (see Table 5.16) that when items are changed according to IRT results, the discrimination index is generally increased (ranging from .008 to .4394).

Table 5.16 Difference between item analyses in amended items (Amend A)

	Old Form (n=650)		New Form (n=307)		Difference	
Item	**Difficulty index (P)**	**Discrimination index (D)**	**Difficulty index (P)**	**Discrimination index (D)**	**Change in (P)**	**Change in (D)**
D_13 Aggressive Hi D 主動積極進取 (1_13)	0.4198	0.3457	0.4276	0.4605	0.0078	0.1148
D_16 Assertive D 果斷有原則 (4_16)	0.0247**	0.0370***	0.2237**	0.3421	0.1990	0.3051
D_24 Restless D 忙不停的 (1_24)	0.0370**	0.0370***	0.0855**	0.0658***	0.0485	0.0288
I_04 Cheerful I 樂觀開朗 (4_04)	0.1543*	0.2840	0.3684*	0.4211	0.2141	0.1371
I_06 Happy Lo I 輕鬆 (3_06)	0.1111**	0.1605***	0.1053**	0.1579***	-0.0058	-0.0026
I_10 Cordial (polite) Lo I 友善 (3_10)	0.2438**	0.1790***	0.2566*	0.1184***	0.0128	-0.0606
I_14 Convincing I 有說服力 (3_14)	0.2747*	0.2901	0.1908**	0.3289	-0.0839	0.0388
I_18 Admirable Hi I 受人喜愛 (1_18)	0.0926**	0.1235***	0.1382**	0.1447***	0.0456	0.0212
S_02 Pleasant Lo S 親切 (4_02)	0.0154**	0.0062***	0.1184**	0.1579***	0.1030	0.1517
S_04 Obliging S 體貼 (2_04)	0.0494**	0.0370***	0.2171**	0.4079	0.1677	0.3709
S_08 Submissive (willing to submit) Lo S 善於禮讓 (3_08)	0.1019**	0.1790***	0.4013	0.6184	0.2994	0.4394
S_13 Soft-touch S 易被利用 (3_13)	0.0525**	0.0062***	0.0789**	0.1579***	0.0264	0.1517
C_12 Diplomatic Hi C 處世圓滑 (3_12)	0.2068**	0.2654	0.0987**	0.1447***	-0.1081	-0.1207
C_15 Agreeable C 容易相處 (3_15)	0.0772**	0.0309***	0.3684*	0.2632	0.2912	0.2323
C_24 Faithful C 忠誠守信 (4_24)	0.4259	0.3333	0.5132	0.3421	0.0873	0.0088

Note: *=Difficulty index (P) below .4, **=Difficulty index(p) below .25, ***=Discrimination Index (D) below .19.( )=negative value

5.6.2 Item Response Theory (IRT): new form

5.6.2.1 RICC method and Kendall's Tau B: result of experimental method

The old form analysis was done via RICC, without any specific parameter-estimation technique. The interpretation of item expressed construct (IEC) was assigned to two psychology master students, who had no pre-knowledge of the PPA pre-assigned construct (PAC). When the RICC 'M' curve swing went up to the right side, it was interpreted as suggesting a positive response towards the competing construct (refer to Chapter 4 section 4.8.1).

The main difference between RICC and IRT was that RICC used the construct percentile as the latent trait. In contrast, the traditional IRT method and GRM used the 'total item response' as the latent trait. The RICC was highly time-consuming and difficult to fully standardise (due to human interpretation), and another method was therefore used in this study to compare IRT and RICC. This alternative method needed to be more efficient, easier to standardise, and with a lower level I of human intervention. The Kendall's Tau-B (τ) rank correlation technique was selected and it was conducted using SPSS (version 12) (Shaw, 2007).

For the calculation of Kendall's Tau-B, this study also further the flattened the FC items into the interval/ordinal values M=1; blank=0; and L=-1. The PPA construct percentiles were calculated through ordinal rank correlation. According to Rupinski and Dunlap (1996), Kendall's Tau-B (τ) has better approximation towards estimation of Pearson's *r* than Spearman's rank correlation ρ (rho) (Rupinski & Dunlap, 1996). It is also suggested by Shaw (2007) that Spearman's ρ cannot be interpreted as clearly as Kendall's Tau, and the distributions do not follow normal distribution when the samples are small. The full table for PPA Kendall's Tau B and Spearman's ρ rank correlation results (all against old form, n=650) are included in the Appendix 3 (Table C.1~C.6, for the old form, on the CD).

Three correlational methods, Kendall's Tau B, Spearman's ρ, and Pearson's *r* were used to compare the results (Appendix 3, Table C.1-C.3 on the CD). In general, RICC and Kendall's Tau B agreed with each other. The areas highlighted in blue were the sections that were similar (RICC and KTB similar). The comparison showed that 46.875% of both results (45/96) agreed with each other (highlighted by blue mark), 21.875% (21/96) was partially in agreement (highlighted by green mark), and 31.25% disagreed with each other (30/96) (please see Appendix 3, Table C.1~C.3 on the CD). All three methods showed certain levels of similarity. The result indicated that Kendall's Tau was closest to RICC (Appendix 3, Table C.1 on the CD).

This study further compared the similarity between RICC and Kendall's Tau B, in two different benchmarks (Appendix 3, Table C.4~C.6, CD), which were (τ >0.09; τ <-0.09), (τ >0.19; τ <-0.19), and (τ >0.249; τ <-0.249). It was worth noticing that similarity was highest in (τ >0.249; τ <-0.249). The results were a 53.125% (51/96) match, 17.7083% (17/28) partial match, and 29.1667% (28/96) were unmatched. It was therefore suggested that RICC performs most similar to Kendall's Tau B within the range τ =.25 ~.10. In high (τ =.25), items considered to be 'contaminated' are not be shown. They are only shown when low Tau (τ =.0.001) was used (see the result summary, Table 5.17).

These findings there showed that an algorithm such as Kendall's Tau is similar to the IRT approach in terms of which the RICC generates information in both 'high' and 'low' ability areas.

However, in the case of Kendall's Tau, data were 'merged' into a single value, which raises the possibility that some important information may have been missed. RICC, on the other hand, gives the full picture of how ability and item function throughout all ability areas.

Table 5.17 Matching between RICC and different Kendall's Tau B Strength (old form n=650)

Kendall's Tau-B (τ) Strength	Match	Partial Match	Not Match
(τ >0.249; τ <-0.249)	51	17	28
	53.1250%	17.7083%	29.1667%
(τ >0.19; τ <-0.19)	45	21	30
	46.8750%	21.8750%	31.2500%
(τ >0.09; τ <-0.09)	47	18	31
	48.9583%	18.7500%	32.2917%
(τ >0.0001; τ <-0.0001)	37	29	30
	38.5417%	30.2083%	31.2500%

It is proposed that the possibility of also generating the RICC in concert with Kendall's Tau should be explored in future. The RICC method seemed to be better at examining the contamination effect of the PPA FC tetrads, but when RICC is not accessible, Kendall's Tau can be used. However, both results (τ >0.249; τ <-0.249), and (τ >0.0001; τ <-0.0001) should ideally be generated to yield the maximum information regarding the 'depth' of the 'contamination' effect.

Future researchers should first conduct Kendall's Tau B with (τ >0.249; τ <-0.249). If the item express construct (IEC) generated from Kendall's Tau (τ >0.249; τ <-0.249) is the same as-the pre assigned construct (PAC) it can be concluded that the item is functioning appropriately. If the IEC shows a difference to PAC, or nothing appears at all, this could be an indication of item contamination. The researcher should then investigate items dynamics and the depth of item contamination.

5.6.2.2 *New item comparison: amended items and Kendall's Tau B*

The 15 amended items generally showed an improvement in constructs. However, other items that did not go through the amendment process remained the same. Samejima's Graded Response Model showed that other items that not undergo any change retained the same item parameter.

This provides some evidence for the validity of IRT, because 15 items undergoing amendment also have an impact on the ICC of other items within the same item set (tetrad). The current research had amended 15 items (15 items in 12 tetrads) from the old form. The results were generally satisfactory. 73.3% (11 out of 15 items) are reported as successful, 26.6% (4 out of 15 items) were still considered problematic (see Table 5.18).

Table 5.18 Summary of Kendall's Tau B (τ) analysis comparison of 11 amended items

Item code	KTB Level (τ >n)	Before	After	Note	Result
D_13	0.001	DI,SI,CI,DII,SII,D,S	DI,CI,DII,D	Higher D link	Successful
	0.099	DI,DII,D	DI,DII,D		
	0.199	DI,D	DI,DII,D		
	0.249				
D_16	0.001	DI,II,DII,III,D,I	DI,II,DII,III,D,I	Lower I	Successful
	0.099	DI,II,DII,D,I	DI,II,DII,D,I		
	0.199	DI,II,D,I	DI,II,D		
	0.249	II	II		
D_24	0.001	DI,SI,DII,SII,CII,D,S,C	DI,DII,CII,D,C	Higher D	Successful
	0.099	DII	DII,CII,D		
	0.199				
	0.249				
I_04	0.001	II,III,I	DI,II,DII,III,D,I	Higher I	Successful
	0.099	II,III,I	II,III,I		
	0.199	III,I	II,III,I		
	0.249				
I_06	0.001	II,SI,III,I,S	II,SI,III,I,S		No difference
	0.099	III,I,	III,I,		
	0.199	III,	III,		
	0.249				
I_10	0.001	SI,III,SII,CII,S,C	SI,SII,CII,S,C,	Higher S	unsuccessful
	0.099	SI,SII,CII,S	SI,SII,CII,S,		
	0.199	S	SII,S,		
	0.249				
I_14	0.001	DI,II,DII,III,D,I	DI,II,DII,III,D,I	Higher I	Successful
	0.099	DI,II,DII,III,D,I	DI,II,DII,III,D,I	,but also higher D	
	0.199	II,I	II,DII,III,D,I		
	0.249	II	II		
I_18	0.001	DI,II,DII,III,D,I	II,III,I	Higher I	Successful
	0.099	II,I	II,III,I		
	0.199	II	III,I		
	0.249				
S_02	0.001	II,SI,CI,III,SII,CII,I,S,C	SI,SII,CII,S	Higher S	Successful

Item code	KTB Level (τ >n)	Before	After	Note	Result
	0.099	III,SII,S	SI,SII,S		
	0.199	SII	SII		
	0.249				
S_04	0.001	SI,SII,CII,S,C	SI,CI,SII,CII,S,C	Higher S	Successful
	0.099	SII,CII,S	SI,SII,CII,S,C	,but also higher C	
	0.199		SI,SII,CII,S,C		
	0.249				
S_08	0.001	SI,CI,SII,CII,S,C	SI,CI,SII,CII,S,C	Higher S	Successful
	0.099	SII,S	SI,SII,CII,S,C	,but also higher C	
	0.199		SI,SII,CII,S,C		
	0.249				
S_13	0.001	DI,SI,DII,SII,D,S	SI,CI,SII,S,C	Higher S	Successful
	0.099	SII,S	SII,S		
	0.199		SII,S		
	0.249				
C_12	0.001	DI,II,CI,III,D,I,C	II,CI,III,CII,I,C	Higher I	unsuccessful
	0.099	II,CI,I,C	II,III,I		
	0.199				
	0.249				
C_15	0.001	II,SI,CI,III,SII,CII,I,S,C	II,SI,CI,III,SII,CII,I,S,C	Higher C	Successful
	0.099	II,III,CII,I,C	CI,SII,CII,C		
	0.199	III,I			
	0.249				
C_24	0.001	SI,CI,SII,CII,S,C	DI,SI,CI,DII,SII,D,S,C		No difference
	0.099	CI,C	CI,C		
	0.199	CI,C	CI		
	0.249	CI	CI		

Note: Overall, 11 items *(73.3%)* are considered as successful. 2 items (13.3%) with no difference, 2 items (13.3%) are considered as unsuccessful. However, within successful items, three items (25%) contain mixed results.

Within the 12 tetrads that underwent amendment, only 9 had one item that was altered and 3 tetrads had two items altered (set 4, 13, and 24). When one item within the tetrad improved, the other items also improved, as demonstrated by tetrad 4, 6, 13, 14, 15, 16, 18, and 24 (see Table 5.19 - 5.30).

However, it was found with tetrads 10 and 12 that when an item was unsuccessful, the other item would also be contaminated (see Table 5.24 and 5.31).

According to the current result, the KTB method could illustrate such an effect (see Table 5.18 for full details and Appendix 3, Table C.7 and Appendix 4 on the CD).

D construct KTB summary

All three amendments were reported to be successful. D_13 (Table 5.19) and D_16 (Table 5.20) were very successful, the amendment strengthened the D item and purified it, even to as low as the $\tau > 0.001$ level. In contrast, D_24 (Table 5.21) only received moderate success firming up to $\tau > 0.099$ level.

However, such amendments seemed to purify item 3_24 indirectly. The I item lost its D connection. Overall the amendment of the D construct was also relatively easy in comparison to other constructs, most likely due to the fact that this construct was more easily defined and relatively ‘culture’ free.

Table 5.19 Kendall's Tau B (τ) analysis comparison after amendment, item set 13, (D_13) amended

KTB Level	**IEC**	**IEC**	**IEC**	**IEC**
Before(n=650)				
τ > 0.001	DI,SI,CI,DII,SII,D,S	II,III,I	DI,SI,DII,SII,D,S	II,SI,CI,CII,C
τ > 0.099	DI,DII,D	II,III,I	SII,S	CII,C
τ > 0.199	DI,D	II,III,I		CII,C
τ > 0.249		II,III,I		CII
After(n=307)				
τ > 0.001	DI,CI,DII,D	II,DII,III,I	SI,CI,SII,S,C	SI,CI,CII,S,C
τ > 0.099	DI,DII,D	II,III,I	SII,S	CII,C
τ > 0.199	DI,DII,D	II,III,I	SII,S	CII,C
τ > 0.249		II,III,I		CII
PAC	Hi-D*	I	S*	Lo-C
Item code	1_13*	2_13	3_13*	4_13
Item (Eng)	Aggressive*	Life of the party	Soft-touch*	Fearful,

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

Table 5.20 Kendall's Tau B (τ) analysis comparison after amendment, item set 16, (D_16) amended

KTB Level	IEC	IEC	IEC	IEC
Before(n=650)				
$\tau > 0.001$	DI,II,DII,III,D,I	SI,SII,CII,S	SI,CI,SII,CII,S,C	DI,CI,DII,CII,D,C
$\tau > 0.099$	DI,II,DII,D,I	SI,SII,S	SI,CI,SII,CII,S,C	DII,D
$\tau > 0.199$	DI,II,D,I	SII,S		DII
$\tau > 0.249$	II	SII,S		
After(n=307)				
$\tau > 0.001$	DI,II,DII,III,D,I	SI,CI,III,SII,CII,S,C	SI,CI,III,SII,CII,S,C	DI,DII,D
$\tau > 0.099$	DI,II,DII,D,I	SI,SII,CII,S,C	SI,CI,SII,CII,S,C	DI,DII,D
$\tau > 0.199$	DI,II,D	SI,SII,S	S,C	DI,DII,D
$\tau > 0.249$	II	SII,S		
PAC	Hi-I	Lo-S	Lo-C	D*
Item code	1_16	2_16	3_16	4_16*
Item (Eng)	Confident	Sympathetic	Tolerant	Assertive*

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

Table 5.21 Kendall's Tau B (τ) analysis comparison after amendment, item set 24, (D_24) amended

KTB Level	IEC	IEC	IEC	IEC
Before(n=650)				
τ > 0.001	DI,SI,DII,SII,CII,D,S,C	II,SI,DII,III,SII,I,S	DI,II,III,I	SI,CI,SII,CII,S,C
τ > 0.099	DII	SI,S	II,III,I	CI,C
τ > 0.199			II,III,I	CI,C
τ > 0.249				CI
After(n=307)				
τ > 0.001	DI,DII,CII,D,C	II,SI,III,SII,I,S	II,III,I	DI,SI,CI,DII,SII,D,S,C
τ > 0.099	DII,CII,D	SI,SII,S	II,III,I	CI,C
τ > 0.199		SI,S	II,III,I	CI
τ > 0.249				CI
PAC	D*	S	I	C*
Item code	1_24*	2_24	3_24	4_24*
Item (Eng)	Restless*	Neighbourly	Popular	Faithful*

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

I construct KTB summary

The amendments of I construct had mixed results. Tetrads considered successful were sets 4 (Table 5.22), 14 (Table 5.25), and 18 (Table 5.26) (the success of set 4 could be due to alteration of S construct). Although their results were shown to be mixed, the alteration seemed to strengthen other items within the tetrads. It was therefore suggested that the alteration of the above items had made the selection process clearer. This would mean that individuals who possess non I traits would not be confused by this item, making their choice of another item easier. However, tetrad 6 (Table 5.23) did not appear to change the dynamics at all (or inclined slightly towards poor alteration). The tetrad 10 (Table 5.24) amendment decreased the clarity within the set and therefore proved an unsuccessful attempt.

Table 5.22 Kendall's Tau B (τ) analysis comparison after amendment, item set 04, (I_04) amended

KTB Level	**IEC**	**IEC**	**IEC**	**IEC**
Before(n=650)				
τ > 0.001	SI,CI,SII,CII,S,C,	SI,SII,CII,S,C,	DI,DII,D	II,III,I
τ > 0.099	SI,CI,S,C,	SII,CII,S,	DI,DII,D	II,III,I
τ > 0.199	CI,		DI,D	III,I
τ > 0.249				
After(n=307)				
τ > 0.001	CI,III,SII,S,C	SI,CI,SII,CII,S,C	DI,DII,D	DI,II,DII,III,D,I
τ > 0.099	CI,C	SI,SII,CII,S,C	DI,DII,D	II,III,I
τ > 0.199	CI	SI,SII,CII,S,C	DI,DII,D	II,III,I
τ > 0.249				
PAC	Hi-C	S*	Lo-D	I*
Item code	1_04	2_04*	3_04	4_04*
Item (Eng)	Open Mind	Obliging*	Will power	Cheerful*

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

Table 5.23 Kendall's Tau B (τ) analysis comparison after amendment, item set 06, (I_06) Amended

KTB Level	**IEC**	**IEC**	**IEC**	**IEC**
Before(n=650)				
τ > 0.001	DI,II,DII,III,D,I	SI,CI,SII,S,C	II,SI,III,I,S	SI,CI,SII,CII,S,C
τ > 0.099	DI,DII,D	SI,SII,S	III,I,	CII,C
τ > 0.199	DI,DII,D	SI,SII,S	III,	CII,C
τ > 0.249	DI,DII,D	SI,SII,S		CII,C
After(n=307)				
τ > 0.001	DI,II,DII,D,I	SI,CI,SII,CII,S,C	II,SI,III,I,S	SI,CI,SII,CII,S,C
τ > 0.099	DI,DII,D	SI,CI,SII,S,C	III,I,	SII,CII,C
τ > 0.199	DI,DII,D	SI,SII,S	III,	CII,C
τ > 0.249	DI,DII,D	SI,SII,S		CII,C
PAC	D	S	LoI*	loC
Item code	1_06	2_06	3_06*	4_06
Item (Eng)	Competitive,	Considerate	Happy*	Harmonious,

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

Table 5.24 Kendall's Tau B (τ) analysis comparison after amendment, item set 10, (I_10) amended

KTB Level	**IEC**	**IEC**	**IEC**	**IEC**
Before(n=650)				
$\tau > 0.001$	DI,II,DII,D,I	II,CI,III,I,C	SI,III,SII,CII,S,C	SI,CI,SII,CII,S,C
$\tau > 0.099$	DI,DII,D	CI,C	SI,SII,CII,S	SII,S
$\tau > 0.199$	DI,DII,D	CI	S	
$\tau > 0.249$	DI,DII,D	CI		
After(n=307)				
$\tau > 0.001$	DI,II,DII,III,D,I	II,SI,CI,III,SII,CII,I,S,C	SI,SII,CII,S,C,	DI,SI,CI,DII,SII,CII,D,S,C,
$\tau > 0.099$	DI,II,DII,D,I	CI,C	SI,SII,CII,S,	
$\tau > 0.199$	DI,DII,D	CI	SII,S,	
$\tau > 0.249$	DI,DII,D	CI		
PAC	D	Hi-C	Lo-I*	S
Item code	1_10	2_10	3_10*	4_10
Item (Eng)	Adventurous,	Receptive	Cordial (polite)*	Moderate,

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

Table 5.25 Kendall's Tau B (τ) analysis comparison after amendment, item set 14, (I_14) amended

KTB Level	IEC	IEC	IEC	IEC
Before(n=650)				
τ > 0.001	SI,CI,SII,CII,S,C,	DI,CI,SII,CII,D,S,C,	DI,II,DII,III,D,I	SI,III,SII,S
τ > 0.099	SII,CII,C,	DI,	DI,II,DII,III,D,I	SI,S
τ > 0.199	CII,C,		II,I	SI
τ > 0.249	CII,C,		II	SI
After(n=307)				
τ > 0.001	SI,CI,SII,CII,S,C	DI,CI,DII,D	DI,II,DII,III,D,I	SI,SII,CII,S,C
τ > 0.099	SI,CI,SII,CII,S,C	DI,DII,D	DI,II,DII,III,D,I	SI,SII,S
τ > 0.199	CII,C	DI,D	II,DII,III,D,I	SI,S
τ > 0.249	CII,C		II	SI
PAC	C	Hi-D	I*	Hi-S
Item code	1_14	2_14	3_14*	4_14
Item (Eng)	Cautious,	Determined	Convincing*	Good-natured

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

Table 5.26 Kendall's Tau B (τ) analysis comparison after amendment, item set 18, (I_18) amended

KTB Level	**IEC**	**IEC**	**IEC**	**IEC**
Before(n=650)				
τ > 0.001	DI,II,DII,III,D,I	SI,SII,CII,S,C	SI,CI,SII,CII,S,C	DI,II,CI,DII,III,D,I
τ > 0.099	II,I	SI,SII,CII,S,C	SII,CII,C	DI,DII,D
τ > 0.199	II	SI,S	CII,C	DI,DII,D
τ > 0.249		SI,S	CII	DI,D
After(n=307)				
τ > 0.001	II,III,I	SI,CI,SII,CII,S,C	SI,III,SII,CII,S,C	DI,II,CI,DII,D
τ > 0.099	II,III,I	SI,SII,CII,S,C	SII,CII,C	DI,DII,D
τ > 0.199	III,I	SI,SII,S	CII,C	DI,DII,D
τ > 0.249		SI,S	CII	DI,D
PAC	Hi-I*	Hi-S	Lo-C	D
Item code	1_18*	2_18	3_18	4_18
Item (Eng)	Admirable*	Kind	Resigned	Force-of-Character

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

S construct KTB summary

The amendment of S construct was generally successful; two items were complete successes: item 2 (Table 5.27) and 13 (Table 5.30). Two items had mixed results 4 (Table 5.28) and 8 (Table 5.29). The tetrads 4 and 8 were reported to be mixed, or contaminated. The contamination was related to the S-C mix. Previous reports suggest that it is not uncommon to find SC mixed within S or C items (Hendrickson, 1983). This could be due to the fact that the S and C construct terminologies are naturally related in the target cultural terminology. However, the amendment of both sets 4 and 8 strengthened the items, and led to clear construct manifestation for other items within the tetrad. Therefore both sets 4 and 8 were defined as successful.

Table 5.27 Kendall's Tau B (τ) analysis comparison after amendment, item set 02, (S_02) amended

KTB Level	**IEC**	**IEC**	**IEC**	**IEC**
Before(n=650)				
τ > 0.001	II,SI,III,SII,CII,I,S	DI,CI,DII,III,D,C	DI,DII,D	II,SI,CI,III,SII,CII,I,S,C
τ > 0.099	II,SI,S	CI	DI,DII,D	III,SII,S
τ > 0.199		CI	DII,D	SII
τ > 0.249			DII,D	
After(n=307)				
τ > 0.001	II,SI,III,SII,CII,I,S	DI,CI,DII,D,C	DI,DII,CII,D	SI,SII,CII,S
τ > 0.099	II,SI,I,S	DI,CI,D	DI,DII,D	SI,SII,S
τ > 0.199			DII	SII
τ > 0.249			DII,D	
PAC	I	C	D	Lo-S*
Item code	1_02	2_02	3_02	4_02*
Item (Eng)	Attractive	Dutiful	Stubborn	Pleasant*

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

Table 5.28 Kendall's Tau B (τ) analysis comparison after amendment, item set 04, (S_04) amended

KTB Level	**IEC**	**IEC**	**IEC**	**IEC**
Before(n=650)				
τ > 0.001	SI,CI,SII,CII,S,C	SI,SII,CII,S,C	DI,DII,D	II,III,I
τ > 0.099	SI,CI,S,C	SII,CII,S	DI,DII,D	II,III,I
τ > 0.199	CI		DI,D	III,I
τ > 0.249	,			
After(n=307)				
τ > 0.001	CI,III,SII,S,C	SI,CI,SII,CII,S,C	DI,DII,D	DI,II,DII,III,D,I
τ > 0.099	CI,C	SI,SII,CII,S,C	DI,DII,D	II,III,I,
τ > 0.199	CI	SI,SII,CII,S,C	DI,DII,D	II,III,I
τ > 0.249				
PAC	Hi-C	S*	Lo-D	I*
Item code	1_04	2_04*	3_04	4_04*
Item (Eng)	Open Mind	Obliging*	Will power	Cheerful*

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

Table 5.29 Kendall's Tau B (τ) analysis comparison after amendment, item set 08, (S_08) amended

KTB Level	**IEC**	**IEC**	**IEC**	**IEC**
Before(n=650)				
τ > 0.001	DI,DII,III,D	II,DII,III,I	SI,CI,SII,CII,S,C	SI,CI,SII,CII,S,C
τ > 0.099	DI,DII,D	II,I	SII,S	CII,C
τ > 0.199	DI,D	II		CII,C
τ > 0.249	DI	II		CII,C
After(n=307)				
τ > 0.001	DI,II,DII,III,D,I	DI,II,CI,DII,III,D,I	SI,CI,SII,CII,S,C	SI,CI,SII,CII,S,C
τ > 0.099	DI,DII,D	II,III,I	SI,SII,CII,S,C	SI,SII,CII,S,C
τ > 0.199	DI,DII,D	II	SI,SII,CII,S,C	CII,C
τ > 0.249	DI	II		CII,C
PAC	Hi-D	Hi-I	Lo-S*	Lo-C
Item code	1_08	2_08	3_08*	4_08
Item (Eng)	Brave,	Inspiring	Submissive (willing to Submit)*	Timid

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

Table 5.30 Kendall's Tau B (τ) analysis comparison after amendment, item set 13, (S_13) amended

KTB Level	IEC	IEC	IEC	IEC
Before(n=650)				
τ > 0.001	DI,SI,CI,DII,SII,D,S	II,III,I	DI,SI,DII,SII,D,S	II,SI,CI,CII,C
τ > 0.099	DI,DII,D	II,III,I	SII,S	CII,C
τ > 0.199	DI,D	II,III,I		CII,C
τ > 0.249		II,III,I		CII,
After(n=307)				
τ > 0.001	DI,CI,DII,D	II,DII,III,I	SI,CI,SII,S,C	SI,CI,CII,S,C
τ > 0.099	DI,DII,D	II,III,I	SII,S	CII,C
τ > 0.199	DI,DII,D	II,III,I	SII,S	CII,C
τ > 0.249		II,III,I		CII
PAC	HiD*	I	S*	loC
Item code	1_13*	2_13	3_13*	4_13
Item (Eng)	Aggressive*	Life of the party	Soft-touch*	Fearful,

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

C construct KTB summary

Three items had been amended in the C construct. Two items were considered as successful 15 (Table 5.32) and 24 (Table 5.33), and one item as unsuccessful (set 12. see Table 5.31). Tetrad 15 showed good striping from I construct, which would leave the C item with only SC implications (before amend =ISC, after amend=SC construct, see Table 5.32). However, set 24 (Table 5.33) was considered as successful due to the clarification on the other items within the tetrad (general improvement on all DISC items). Tetrad 12 (Table 5.31) demonstrated a poor alteration in that the I and C constructs had switched places. It was therefore suggested that set 12 (Table 5.31) should undergo further amendment.

Table 5.31 Kendall's Tau B (τ) analysis comparison after amendment, item set 12, (C_12) amended

KTB Level	**IEC**	**IEC**	**IEC**	**IEC**
Before(n=650)				
τ > 0.001	SI,CI,III,SII,CII,S,C	DI,DII,D	DI,II,CI,III,D,I,C	SI,SII,CII,S,C
τ > 0.099	SI,SII,CII,S,C	DI,DII,D	II,CI,I,C	SI,SII,CII,S
τ > 0.199	SI,S	DI,DII,D		SII,S
τ > 0.249		DI,DII,D		SII,S
After(n=307)				
τ > 0.001	SI,III,SII,CII,I,S,C	DI,II,DII,D,I	II,CI,III,CII,I,C	SI,CI,SII,CII,S,C
τ > 0.099	SI,SII,S	DI,DII,D	II,III,I	SI,SII,CII,S,C
τ > 0.199		DI,DII,D		SII,S
τ > 0.249		DI,DII,D		SII,S
PAC	LoI	D	HiC*	S
Item code	1_12	2_12	3_12*	4_12
Item (Eng)	Polished	Daring	Diplomatic*	Satisfied

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

Table 5.32 Kendall's Tau B (τ) analysis comparison after amendment, item set 15, (C_15) amended

KTB Level	IEC	IEC	IEC	IEC
Before(n=650)				
τ > 0.001	SI,SII,CII,S,C	DI,DII,D	II,SI,CI,III,SII,CII,I,S,C	DI,II,DII,III,D,I
τ > 0.099	SI,SII,S	DI,DII,D	II,III,CII,I,C	DI,DII,D
τ > 0.199	SI,S		III,I	
τ > 0.249	SI,S			
After(n=307)				
τ > 0.001	DI,SI,CI,SII,S	DI,CI,CII,C	II,SI,CI,III,SII,CII,I,S,C	DI,II,DII,III,D,I
τ > 0.099	SI,S	CII,C	CI,SII,CII,C	DI,II,DII,III,D,I
τ > 0.199	SI			I
τ > 0.249	SI,S			
PAC	HiS	--	C*	LoD
Item code	1_15	2_15	3_15*	4_15
Item (Eng)	Willing,	Eager	Agreeable*	High Spirited

Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

Table 5.33 Kendall's Tau B (τ) analysis comparison after amendment, item set 24, (C_24) amended

KTB Level	IEC	IEC	IEC	IEC
Before(n=650)				
τ > 0.001	DI,SI,DII,SII,CII,D,S,C	II,SI,DII,III,SII,I,S	DI,II,III,I	SI,CI,SII,CII,S,C
τ > 0.099	DII	SI,S	II,III,I	CI,C
τ > 0.199			II,III,I	CI,C
τ > 0.249				CI
After(n=307)				
τ > 0.001	DI,DII,CII,D,C	II,SI,III,SII,I,S	II,III,I	DI,SI,CI,DII,SII,D,S,C
τ > 0.099	DII,CII,D	SI,SII,S	II,III,I	CI,C
τ > 0.199		SI,S	II,III,I	CI
τ > 0.249				CI
PAC	D*	S	I	C*
Item code	1_24*	2_24	3_24	4_24*
Item (Eng)	Restless*	Neighbourly	Popular	Faithful*

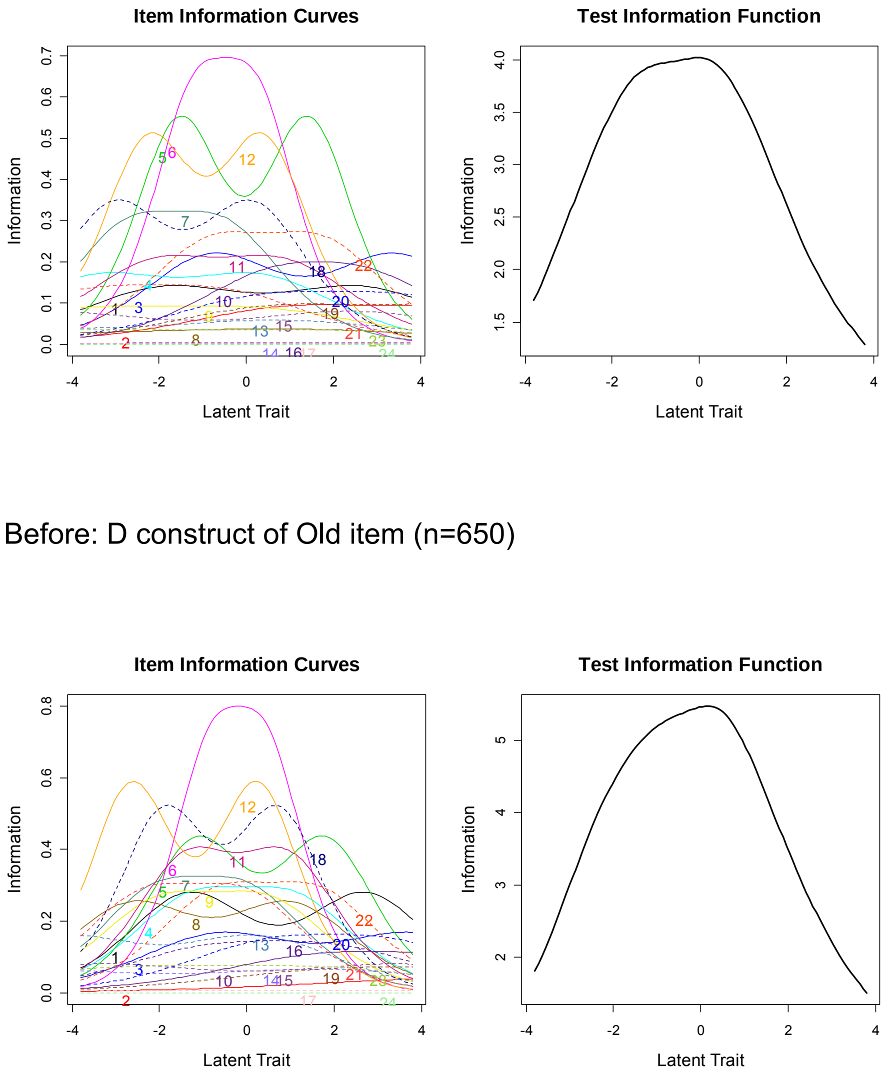
Note: IEC=Item Expressed Construct, PAC=Pre-assigned construct. *=Amended item within tetrad.

5.6.3 GRM comparison interpretation

5.6.3.1 Item Information Curve (IIC) and Test Information Function (TIF)

The full GRM IRCCC analyses are presented in Appendix 4 (on the CD). This section therefore presents the summary forms of GRM IRCCC, which are Item Information Curve (IIC), Test Information Function (TIF) and Item Response Category Characteristic Curves summary (IRCCC-s).

In IIC and TIF, the D construct showed an increment of testing information from 4 to 5.5. The I construct showed a left shift but increased item information from 3 to 4. The S construct showed a general improvement of 1.0~1.2 information. The C construct also showed a general improvement of item information (1.0~1.4) (for detail explanation of GRM morel please see section 4.9 ~ 4.11 in Chapter 4).



Before: D construct of Old item (n=650)

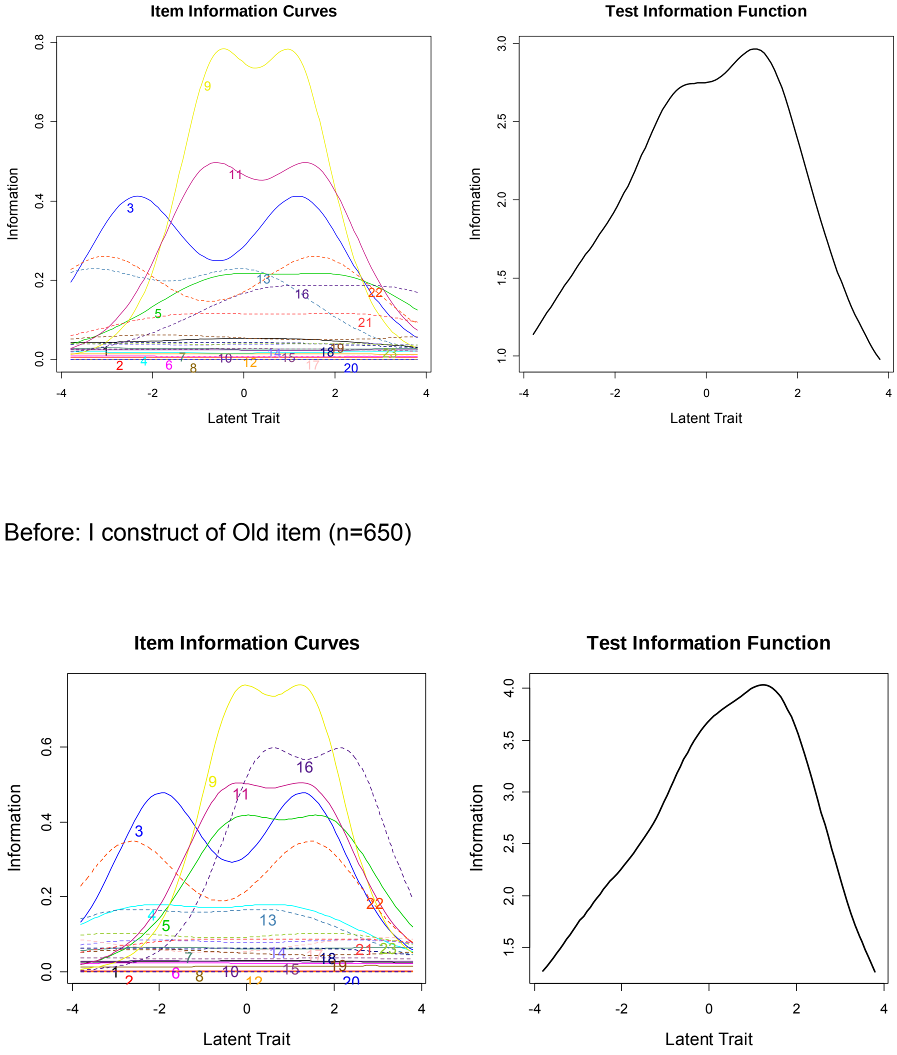
After: D construct of New item (n=307)

Figure 5.2 Comparing D construct in TIC and IIC after Amend A

D in TIF and IIC

According to the above IIC and TIF (Figures 5.2 and 5.3), the old and new D constructs all function well. The D construct showed good sample discrimination as illustrated by the centralised TIF curve (optimal information around theta -2~1.5, information level around 3.8~4.0, see above TIF).

The New D construct had also improved its item information value around theta 1.5 with the information value level increasing up to 5.5 (see the New D TIF). This showed that D construct's had improved after *Amend A*.



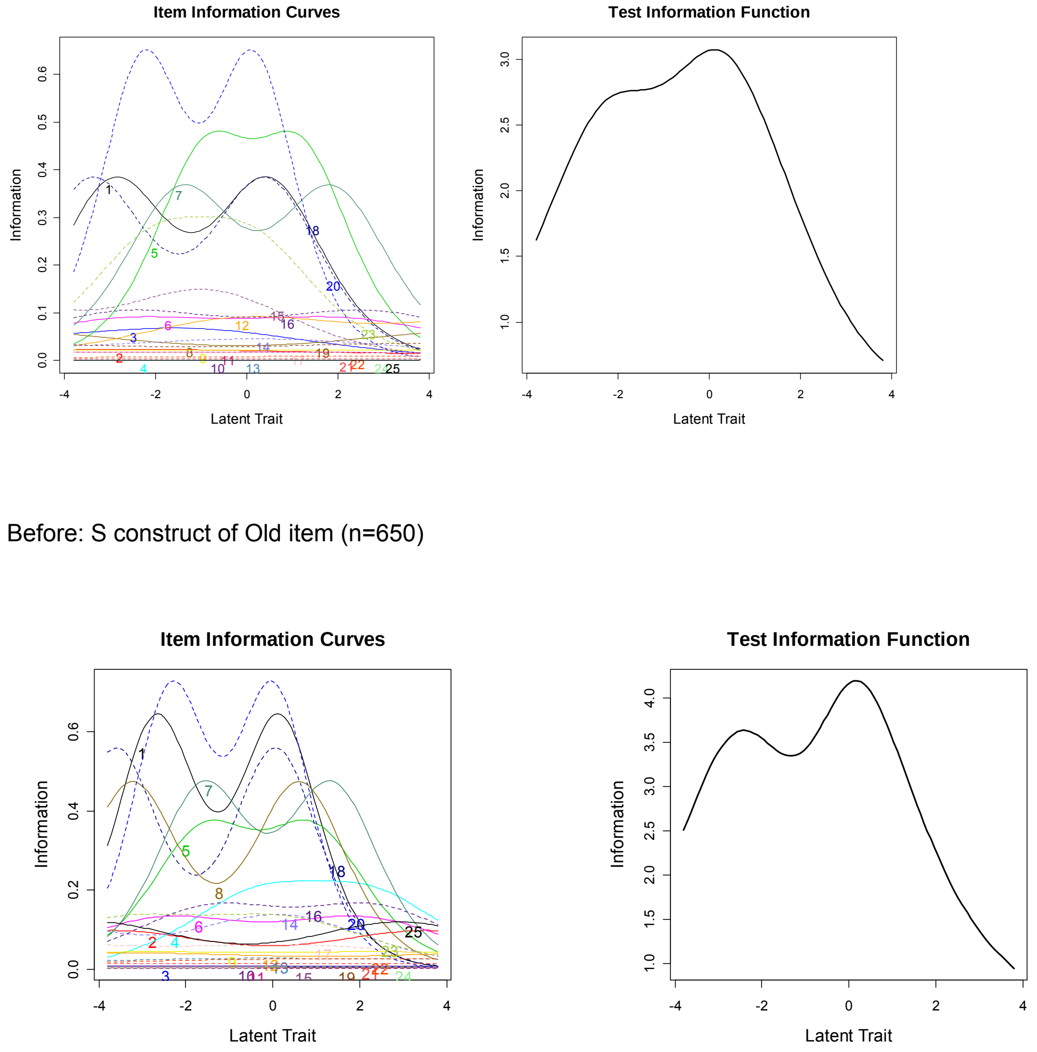
Before: I construct of Old item (n=650)

After: I construct of New item (n=307)

Figure 5.3 Comparing I construct in TIC and IIC after Amend A

I in TIF and IIC

In above graphs (Figure 5.3), the old I construct's item information value peaks in theta 1.7~2.0. This shows that I items generate more information for individuals who really are high in I construct. This in turn implies that I items in general could be too simple, or socially preferable, to be marked as 'M (most)'. According to above the Test Information Function (TIF), only small numbers of respondents would mark items such as item 9, 16, 11, and 5 as 'M (most)'. This implies that the current item sets generate information in the high I group. The new I performs in a similar manner, although the item information value had increased from 3 to 4.



Before: S construct of Old item (n=650)

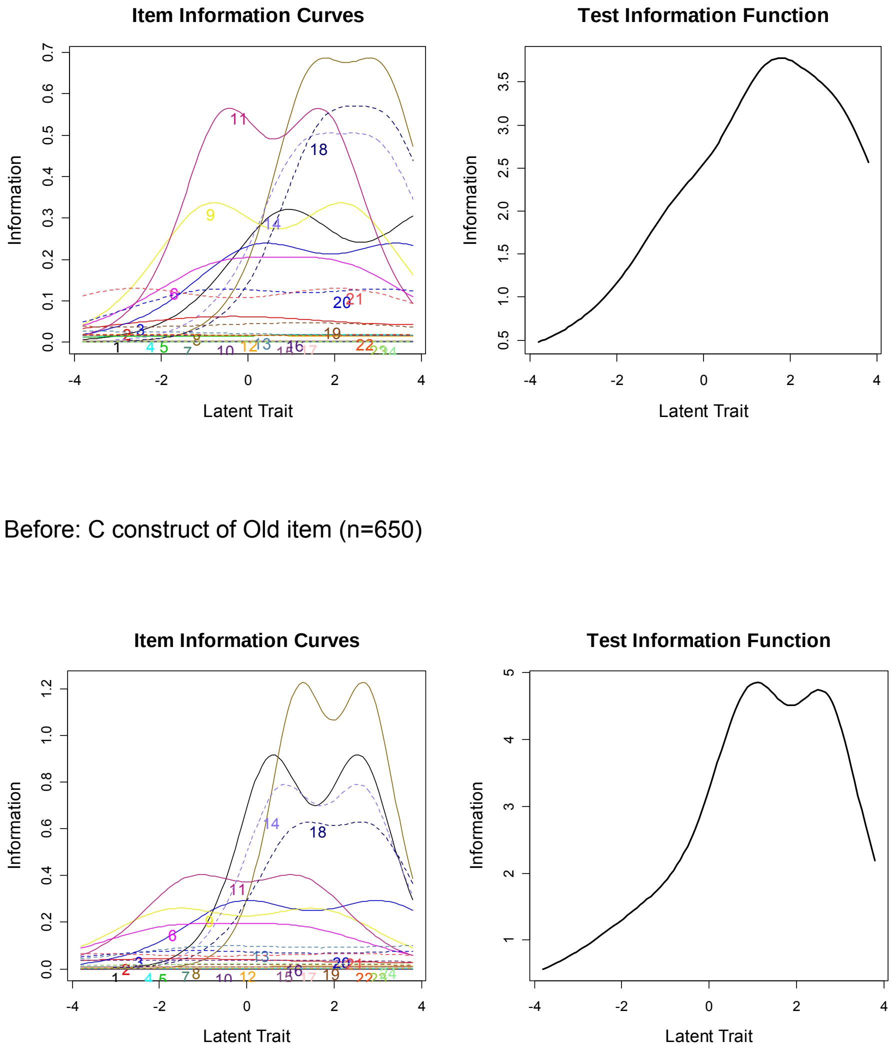
After: S construct of New item (n=307)

Figure 5.4 Comparing S construct in TIC and IIC after Amend A

S in TIF and IIC

In the above graphs (Figure 5.4), the S construct indicates two peaks of information. One peak is in theta -3~-2 and another is in -1~ 0.6. This shows that the S constructs differentiate well for individuals who are either not in the S construct (-3~-2) or have an average (-1~0.6) rating in the S constructs. However, the S construct do not to differentiate well for individuals high in the S construct.

The new items in the S construct indicate a general increase in testing information; the first peak increases from 2.7 to 3.6. The second peak increases from 3.0 to 4.3.



Before: C construct of Old item (n=650)

After: C construct of New item (n=307)

Figure 5.5 Comparing C construct in TIC and IIC after Amend A

C in TIF and IIC

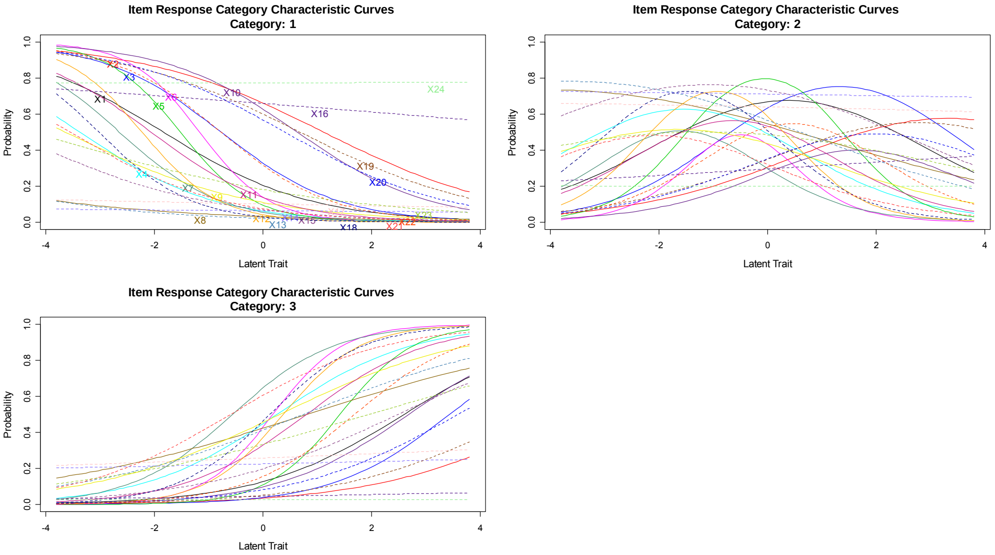
In the above graphs (Figure 5.5), the C construct is shown to be able to differentiate between individuals who are really high in the C construct.

The peak range of C construct is around 1~3 theta. Both old and new sets indicate similar results. The C construct also indicates a general improvement of item information (1~1.4). This result is considered to be an improvement.

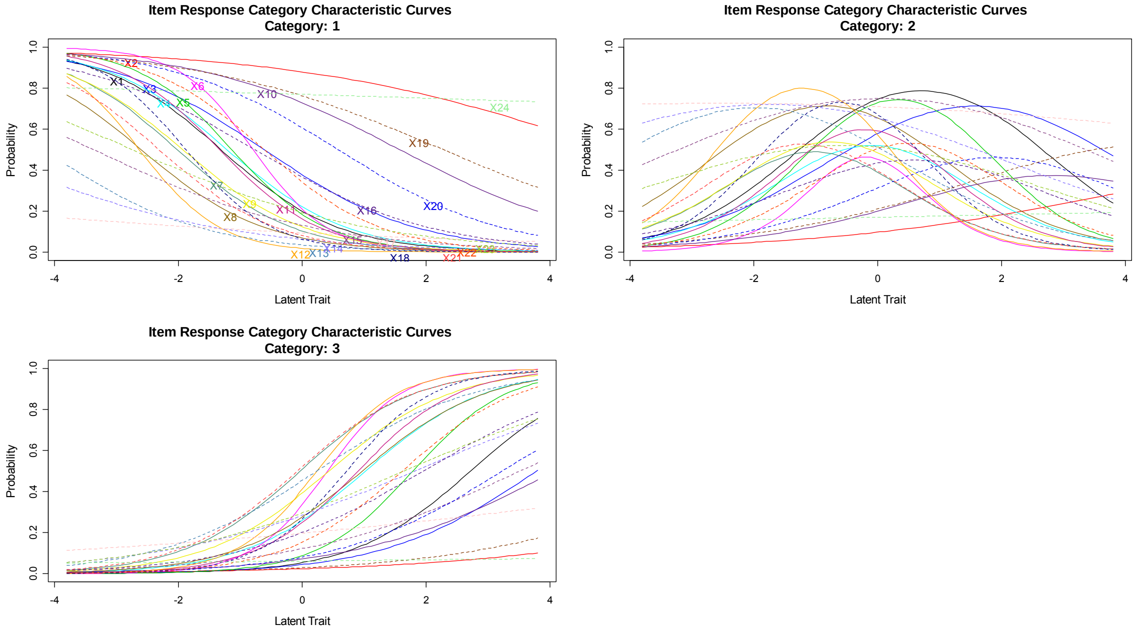
5.6.3.2 IRCCC-S analysis overall

The following sections (Figure 5.6~5.9) are the 'before' and 'after' analysis of Amend A. The main analysis that been conducted here is item response category characteristic curves summary (IRCCC-S). The results could be interpreted in detail for different DISC constructs in the following sections. The D constructs are generally functioning better except for a few items. The I construct, on the other hand, has 9 items that show improvements and 3 items that do not. It is also possible that the I construct could generate an inaccurate work mask. In the S construct, there are 9 items showing improvement, 5 items showing as decrement, and one item remaining poor. The C Construct IRCCC-s shows that there are 5 improved items, 4 items decrease, and 6 items remain poor. Overall, the DIS constructs show general improvement by the GRM model, while the C construct remains problematic.

Item Response Category Characteristic Curves (summary) IRCCC-S



Before: D Construct IRCCC-S of Old item (n=650)



After: D Construct IRCCC-S of New item (n=307)

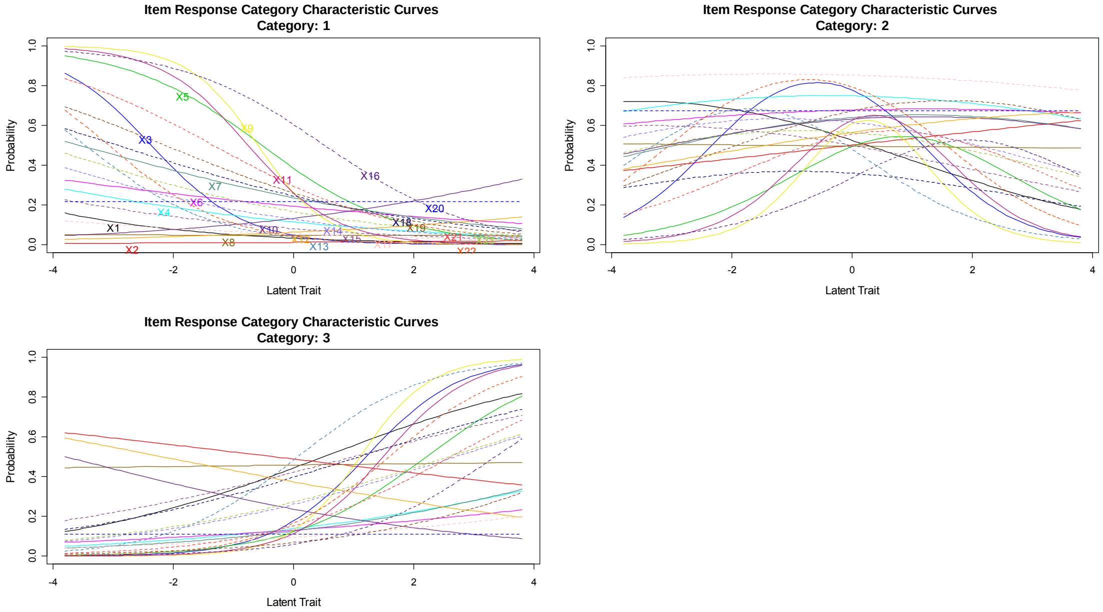
Figure 5.6 Comparing D construct in IRCCC-S

D Construct IRCCC-S

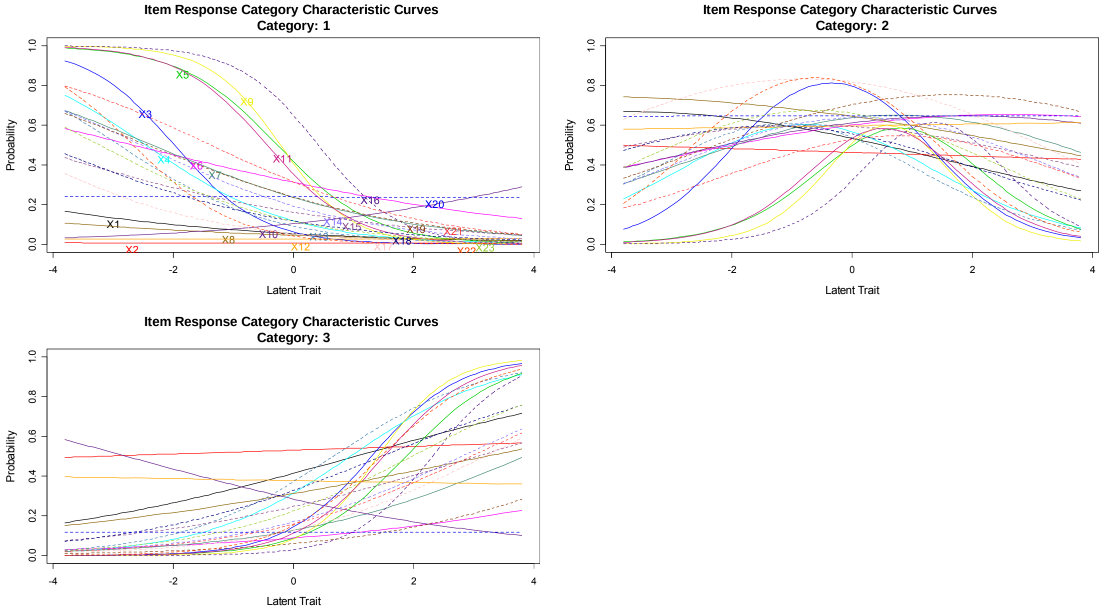
According to the above IRCCC-s for the D construct (Figure 5.6), items in both before and after *Amend A*, the D constructs can be interpreted as functional. The IRCCC-S shows that many D items function better in the 'Most' negative discrimination, this could lead to better performance in the work mask. The 'Least' curves show that item 24 still does not function properly, although many items (such as 16) start to function.

However, item 2 seems to be slightly unstable due to the increment of un-popularity (higher 'L' marking). The result suggests that although there is no change in the item, the response pattern changes. This result is possibly due to the historical effect (effect that happened due to natural development through time). The item 02, 'Stubborn', has become even more un-popular in the current population. This effect could stem from the popularity of contemporary Western business culture due to the influence of media and the educational systems.

In the Western business culture, the term ‘stubborn’ has a strong negative connotation. This view could have influenced the contemporary Chinese business culture, and led to negative responses on the PPA. Overall, the D construct has better discrimination with items such as 08, 09, 13, 14, 16, and 23 yielding a much better GRM curve after amendment. However, the amendment of A is not perfect as items 02, 17, and 24 still pose relative threats to the D construct. It is therefore suggested that CPPA undergo further amendment.



Before: I Construct IRCCC-S of Old item (n=650)



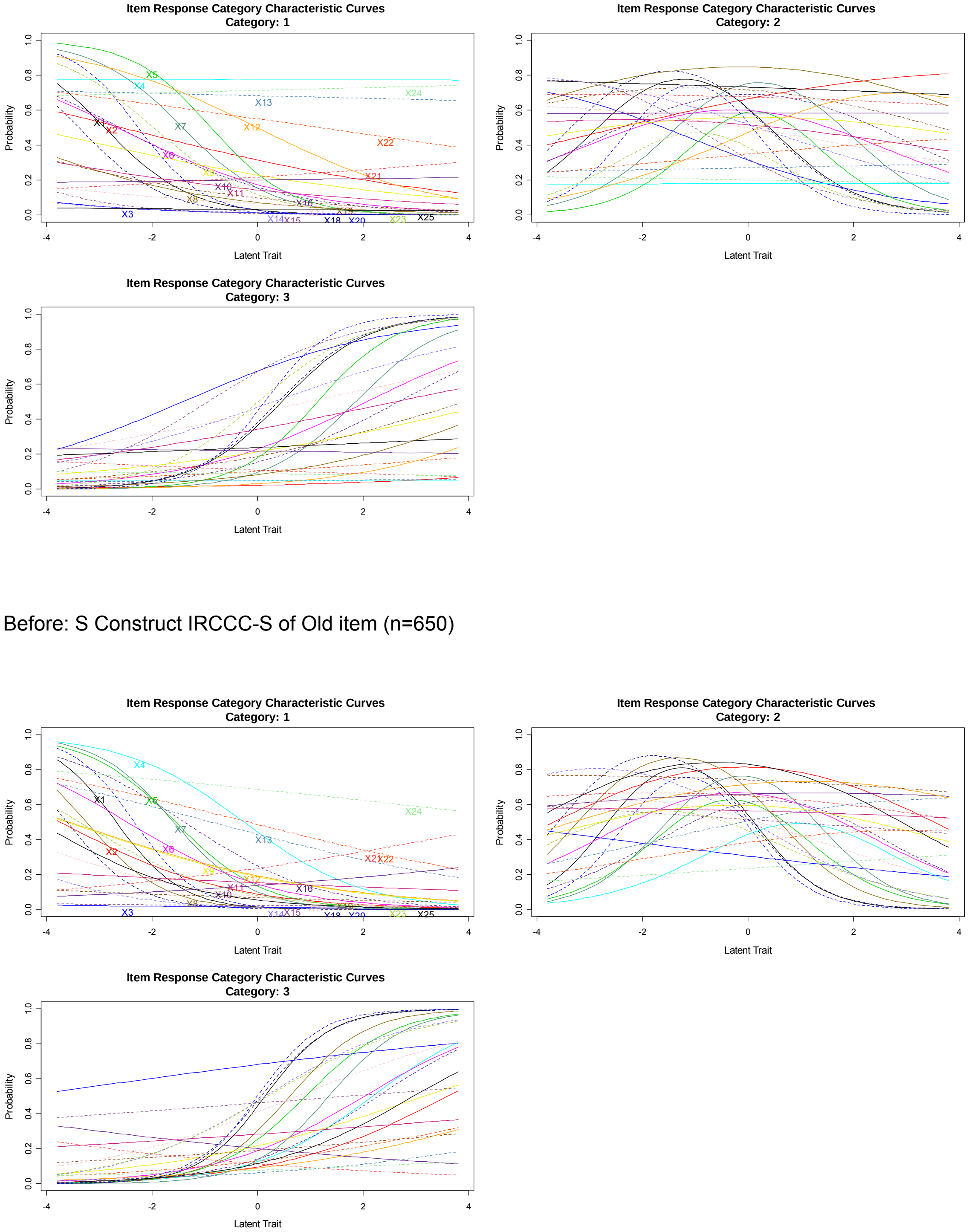
After: I Construct IRCCC-S of New item (n=307)

Figure 5.7 Comparing I construct in IRCCC-S

I Construct IRCCC-S

In the above graphs (Figure 5.7), compared to the D construct, the I construct is still relatively unstable. The *Amend A* seems to improve a few items' curve in the 'Most' graph. However, these results suggest that further amendments may be required. The original form has 4 items that do not function properly in the least graph, which decreases to 3 in the new form. Improvement of performance is shown in item 04, 06, 07, 08, 14, 16, 17, 18, and 19.

A decrease of performance is shown by item 02 and there is no improvement shown in the performance of items 12 and 21 (already poor items). Most items do show an improved performance, but the results suggest that the Work mask of I construct is relatively unstable. This could be due to the popularity of various 'I' items in the contemporary business environment, which could have led to low test/retest reliability in the I construct.



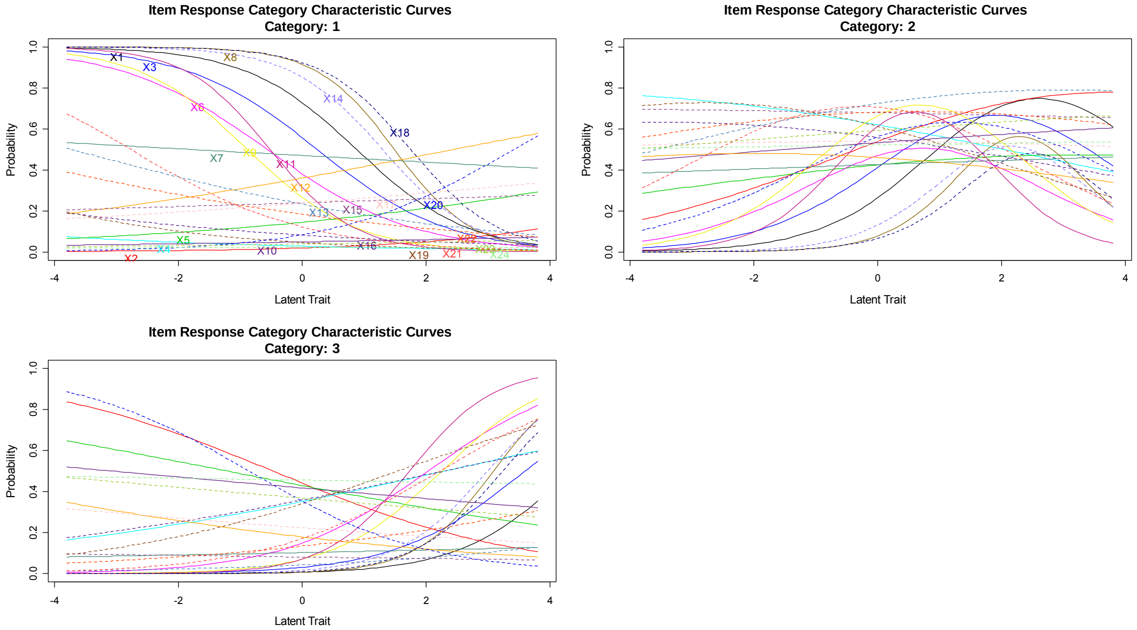
Before: S Construct IRCCC-S of Old item (n=650)

After: S Construct IRCCC-S of New item (n=307)

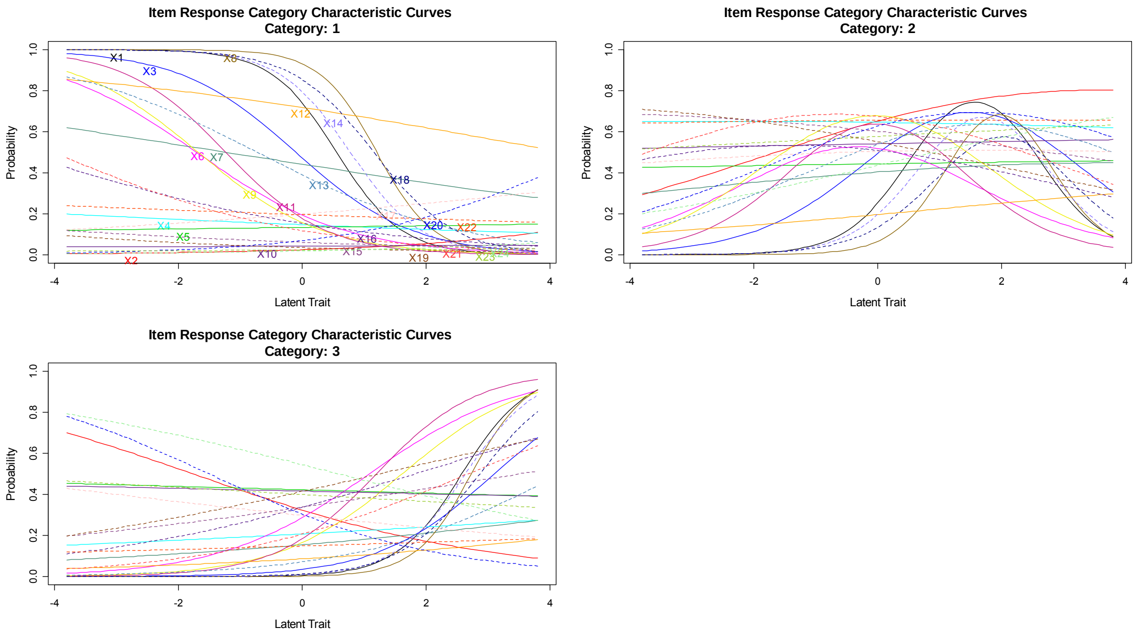
Figure 5.8 Comparing S construct in IRCCC-S

S Construct IRCCC-S

In the above S graphs (Figure 5.8), the old S construct is unstable in the 'Least' graph, or the pressure mask. The new S construct shows improvement in the pressure mask, but the work mask is relatively unstable in the new S construct. An improved performance is shown in items 02, 04, 08, 09, 13, 14, 17, 21, and 24. A decrease of performance is shown in item 03, 10, 11, 15, and 19, while there is no improvement in the performance of item 23 (already a poor item). After the *Amend A*, 9 items showed improvement, 5 items showed decreased performance, and one item remained poor. The result therefore suggested that another amendment should target items 03, 11, 15, 19, and 23. The result indicated that, under the current S construct system, the pressure mask and work mask were both unstable.



Before: C Construct IRCCC-S of Old item (n=650)



After: C Construct IRCCC-S of New item (n=307)

Figure 5.9 Comparing the C construct in IRCCC-S

C Construct IRCCC-S

According to above graphs (Figure 5.9), 'before' and 'after' amendments the C construct 'can be considered unstable. The results suggest that C contains a construct that no longer functions according to the original assumption. This further implies that the items should be classified into two or three constructs instead of the one C construct. Improvement of performance was shown in items 01, 07, 13, 15, and 16. A decrease in performance was shown in items 04, 19, 22, and 24, while there was no improvement of performance shown (already poor items) in item 02, 05, 10, 12, 17, and 23. There are 5 items showing improvement, 4 items showing decreases, and 6 items remained poor. The results suggest that item 04, 19, 22, 24, 02, 05, 10, 12, 17, and 23 should be targeted for the next amendment. The full PPA GRM IRT comparison of Amend A is in Appendix 4, on the CD.

5.7. Results summary

5.7.1 Relationship between measurements

In most cases the method for examining item discrimination GRM-IRCCC agreed with CPE-ICC's result. The contamination-checking methods (KTB and RICC) yielded similar results. The negative discrimination of all methods showed contamination, and were generally similar, suggesting s that the methods confirmed one another.

However, there were still incidences of disagreement. The following interpretation assumptions were made (see Table 5.34). In cases of disagreement, this study used GRM as the primary method. The other methods (CPE, RICC, KTB, and FC-MCQ) were used as supporting methods. However, when the majority disagreed with GRM, the study followed the supporting methods' interpretation.

Table 5.34 Summary table: Forced choice analysis methods comparisons

	Methods for examine discrimination		**Method for examine contamination**		**Method for examine item option spread frequencies**
	GRM	CPE	RICC	KTB	FCMCQ
Theoretical aspects					
Measurement Sensitivities					
Sensitivity to weak item	Yes	No	No	Yes	Yes
Sensitivity to strong item	Yes	Yes	Yes	Yes	Yes
Item Swing sensitivity	Low	High	High	Low	No
item contamination					
Contamination detail	No	No	Yes	Yes	Yes
Item construct nature	partially	No	Yes	Yes	No
Construct strength	No	No	Partially	Yes, full detail	Yes
Construct dynamics	partially	No	Yes	Yes	Yes, full detail
Item Response Theory					
Definition of ability	(Defined by 3PL, theta)				
Suitability affective psychometric	Yes	Yes	Yes	Yes	No
Custom made definition	No	Yes	Yes	Yes	No
Suitable for explorative study	Yes	Yes	Yes	Yes	Yes
Item discrimination	(Defined by 3PL, A parameter)				
Item parameter	Yes	Yes	No	No	No
Item difficulty	(Defined by 3PL, B parameter)				
Item parameter	Yes	Yes	No	No	No
Social Desirable Response	(Defined by 3PL, C parameter)				
Item parameter	Yes	No	No	No	No
SDR in frequency	No	No	No	No	Yes
Level of measurement (algorithm assumption)	Ordinal	Nominal	No assumption	Ordinal	Nominal
Result interpretation					

Overall method Reliability	High	Median	Low	High	High
Interpretation Difficulty	Easy	Easy	Easy	Easy	Difficult
Practical aspects					
Time Cost (with current methods)	20 minutes	1 day	1~2 weeks	30 minutes	5 minutes
Software capable for performing analysis	R- ltm package (Royalty free), with SPSS data refinement	MS excel, with SPSS data refinement	SPSS calculation, Excel graphing, and Human interpretation	SPSS calculation, with Excel interpretation sequence	SPSS, additional syntax
Possibility of human error	Low	Median	High	Low	Low
Method reliability	High	Median	Low	High	High
Total Cost	Low	Median	High	Low	Low

Note: GRM=General Response model; CPE=Correlational Parameter Estimation; RICC=Raw Item Characteristic Curve; KTB=Kendal's Tau B ordinal correlation modification; FCMCQ=Forced choice to Multiple Choice Question flattening model.

This study suggests that Samejima's (1969, 1972, 1973a, 1973b, 1974, 1977a, 1977b, 1996, 1999) Graded (General) response model is the most accurate model for PPA research, but the effectiveness of other supporting methods is not discounted. The study further suggests that all these methods could be employed for a comprehensive PPA analysis. The following protocol is recommended.

5.7.2 Protocol: Check list for forced choice analysis

Textual: Check for translation error. The textual error is the primary mistake of FC analysis. When the translation does not agree with the original item it is very likely to be non-functional. Following the original text is important, but following the original construct is even more crucial. This step could also be the last.

Functional: Check for item discrimination index. The GRM-IRCCC or CPE-ICC should be used when items demonstrate anything other than positive discrimination. This result suggests that items contain other constructs or contamination.

Dynamical: Check for contamination and item set system dynamic. The KTB or RICC should be used and the FC-MCQ generated to support the result. The contamination originated from poorly selected items or uncontrolled social desirable responses. The item quality can be examined via GRM-IRCCC, CPE-ICC, KTB, or RICC. The socially desirable response can be examined by GRM-IRCCC and FC-MCQ.

Constructual: Check for construct functionality. Researchers should regroup items via textual meaning and run GRM-IRCCC-S to reconfirm constructs. This would help researchers to examine the real item functionality within the target culture.

5.7.3 Decision and interpretation of results

The current study suggests that results yielded from above the protocol can be interpreted in the following system (see Tables 5.34~5.37) and that alterations or amendments could be made where applicable. The current protocol places the textual interpretation in the first level of exemption.

However, it is also suggested that all levels can be revisited when the need arises. The support of a statistical result would increase the success of textual interpretation. The early interpretation/alteration of the textual content could lead to subjective judgments. It is not uncommon that mental health specialists would consider their opinions as a representative voice of the target population. However, the current results suggest that mental health specialists have a higher degree of education, training, and exposure to psychometrics, and for this reason that they may not view the items in the same way as the lay public.

It was therefore suggested that researchers should let the statistics, as the public's real voice, speak for itself.

Table 5.35 Interpretation Decision system in current study

Support Methods	Primary method	GRM		
	Outcome	**Positive**	**Flat**	**Negative**
CPE	Positive	G	SC	SC
	Flat	C	N	N
	Negative	C	N	N
RICC	No contamination	I	SC	SC
	With contamination	C	N	N
KTB	No contamination	I	SC	SC
	With contamination	C	N	N
FC MCQ	Even spread	I	LC	LC
	Extreme Spread	SO	SO	SO

Note: Acronyms are interpreted in the following legend table

Table 5.36 Legend to interpret Table 5.35

Judgment legend	Detail
C	**Contamination**: Contamination exists. However items are not entirely non-functional
SC	**Sub Construct**: Item is functioning. However, sub constructs exist. Item should be regrouped. Sub-structure should be investigated. It can also be due to incoherence of the construct.
N	**Not functional**: Affirmative contamination. Items are not functional. Amendment should be followed
G	**Good item**: Item is functional
I	**Ideal**: Item is ideal
LC	**Low Contamination**: There is low contamination within tetrads. However, item might not be functioning
SO	**Socially Desirable item:** Highly socially (un) desirable item exist within the item set. Alteration is necessary.

Table 5.37 Four types of outcomes

Forced choice analysis	**Item contamination (Can be identified by RICC or MTB)**		
Item discrimination (Can be identified by GRM-IRCCC or CPE-ICC)		**Yes** (item showing mix constructs)	**No** (item showing one construct only)
	Yes	**Type A:** Item is functional but contamination exists.	**Type C**: Item is functional in its optimum.
	No	**Type B**: Item is poorly constructed.	**Type D**: Item is poorly constructed. Original construct has more than one sub construct.

5.8. Chapter summary and discussion

The current research yielded results from the old form (n=650) via classical test theory (CTT) as well as Item Response Theory (IRT). The reliability result (ranging 0.227~0.725) suggests that the old form is relatively unstable, especially SC construct. The construct inter-correlation suggested that the SA model still functions similar to international studies.

This research result suggests that both CTT and IRT methods can be used for PPA analysis and therefore can be explored in the Chinese population. The traditional item analysis (item discrimination and difficulty index) revealed that D I construct was functional but slightly difficult for the Chinese population. The SC construct also had higher numbers of items that were considered as 'too difficult to answer' for the current population. The IRT RICC result showed various items as problematic and suggested *Amend A* (15 items for amendment, 4_02, 2_04, 4_04, 3_06, 3_08, 3_10, 3_12, 1_13, 3_13, 3_14, 3_15, 4_16, 1_18, 1_24, & 4_24). After the *Amend A* been conducted CTT and IRT methods were again employed to investigate the difference.

The comparison of results yielded by item discrimination and difficulty suggested that the new form (n=307) had improved 12 items out of 15 (80%) amended items. The experimental Kendall's Tau B modification analysis was also used to compare the difference, and the result suggested that 73.3% (11 items) showed improvement, while 26.6 (4 items) were still problematic.

The IRT Graded response model had been used to compare results. The GRM TIF analysis showed a general improvement 1~1.5 on the test information. This inferred that items yielded more information (higher discrimination) in the measurement process. In IRCCC-S analysis, the DIS construct showed a general improvement by *Amend A*, while C construct remained problematic.

The current study also suggested a standard protocol that could be used in future PPA research. The protocol used the GRM model as the backbone supported with CPE, RICC, KTB, and FC-MCQ to investigate various dynamical detail of PPA. This study compared the theoretical and practical aspects of all the above statistical models. The research protocol suggests four levels of investigation, which are textual, functional, dynamical, and constructual. The four level analysis should not be considered as hierarchical analyses, but rather as different aspects. There is no absolute order that one should follow. They are four different types of analysis that would support one another and provide greater understanding. The research protocol suggested that all four levels be investigated for a better illustration of PPA functionality in cross-cultural settings. This protocol would generate the result and support with interpretation methods for PPA item evaluation and improvement.

However, current research was not without its own inherent issues. The 2nd sample group (n=307) is be relatively small (n<500) for IRT analysis (Jooste, 2003), which rendered the IRT analysis somewhat unstable. In this dissertation this part of the study was backed up by with KTB correlation analysis (non-parametric ordinal correlation) to avoid the risk of misinterpretation. However, the small sample could still stand as a potential threat to the research results. This dissertation was an exploratory study that was constructed via various small studies and statistical methods.

Although the study was an exploratory one, the use of too many small studies could lead to a superficial analysis, lacking sufficient depth on the different sections. It is therefore suggested that the results should be viewed bearing the above restrictions in mind.

CHAPTER 6. DISCUSSION

6.1. Introduction

The experience of this current exploratory study was of a very practical nature. This method exhausted all the available routes at once. It was relatively cost effective to gain maximum information from limited research resources and assisted practical decision-making. However, the downside of this study was that this type of research could yield too much information, and lead to a superficial understanding of the results. As Chapter 3 stated, the research focus of this study is. ***'Can item response theory (IRT) be used in forced-choice psychometric (FC) adaptation?'*** However, in terms of adaptation, this study still presents many limitations. This chapter focuses on various aspects of this explorative study with the various limitations being introduced via the ethical framework, followed by suggestions of various areas for future research.

6.2. Limitation of current psychometric adaptation methods

The limitations with regard to the adaptation of this study could be roughly divided into three stages of psychometric assessment, (1) before assessment, (2) during assessment, and (3) post assessment. Therefore this study discusses the various problematic aspects according to these three stages.

First, 'before assessment,' the most important issue with psychometric research is with regard to the issue of fair practice. The instrument needs to suggest that it has a good theoretical grounding, reliability, and validity (Eckert, Hintze, & Shapiro, 1999). *Second*, 'during assessment,' this aspect would mainly be the standardised administration procedure or fair administration practice. *Third*, 'post assessment,' this would be the social consequences of assessment, how the test is applied and interpreted; and the social or psychological impact on the test taker (Eckert et al., 1999).

This research explored the possible statistical methods that can assist with the adaptation of forced-choice (FC) psychometric techniques. However, the test was conducted in a Chinese setting making valid-translation the first problematic issue. In the field of psychometric translation, the Brislin's (1970, 1980) back-translation model has been widely used (Eremenco, Cella, & Arnold, 2005; Lee, Li, Arai, & Puntillo, 2009; Sireci et al., 2006; Wang & Lee, 2006). However, the current study did not use this model only. This study also researched and developed alternative methods to examine construct equivalence aspects.

The study focused on refining and creating methods for the practical aspects of the first stage. This method/protocol was designed to increase the validity of translation, and also to boost the efficiency of adapting an old test to an international setting. To ensure the functionality of the instrument, this chapter follows the above three aspects as a framework for discussing the possible issues of current research.

6.3. First stage: Before assessment

6.3.1 Translation equivalence

The current study compares the two Chinese translations of the PPA. However, using immature translation forms would encounter several innate problems. Past researchers suggested that the goal of a congruent cross-cultural translation process is to achieve two levels of equivalence: *textual* and *statistical*: *The textual level*: semantic and functional equivalence. *The statistical level*: differential item functioning (DIF) (Eremenco et al., 2005; Flaherty et al., 1988; Sireci, et al., 2006; Yasuda, Lei, & Suen, 2007). The DIF methods can vary from factor analysis (Yasuda et al., 2007), logistic regression (Sireci et al., 2006), Mantel Haenszel (Linacre, 2009; Linacre & Wright, 1987), or item response theory models (Nathanson & Paulhus, 2004; Puhan & Gierl, 2006).

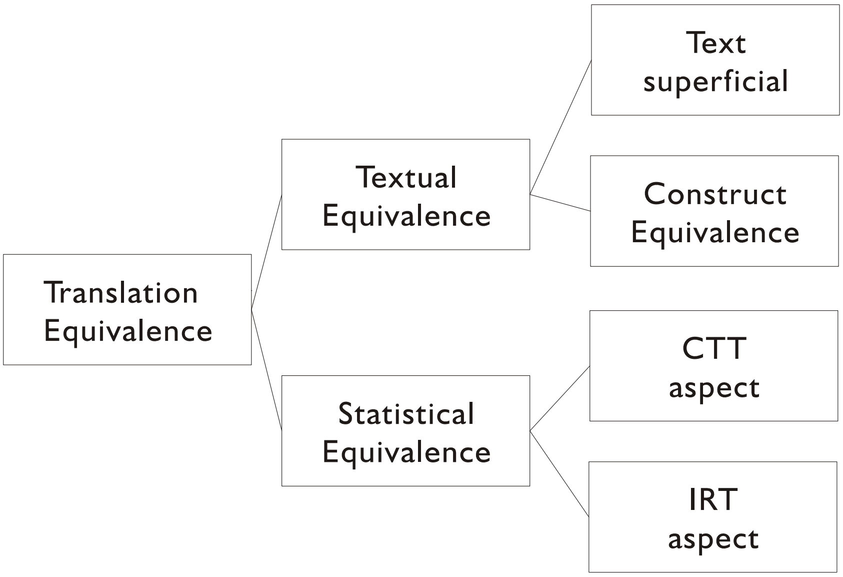


Table 6.1 Level of equivalence in translation

The current research's textual equivalence was only based on two bilingual researchers' opinions for alteration (Chapter 5, section 5.5). The amendment (Amend A) suggestion was decided via consent from both researchers and the statistical result. The final decision was only made when all the statistical analyses showed congruent results.

6.3.2 Translation: Textual equivalence

The current research textual equivalence was different from the popular back-translation adaptation model proposed by Brislin (1970). This study differed from the common practice in the way that it operated mostly in simplified Chinese, without English back-translation. The reason behind this was that Brislin's (1970) model cannot ensure that the functionality of the item still remains the same.

In Brislin's (1970) assumption, a scale remains functional if the translation is equal to the original setting. This assumption ignored the possibility that even though the text was equal in meaning, it would not function in the same way in different times and cultures. For example, the importance of being ‘assertive’ is very different in western and eastern cultures. ‘Assertive’ is generally considered a positive term in western industrial culture, but as a negative word in eastern culture. Using a bilingual specialist to translate would be more likely to retain the functionality of the scale. However, this top-down procedure would also lead to the dilemma of selecting words not commonly used by the target sample. Although the decision was only made when the bottom-up statistic result also concurred, the true voice from the general public does not come through clearly. This study therefore suggests a focus group on ‘item definition’ should be conducted for future research.

6.3.2.1 Textual equivalence: Construct equivalence in forced choice

Early researchers raised the question of the cultural fairness of psychometrics (Torrance, 1981, 1993). Researchers frequently suggested that all psychometric testing is biased in favour of individuals from the dominant culture who designed the test (Gipps, 1999; Gipps & Murphy, 1994; Wigdor & Garner, 1982). This is the first theoretical, as well as an ethical, issue that a researcher needs to consider when using psychometrics in cross-cultural research. Even when the equivalence among forms can be established, it is still difficult to suggest the existence of constructs in alternative cultures (the true cultural difference could exist) (Kim & Han, 2004; Wang & Lee, 2006).

In a psychometric study, a psychological construct is ‘constructed’ via all items’ textual semantic interpretation and response patterns. If any fundamental elements are different, the equivalence of construct to the original culture cannot be established. Also, even without translation, it is difficult to assume that a construct functions exactly the same (i.e. equivalently in all aspects) in another culture. Furthermore, when using the FC tool, an additional facet should be noticed. The ‘popularity’ (social preference) of each item would also be different. The alteration of this characteristic could severely handicap the function of psychometric testing, especially for FC scale, because FC in essence operates on preference.

Contemporary psychometric translation has progressed from the ‘meaning essential’ (Catford, 1965; Nida, 1964) to ‘multiple levels of equivalence’ (Flaherty et al., 1988; Lee et al., 2009; Wang & Lee, 2006). However, to ensure psychometric equivalence and functionality, more aspects need to be explored. In the psychometric registration guideline of the Health Professional Council South Africa (HPCSA, 2006a, 2006c, 2008, 2009), the localised construct model is considered as important evidence that psychometrics is still functioning within South African culture. The statistical analysis of a construct model normally works through a confirmatory factor analysis (CFA). However, such methodology (the original CFA form) is suggested as being statistically inapplicable to FC psychometrics (Bess, Harvey & Swartz, 2003). Therefore, this research cannot provide the evidence for construct equivalence in the popular CFA form. This would be an inherent limitation with regard to forced-choice research.

To overcome the above limitation, this research suggests several methods to examine the construct coherence, such as Kendall's Tau B (KTB) and Grade Response Model Item Response Category Characteristic Curve Summary (GRM-IRCCC-S). The result of the above methods in the old Chinese translation showed that the Item Expressed Construct (IEC) was very often different from the item Pre-Assigned Construct (PAC). Furthermore, the GRM-IRCCC-S revealed in general that the D, I, and S constructs functioned similarly to the original form. However, the C construct seems to be problematic.

This result could lead to several interpretations. The non-convergent behaviour of C items could originate from item or construct problems, or both. The poorly translated items lead to non-convergent constructs. The construct error suggested that some of the PPA construct could be culturally specific, instead of universal.

The result suggested that previously assumed universal PPA constructs were subsequently found to be culturally specific. However, in the psychometrics world, it is not uncommon in cross-cultural translation research. The well-known example was the big-five theory. Research established that there is a true difference among Asian and European in the big five measurements, and also found the existence of a LOF (loss of face) construct among Asian, but which is not strongly represented in Europeans (Eap et al., 2008). Szirmák and Raad (1994) also found no openness dimension in a Hungarian sample.

However, researchers have suggested that such differences could originate from the ways of expression (Brown & Harvey, 2003; Eap et al., 2008; Middleton & Jones, 2000) or response styles (Billiet & Davidov, 2008; Cheung & Rensvold, 2000; Meisenberg & Williams, 2008; Soto, John, Gosling, & Potter, 2008).

In this research it was reported that 15 items from the old Chinese form do not function properly (IEC=/=PAC). However, it was difficult to attribute the cause of this of IEC and PAC to poor item structure only. The true cultural difference can originate from the item level. An item, despite equivalence of textual translation, could still represent, and be linked to, completely different strings of historical and cultural-system semantics. items with the same translational meaning in two different language systems, could therefore represent different constructs (Lee et al., 2009).

The possibility of reflecting the true difference among cultures was not fully explored in the current study. For an item to function, the item needs to be select correctly, and the target culture needs to able to responds correctly. This study was designed for measuring the functionality mostly at the item level, and presumed that differences can be dealt with at the item level. However, the 'cultural' part of the study is lacking in current study. It is therefore suggested that a future study could focus on this aspect.

6.3.3 Translation. Statistical equivalence

Although it has been suggested that statistical methods for FC scales are achievable (Harvey & Thomas, 1996; McCloy et al., 2005; Wagner & Harvey, 2003), the practical methodology that would guarantee success for FC scale has not yet been fully set out in the available literature. Therefore, this research was aimed at exploring a suitable algorithmic protocol as the statistical-level equivalence benchmarking tool (as the differential item functioning DIF tool). However, these methods have not yet been tested by other researchers and so the results of the study may need to be validated in further research.

Several methods (KTB/GRM/IRCCC/ICC/FCMCQ) were used to study the differences between before and after amendment of problems.

The difference of IEC between before and after forms could, in essence, be as a result of the instability of the statistical method. Other factors could include the methods' sensitivity to sample size (sample sizes smaller than 500), true sample differences, or historical effects due to the time difference. The interpretation method is also highly experimental and intuitive; therefore it comes with no guarantee of correctness. Furthermore, there were no previous studies suggesting statistical equivalence of the two forms to the original English form. It is therefore suggested that this study only be regarded as an exploratory or pilot study.

6.3.3.1 Construct in-depth: PPA model eccentricity - CTT aspect

The Thomas PPA construct model does not come without any pre-existing problems. The most problematic issue concerns the relationship between the S and C constructs (Irvine, 2003). Throughout most of the Thomas' internal studies, the SC constructs in international studies remain at a moderate (.1~.4) level of correlation with one another (Irvine, 2003).

This also appears to be the case in the current study. Such evidence has been interpreted as 'the universal stability of construct structure' across different cultures. However, this interpretation is doubtful due to the forced choice nature of the PPA. The correlational result could be 'artificially created' from the pre-existing forced choice options. Further, if such correlation was true, it could also reflect the fact that the SC constructs are not 'constructually clear' during the theoretical formulation. The SC constructs, instead of two constructs, are more likely to be two sub-constructs under one 'passive' mother construct. However, because the correlational result for forced choice cannot be truly accounted for due to the artificial correlational relationship between constructs (as stated in Chapter 4, section 4.5) such a postulate is only speculation.

Furthermore, the first part of this research study (see Table 5.2~5.5) demonstrated that the PPA working mask does not have the SC correlation (*r*=-.041, *p*=.295). However, the overall 'self mask' shares the correlation (*r*=.121, *p<.01*) that could derive from the pressure mask (*r*=.141, *p<.01*).

When studying the construct consistency (same construct in different masks) throughout three different masks, it was also interesting to note that most of the DIS constructs share moderate correlations with the other DIS constructs among the work and pressure masks (*r*=.459~.577). The C construct, however, only shares a minor relationship (*r*=.118) throughout all three masks. Based on the above information, it is suggested that the C constructs within the original Chinese forms are relatively unstable.

6.3.4 PPA model eccentricity: GRM-IRT aspect

This research was conducted under the assumption that the GRM-IRT model was applicable to the FC-psychometric adaptation process. This research concluded that the current method had improved on several aspects of the old form based on various assumptions set out in Chapter 3.

The major assumption is that these improvements are only attributable to the amendment. It is also possible that these differences originated from the current rapid and large-scale acculturation phenomenon of Chinese culture with global, Western industrial values. Firstly, the two sample groups were collected separately in an interval of 4~1 year (s) (first. 01/04/04~01/11/07, second. 03/11/07~20/05/08), the amount of acculturation/globalisation change during an economic/information boom-time in China could be larger than expected, and this aspect should therefore be given serious consideration.

The improvement might also have originated from the cultural-shift in the macro environment, instead of the proposed method in the current study. Also, the difference between the English and Chinese forms could indicate the true difference between cultures, instead of translational non-equivalence (as stated in Chapter 3 Area C, E.1, and E.2). This study therefore suggests that researchers could target a further refinement of the above issues.

6.3.5 Research limitations

Firstly, the current study presented a fundamental threat; **too many methods of analysis were used.** This research involved four CTT methods and five IRT methods. The CTT methods were. Cronbach's Alpha, discrimination index (D), difficulty index (P), and correlation (r). The IRT methods were. RICC, KTB, GRM, FCMCQ, and CPE. In the five IRT methods, four (RICC, KTB, FCMCQ, & CPE) were experimental methods.

Secondly, the research structure was too complex, comprising two parts and two families of methods and it is also potentially jeopardised by **too many assumptions**. Chapters 3 and 4 presented various postulates and assumptions, the validity of which have not yet been investigated in any depth. In the results, Chapter 5, it was merely suggested they have some validity because of their similarity to results obtained with other established statistical methods. However, validity of these methods should be investigated in more depth.

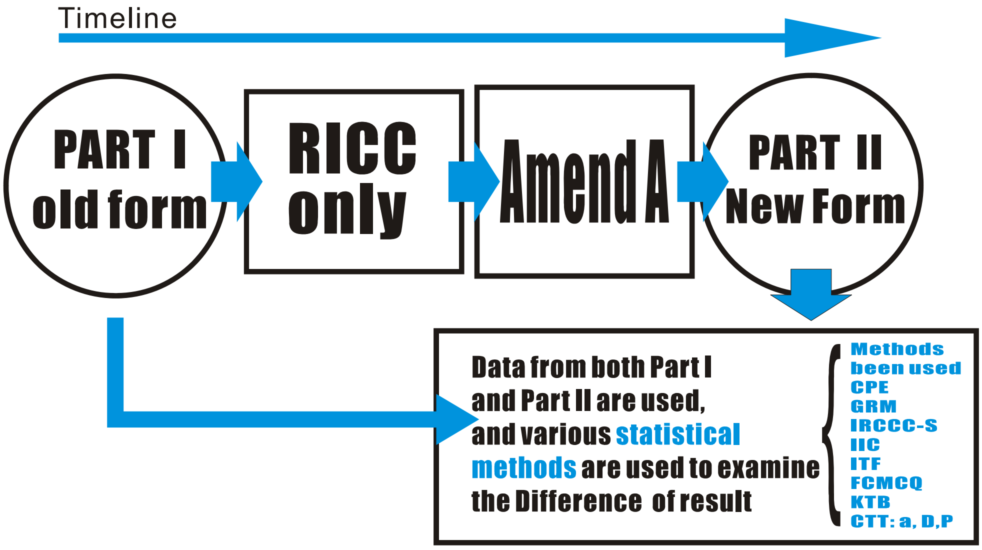


Figure 6.1 The research process

Thirdly, standardisation of the research is negatively affected by **changing the research methods.** Research methods such as RICC methods are different throughout Parts I and II. The original RICC method was only conducted in Part I and replaced by the RICC-CPE method in Part II (see Figure 6.1). Although data of both Parts I and II went through RICC-CPE analysis for a comparison result, the original RICC was only conducted in Part I.

Despite the fact that methods such as KTB, CPE, and FCMCQ were constructed to re-examine the results, they were not used in Part I. The *Amend A* was purely a product of the original RICC and item analysis. The original RICC method was removed from the research in Part II mainly due to the high time costs of manual interpretation. In Part II, the original RICC form was replaced by the relatively efficient RICC-CPE form, and supported with KTB and FCMCQ. Although the result suggested a general improvement, it could be a reflection of the methods and not a true improvement.

6.3.5.1 Result interpretation

This study was aimed at exploring various areas of the topic, and therefore contained too many analyses. However, the true practical strength and weakness, and actual applicability cannot be known before the research. Therefore this research selected various methods, and ended up with the knowledge that some of the methods are very impractical and should be excluded in the first place.

The impractical methods as RICC, CPE, and alternative forms of correlational methods (Pearson's r) that used to re-confirm the result of Kendall's Tau B (KTB). To shorten the research process, these methods should be excluded. This is probably the most valuable knowledge gained from this explorative study. The above methods were also created to aid the understanding and confirmation of other methods, such as RICC, CPE, and KTB. However, the inclusion of multiple results could lead to superficial explanation and discussion. On the other hand, there were still areas not fully explored by this study, such as the full mathematical logic underlying IRT, various IRT assumptions such as unidimensionality, ability estimation, 2PL, 1PL, and Rasch's models. The true mechanism of IRT in FC testing also still needs to be explored.

6.3.5.2 Sampling

Unclear sampling frame: APA psychometric standard suggested that selected samples should be sufficiently representative of the target population (as cited in CSPB, 2003). In this study a sample of n=650 was collected in the first part of research and n=307 in the second part of the research around the Beijing area. Various details of the sampling frame remain unclear, and as a result this research has only limited biographical information about the respondents. The age group and original province were not provided in these data (see Chapter 4), and these could both be important indicators of the type of Chinese dialect used by the respondents. The age group could also indicate the level of familiarity with the official Chinese dialect (Beijing), which differed at times due to changes in education styles.

The Chinese language system contains many different dialects (Hashimoto, 1973) that could represent fundamentally different cultural views (DeFrancis, 1990). Individuals working in Beijing would all use and understand the official language (Beijing dialects). In China's official educational system, individuals from all provinces and major cities (except Hong Kong) would be taught to read and write in Beijing dialects during elementary and high school education (CERN, 1956). However this could still be a potential threat to the validity of the study. Therefore, the influence of these variables on current findings cannot be excluded. Readers should examine the findings cautiously bearing the above warning in mind. The general profile of the sample probably represents white-collar inhabitants near the Beijing area more than the general Chinese population.

6.4. Second stage: During assessment

6.4.1.1 Administration method

The current study's integrity was also endangered by **non-standardised forms and administration methods.** According to APA psychometric standard, the administration procedure and test user's qualification should be standardised (as cited in Kaiser & Smith, 2001).

Although all administrators received a Thomas' administration training, this does not completely guarantee the quality of the test administration.

All Thomas psychometric administrators were given oral/telephonic/written briefs regarding the objectives and time frame of the tests. They were also given one item as a practice to check their understanding, prior to the actual test. However, there were still minor differences. Although the A25 and the C7 form series are textually equivalent, the data was originated from different types of assessments. Some forms were administered for recruitment and some for development; some were completed online and others via fax. Some tests were conducted by external personnel (although all had been trained in the Thomas system, not all were qualified mental health professionals) under different conditions and in different environments. It cannot be excluded that various types of administration methods and purposes would influence a subject's interpretation of the terminology and how to apply them.

6.5. Third stage: Post assessment

6.5.1 Feedback

In a normal Thomas process, a textual report would be generated through the Thomas software system as the feedback report. Textual interpretation would be given via a mental health specialist (clinical, industrial, counselling, or educational psychologist and psychometrcist). As an extra incentive, candidates participating in this study would receive a textual report emphasising that this was for research purposes only. No verbal feedback would be given. Although candidates had received the report clearly stating that it was for research purpose only, Thomas International was unable to limit usage of the report in China.

The actual application of the textual report is entirely dependent upon the integrity of the test administrator. This poses the potential threat of using an un-validated test report for industrial purpose, such as candidate selection. The study therefore recommends that the test reports should not be generated for the future research.

Further, according to the current psychometric quality of the CPPA, it is not suitable for practical purpose, and application of the CPPA should be limited to developmental and research purposes via qualified mental health professionals only.

6.6. Future research

Based on this study, it is recommended that for future PPA-IRT research, the sample size should be at least 500 or more. An English dominated sample should be gathered to ensure the stability of the result. This study also suggests that GRM, KTB, and FCMCQ are the most efficient methods for conducting future studies. The mathematical details and logic of the GRM should be further explored, and the true level of measurement of FC should also be examined to ensure that correct statistics are used. Other models and various assumptions should also be explored, such as Rasch, 1PL, and 2PL models. In addition, the various assumptions such as unidimensionality, ability estimation, and true implication of MMLE in FC should be investigated. Furthermore, the current research was inspired by the question of universal applicability of psychometric across time, culture, and language, and how psychometric methods could be developed to counter various SDRs (social desirable responses). The true ability of FC to deal with the above two major issues was not answered in the current study. It is therefore suggested that experimental study should be conducted to further investigate the above issues.

6.7. Concluding remarks

Much of this study focused on the ‘before’ and ‘during’ assessment areas and was explorative in nature yielding large amounts of result data. The product of this study was a research protocol indicating what would be efficient and beneficial for future research. However, this study was also threatened by the superficially large amount of assumptions rooted in the very nature of exploratory research. The value of the current research still needs to be reconfirmed by future research. It is hoped that the current study will provide a starting point from which future studies might continue such further expoloration.

REFERENCES

Ahmanan, J.S. & Glock, M.D. (1981). *Evaluating Student Progress. Principles of tests and measurement* (6th ed.). Boston. Allyn & Bacon.

Aiken, L.R. (1991). *Psychological testing and assessment* (7th ed.). Boston. Allyn and Bacon.

Aiken, L.R. (2000). *Psychological testing and assessment* (10th ed.). Boston. Allyn and Bacon.

Aiken, L.R. (2003). *Psychological testing and assessment* (11th ed.). Boston. Allyn and Bacon.

Alagumalai, S. & Curtis, D. (2005). Classical Test Theory *Applied Rash Measurement. A Book of Exemplars* (Vol. 4). Netherlands. Springer.

Allen, M.J. & Yen, W.M. (1979). *Introduction to Measurement Theory*. Belmont. Wadsworth. Inc.

Ault, J.T. & Barney, S.T. (2007). Construct Validity and Reliability of Hartman's Color Code Personality Profile. *International Journal of Selection and Assessment, 15* (1), 72-81.

Baier, D.E. (1951). Reply to Travers' "A critical review of the validity and rationale of the forced-choice technique.". *Psychological Bulletin, 48* (5), 421-434.

Baker, F.B. (2001). *The Basic of Item Response Theory* (2nd ed.). USA. ERIC Clearinghouse on Assessment and Evaluation.

Baker, F.B. (2004). *Item response theory: Parameter Estimation Techniques* (2nd ed.). New York. Marcel Dekker.

Baker, F.B. (Ed.). (1992). *Item Response Theory: Parameter Estimation Techniques* New York. Marcel Dekker, Inc.

Barclay, J.R. (1991). Psychological assessment: A theory and systems approach. Malabar. R.E. Krieger Pub. Co.

Baron, H. (1996). Strengths and Limitations of Ipsative Measurement. *Journal of Occupational and Organizational Psychology, 69*, 49-56.

Barrick, M.R. & Mount, M.K. (1991). The Big Five personality dimensions and job performance. A meta-analysis. *Personnel Psychology, 44* (1), 1-26.

Bartram, D. (2007). Increasing Validity with Forced-Choice Criterion Measurement Formats. *International Journal of Selection and Assessment, 15* (3).

Bass, B.M. (1957). Faking by sales applicants of a forced choice personality inventory. *Journal of Applied Psychology, 41* (6), 403-404.

Bendig, A.W. (1956). The development of a short form of the Manifest Anxiety Scale. *Journal of Consulting Psychology, 20* (5), 384.

Berkshire, J.R. & Highland, R.W. (1953). Forced-choice performance rating--a methodological study. *Personnel Psycholog., 6*(3), 355-378.

Bernreuter, R.G. (1933). Validity of the personality inventory. *Personality Journal, 11*, 383-386.

Berry, J.W. (1976). *Human ecology and cognitive style*. New York. Sage.

Berry, J.W. (1980). Ecological analyses for cross-cultural psychology. In N. Warren (Ed.), *Studies in cross-cultural psychology* (Vol. 2). London. Academic.

Berry, J.W. & Associates. (1986). On the edge of the forest: cultural adaptation and cognitive development in Central Africa. Lisse. Swets & Zeitlinger.

Berry, J.W. & Irvine, S.H. (1986). Bricolage: savages do it daily. In N. Warren (Ed.), *Studies in cross-cultural psychology* (Vol. 2). London. Academic.

Bess, T.L., Harvey, R.J. & Swartz, D. (2003). *Hierarchical Confirmatory Factor Analysis of the Myers-Briggs Type Indicator*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology.

Billiet, J.B. & Davidov, E. (2008). Testing the Stability of an Acquiescence Style Factor Behind Two Interrelated Substantive Variables in a Panel Design. *Sociological Methods & Research, 36* (4), 542-562.

Birnbaum, A. (1957). *Efficient design and sue of test of a mental ability for various decision making problems* (No. 7755-23 (project), Series report No.18-16). Texas. USAF School of Aviation Medicine, Randolph Air Force Base.

Birnbaum, A. (1958a). *Further considerations of efficiency in tests of a mental ability* (No. 7755-23 (project), Series report No.17). Texas. USAF School of Aviation Medicine, Randolph Air Force Base.

Birnbaum, A. (1958b). *On the estimation of mental ability.* (No. 7755-23 (project), Series report No.15). Texas. USAF School of Aviation Medicine, Randolph Air Force Base.

Birnbaum, A. (1968). Some Latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. MA. Addison Wesley.

Blackwell, H.R. (1952). The influence of data collection procedures upon psychophysical measurement of two sensory functions. *Journal of Experimental Psychology., 44* (5), 306-315.

Braun, H.I., Jackson, D.N. & Wiley, D.E. (2002). *The Role of Constructs in Psychological and Educational Measurement*. Mahwah. Lawrence Erlbaum Associates.

Brislin, R.W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1* (3), 187-216.

Brislin, R.W. (1980). Translation and content analysis of oral and written materials. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (Vol. 2, pp. 389-444). Boston. Allyn & Bacon.

Brown, R.D. & Harvey, R.J. (2003). *Detecting Personality Test Faking with Appropriateness Measurement. Fact or Fantasy?* Paper presented at the 2003 Annual Conference of the Society for Industrial and Organizational Psychology.

Camara, W.J. (2007). Standards for Educational and Psychological Testing. Influence in Assessment Development and Use Unpublished Paper. The College Board.

Catford, J.A. (1965). *A linguistic theory of translation*. London. Oxford University Press.

CERN. (1956). Retrieved 16.24 20th Oct, 2009, from http.//www.edu.cn/20011114/3009780.shtml.

Chase, C.J. (1978). *Measurement for Educational Evaluation* (2nd ed.). Reading. Addison-Wesley.

Cheung, G.W. & Rensvold, R.B. (2000). Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equations Modeling. *Journal of Cross-Cultural Psychology, 31*, 187-212.

Christiansen, N.D., Burns, G.N. & Montgomery, G.E. (2005). Reconsidering Forced-Choice Item Formats for Applicant Personality Assessment. *Human Performance, 18* (3), 267 - 307.

Christie, R. & Budnitzky, S. (1957). A short forced-choice anxiety scale. *Journal of Consulting Psychology., 21* (6), 501.

Clemans, W.V. (1996). An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monographs, 14*.

Closs, S.J. (1996). On the factoring and interpretation of ipsative data. *Journal of Occupational Psychology, 69*.

Cohen, R.J., Swerdlik, M.E. & Smith., D.K. (1992). *Psychological testing and assessment: an introduction to tests and measurement*. Mountain View. Mayfield Pub. Co.

Cornwell, J.M. & Dunlap, W.P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville & Willson (1991). *Journal of Occupational & Organizational Psychology, 67* (2), 89-100.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York. CBS College Publishing.

Crocker, L.M. (1986). *Introduction to classical and modern test theory*. New York. Holt, Rinehart & Winston.

Cronbach, L.J. (1990). *Essentials of psychological testing* (5th ed.). New York. HarperCollins.

Crowne, D.P. (1960). A New Scale Of Social Desirability Independent Of Psychopathology. *Journal of Consulting Psychology, Vol. 24* (4,), 349-354.

CSPB. (2003). Appendix F. Summary of the Standards for Educational and Psychological Testing *Merit Selection Manual. Policy and Practices*. California. California State Personnel Board.

Day, A.L. & Carroll, S.A. (2008). Faking emotional intelligence (EI). Comparing response distortion on ability and trait-based EI measures. *Journal of Organizational Behavior, 29*, 761–784.

DeFrancis, J. (1990). *The Chinese Language: Fact and Fantasy*. Honolulu. University of Hawaii Press.

Denton, J.C. (1954). Building A Forced-Choice Personality Test. *Personnel Psychology., 7*(4), 449-459.

DeVellis, R.F. (2003). *Scale Development. Theory and Applications* (2nd ed.). London. Sage Publications, Inc.

Dicken, C. (1963). Good Impression, Social Desirability, and Acquiescence as Suppressor Variables. *Educational and Psychological Measurement, VOL. XXII I* (No. 4), 699-720.

Dunn, D.S., Mehrotra, C.M. & Halonen., J.S. (Eds.). (2004). *Measuring up. educational assessment challenges and practices for psychology*. Washington. American Psychological Association.

Eap, S., DeGarmo, D.S., Kawakami, A., Hara, S.N., Hall, G.C. N. & Teten, A.L. (2008). Culture and Personality Among European American and Asian American Men. *Journal of Cross-Cultural Psychology, 39*(5), 630-643.

Ebel, R.L. & Frisbie, D.A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs. Prentice-Hall.

Eckert, T.L., Hintze, J.M. & Shapiro, E.S. (1999). Development and Refinement of a Measure for Assessing the Acceptability of Assessment Methods. The Assessment Rating Profile-Revised. *Canadian Journal of School Psychology, 15* (1), 21-42.

Edwards, A.L. (1957). The Social Desirability Variable in Personality and Assessment and Research. New York. Dryden.

Edwards, A. L. (1970). The Measurement of Personality Traits by Scales and Inventories. New York. Holt.

Edwards, A.L. & Diers, C.J. (1962). Social Desirability and the factorial interpretation of the MMPI. *Educational and Psychological Measurement, 22* (3), 501-509.

Edwards, A.L. & Walsh, J.A. (1964). A factor analysis of scores. *The Journal of Abnormal and Social Psychology, 69* (5), 559-563.

Eid, M. & Diener, E. (Eds.). (2005). *Handbook of multimethod measurement in psychology*. Washington. American Psychological Association.

Eremenco, S.L., Cella, D. & Arnold, B.J. (2005). A Comprehensive Method for the Translation and Cross-Cultural Validation of Health Status Questionnaires. *Evaluation & the Health Professions, 28* (2), 212-232.

Falk, G.H. & Bayroff, A.G. (1954). Rater and technique contamination in criterion ratings. *Journal of Applied Psychology, 38* (2), 100-102.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58* (3), 357-383.

Flaherty, J.A., Gaviria, F.M., Pathak, D., Mitchell, T., Wintrob, R., Richman, J.A. (1988). Developing instruments for crosscultural psychiatric research. *The Journal of Nervous and Mental Disease, 176* (5), 257-263.

Ford, L.H., Jr. (1964). A forced-choice, acquiescence-free, social desirability (defensiveness) scale. *Journal of Consulting Psychology, 28* (5), 475.

Foxcroft, C. & Roodt, G. (2007). *An introduction to Psychological Assessment in the South African Context* (2nd ed.). Cape Town. Oxford University Press.

Ghiselli, E.E. (1954). The forced-choice technique in self-description. *Personnel Psychology, 7* (2), 201-208.

Gipps, C. (1999). Chapter 10. Socio-Cultural Aspects of Assessment *Review of Research in Education, 24* (1), 355-392.

Gipps, C. & Murphy, P. (1994). *A fair test? Assessment, achievement and equity.* Buckingham. Open University Press.

Goldman, I.J. (1964). Effectiveness of the forced-choice method in minimizing social desirability influence. *Journal of Consulting Psychology, 28* (3), 289.

Gordon, L.V. (1951). Validities of the forced-choice and questionnaire methods of personality measurement. *Journal of Applied Psychology, 35* (6), 407-412.

Gordon, L.V. (1976). *Survey of Interpersonal Values: Revised Manual*. Chicago. Science Research Associates.

Gordon, L.V. & Stapleton, E.S. (1956). Fakability of a forced-choice personality test under realistic high school employment conditions. *Journal of Applied Psychology, 40*(4), 258-262.

Gough, H.G. (1947). Simulated patterns on the Minnesota Multiphasic Personality Inventory. *Journal of Abnormal and Social Psychology, 42* (2), 215-225.

Gregory, R.J. (1992). Psychological testing. history, principles, and applications. London. Allyn and Bacon.

Griffith, R.L., Chemielowski, T. & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36* (3), 341-355.

Gulliksen, H. (1950). *Theory of Mental Tests*. New York. John Wiley & Sons Inc.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educationtional Measurement* (3rd ed., pp. 147-200). New York. Macmillan.

Hambleton, R.K. (Ed.). (1983). *Application of item response theory*. Vancouver, BC. Educational Research Institute of British Columbia.

Hambleton, R.K. & Cook, L.L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14*, 75-96.

Hambleton, R.K. & Jones, R.W. (1993). An NCME Instructional Module on Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. Unpublished An NCME Instructional Module. University of Massachusetts at Amherst.

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Hingham. Kluwer-Nijhoff Publishing.

Hambleton, R.K., Swaminathan, H. & Jane Rogers, H. (1991). *MMSS: Fundamentals of Item Response Theory*. London. Sage Publications, Inc.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park. Sage Publications.

Hammer, A.L. & Mitchell, W.D. (1996). The distribution of MBTI types in the US by gender and ethnic group. *Journal of Psychological Type, 37*, 2-15.

Harth, E. (2004). Art and Reductionism. *Journal of Consciousness Studies, 11* (3-4), 111–116.

Harvey, R.J. & Thomas, L.A. (1996). Using Item Response Theory to Score the Myers-Briggs Type Indicator: Rationale and Research Findings. Paper presented at the SIOP 2007 NYC.

Hashimoto, M.J. (1973). The Hakka dialect. a linguistic study of its phonology, syntax, and lexicon. Cambridge. University Press.

Haynes, S.N. & O'Brien, W.H. (2000). *Principles and practice of behavioral assessment*. Dordrecht. Kluwer Academic/Plenum.

Heineman, C.E. (1953). A forced-choice form of the Taylor Anxiety Scale. *Journal of Consulting Psychology, 17*(6), 447-454.

Hendrickson, T.M. (Undated/1958). Personal profile analysis: a technical manual. Marlow. Thomas International Systems (Europe) Ltd.

Hendrickson, T.M. (1983). Personal Profile Analysis: A Technical Manual. Marlow. Thomas International Systems (Europe) Ltd.

Hendrickson, T.M. (2007). UK Patent No. ISBN 185433512X. British Psychological Society Psychological Testing Centre. B. P. S. P. T. Centre.

Henry, G.T. (1990). *Practical Sampling* (Vol. 21). London. Sage Publications, Inc.

Hicks, L.E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74* (3), 167-184.

Hirsh, J.B. & Peterson, J.B. (2008). Predicting creativity and academic success with a "fake-proof" measure of the Big Five. *Journal of Research in Personality, 42* (5), 1323-1333.

Hogan, T.P. (2003). Psychological testing: a practical introduction. New York. Wiley & Sons.

Hogan, T.P. (2007). *Psychological testing: a practical introduction* (2nd ed.). Hoboken. John Wiley & Sons.

Hooft, G.T. (2001). The obstinate reductionist's point of view on the laws of physics. Unpublished Lecture given at the 2001 Technology Forum. Institute for Theoretical Physics Utrecht University.

Howe, E.S. & Silverstein, A.B. (1960). Comparison of two short-form derivatives of the Taylor Manifest Anxiety Scale. *Psychological Reports, 6*, 9-10.

HPCSA. (2006a). Policy on The Classification of Psychometric Measuring Devices, Instruments, Methods and Techniques (Form 208). Pretoria. HPCSA.

HPCSA. (2006b). The Professional Board For Psychology, HPCSA, List of Tests Classified As Being Psychological Tests (Form 207). Pretoria. HPCSA.

HPCSA. (2006c). South African Guidelines on Computerised Testing (Form 257). Pretoria. HPCSA.

HPCSA. (2008). The Professional Board For Psychology, HPCSA, Classification Review Form (Form 205). Pretoria. HPCSA.

HPCSA. (2009). The Professional Board For Psychology, HPCSA, Classification Review Form (Form 205). Pretoria. HPCSA.

Humm, D.G. (1939a). Discussion of "A statistical analysis of the Humm-Wadsworth Temperament Scale". *Journal of Applied Psychology, 23* (4), 525-526.

Humm, D.G. (1939b). Dysinger's critique of the Humm-Wadsworth temperament scale. *The Journal of Abnormal and Social Psychology, 34* (3), 402-403.

Humm, D.G., & Wadsworth, G.W. (1933). *A Diagnostic Inventory of Temperament, Preliminary Report.* Paper presented at the Western Psychological Association Meetings, University Of Southern California, Los Angeles.

Humm, D.G. & Wadsworth, G.W.J. (1939). *The Humm-Wadsworth Temperament Scale*. Los Angeles. Doncaster G Humm.

Irvine, S.H. (2003). Personal Profile Analysis: The Technical Resource Book. Marlow. Thomas International.

Irvine, S.H. (2007). PPA Alternate Form Reliability Study Unpublished Research. Thomas International Inc.

Irvine, S.H., Mettam, D. & Syrad, I. (1994). Valid and more valid? Keys to understanding personal appraisal at work. *Current Psychology. Developmental, Learning, Personality, Social, 13* (1), 27-59.

Johnson, C.E., Wood, R. & Blinkhorn, S.F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61* (2), 153-162.

Jooste, M.J.L. (2003). Introduction To Psychological Testing in South Africa. Unpublished Class Note. University of Johannesburg.

Kaiser, P.D. & Smith, K. (2001). *The Standards for Educational and Psychological Testing. Zugzwang for the Practicing Professional?* Paper presented at the The International Personnel Management Association Assessment Council.

Kaplan, R.M. & Saccuzzo, D.P. (2005). *Psychological testing: principles, applications, and issues* (6th ed.). Belmont. Wadsworth/Thomson Learning.

Kim, M.T. & Han, H.R. (2004). Cultural considerations in research instrument development. In M. Frank-Stromberg & S. J. Olsen (Eds.), *Instruments for clinical health care research* (3rd ed., pp. 73-81). Boston. Jones & Bartlett.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). New York. Routledge.

Kline, T.J.B. (2005). Psychological Testing: A Practical Approach to Design and Evaluation. Thousand Oaks. Sage Publications.

Komar, S., Brown, D.J., Komar, J.A. & Robie, C. (2008). Faking and the Validity of Conscientiousness. A Monte Carlo Investigation. *Journal of Applied Psychology, Vol. 93* (No. 1), 140–154.

Kruger, B.L. (1938). A statistical analysis of the Humm-Wadsworth temperament scale. *Journal of Applied Psychology, 22*(6), 641-652.

Lönnqvist, J.-E. (2008). Issues in socially desirable responding and personality research.

Lawley, D.N. (1943). On Problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 6*, 273-287.

Lawley, D.N. (1944). The Factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburgh, 62-A*, 74-82.

Lazarsfeld, P.F. (Ed.). (1950). *The logical and mathematical foundation of latent structure analysis*. Princeton. Princeton University Press.

Lee, C.-C., Li, D., Arai, S. & Puntillo, K. (2009). Ensuring Cross-Cultural Equivalence in Translation of Research Consents and Clinical Documents. *Journal of Transcultural Nursing, 20* (1), 77-82.

Lepkowski, J.R. (1963). Development of a forced-choice rating scale for engineer evaluation. *Journal of Applied Psychology., 47* (2), 87-88.

Levashina, J. & Campion, M.A. (2007). Measuring faking in the employment interview. Development and validation of an interview faking behavior scale. *Journal of Applied Psychology, 92* (6), 1638-1656.

Linacre, J. M. (2009). A User's Guide to Winstesp / Ministep Rasch-Model computer Program (Program Manual 2.68.0) (Version 2.68) [Computer Software]. Chicago IL. winsteps.com.

Linacre, J.M. & Wright, B.D. (1987). *Mantel-Haenszel and the Rasch Model*. Paper presented at the MESA Psychometric Laboratory, Department of Education, University of Chicago.

Lord, F.M. (1952). A Theory of test scores. *Psychometric Monograph, 7*.

Lord, F.M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 13*, 57-75.

Lord, F.M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

Lord, F.M. (1974). Individualized testing and item characteristic curve theory. In D.H. Krantz, R.D. Atkinson, R.D. Luce & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. II). San Francisco. Freeman.

Lord, F.M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95-100.

Lord, F.M. (1980a). *Applications of Item Response Theory To Practical Testing Problems*. Hillsdale. Lawrence Erlbaum Associates, Inc.

Lord, F.M. (1980b). Some how and which for practical tailored testing. In L.J.T. Van der Kamp, W.F. Langerak & D.N.M. de Gruijter (Eds.), *Psychometrics for Educational Debates*. New York. Wiley.

Lorenz, M.G. & Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol Rev., 58* (3), 563–602.

Martinussen, M., Richardsen, A.M. & Varum, H.W. (2001). Validation of an ipsative personality measure (DISCUS): Scandinavian Journal of Psychology, 42, 411-416.

Matthews, G. & Oddy, K. (1997). Ipsative and Normative Scales in Adjectival Measurement of Personality. Problems of Bias and Discrepancy. *International Journal Of Selection And Assessment, 5* (3), 169-182.

McCloy, R.A., Heggestad, E.D. & Reeve, C.L. (2005). A Silk Purse From the Sow's Ear. Retrieving Normative Information From Multidimensional Forced-Choice Items. *Organizational Research Methods, 8* (2), 222-248.

McKee, S.P., Klein, S.A. & Teller, D.Y. (1985). Statistical properties of forced-choice psychometric functions. implications of probit analysis. *Perception & Psychophysics, 37* (4), 286-298.

McKinley, J.C., Hathaway, S.R. & Meehl, P.E. (1948). The Minnesota Multiphasic Personality Inventory. VI. The K Scale. *Journal of Consulting Psychology, 12*(1), 20-31.

Meisenberg, G. & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences, 44*, 1539–1550.

Mellenbergh, G.J. (1989). Item Bias and Item Response Theory. *International Journal of Educational Research, 13*, 127-143.

Middleton, K.L. & Jones, J.L. (2000). Socially Desirable Response Sets: The Impact of Country Culture. *Psychology & Marketing, 2*, 149-163.

Mills, C.J. & Parker, W.D. (1998). Cognitive-Psychological Profiles of Gifted Adolescents From Ireland and the U.S. Cross-Societal comparisons. *International Journal of Intercultural Relations, 22* (1), 1-16.

Morgeson, F.P., Campion, M.A., Dipboye, R.L. & Hollenbeck, J.R. (2007). Reconsidering The Use Of Personality Tests In Personnel Selection Contexts. *Personnel Psychology, 60* (3), 683-729.

Murphy, K.R. (1988). *Psychological testing: principles and applications*. Englewood Cliffs. Prentice-Hall.

Murphy, K.R. & Davidshofer, C.O. (1998). *Psychological Testing: Principle and Applications* (5th ed.). London. Prentice-Hall Inc.

Nathanson, C. & Paulhus, D.L. (2004). A Differential Item Functioning (DIF) Analysis of the Self-Report Psychopathy Scale. Poster presented at the 1st biannual meeting of the Society for the Scientific Study of Psychopathy.

Nederhof, A.J. (1985). Methods of coping with social desirability bias: A review. *European Journal of social Psychology, 15* (3), 263-280.

Nida, E. (1964). Toward a science of translating. In E. J. Brill (Ed.). *With special reference to principles and procedures involved in Bible translating*. Leiden. Netherlands.

Ockham's_Razor. (2009). Encyclopædia Britannica. Retrieved November 02, 2009, from Encyclopædia Britannica Online. http.//www.britannica.com/EBchecked/topic/424706/Ockhams-razor.

Paulhus, D.L. (1981). Control of Social Desirability in Personality Inventory Principal-Factor Deletion. *Journal of Research in Personality, 15*, 383-388.

Paulhus, D.L. (1991). Measurement and Control of Response Bias. In R.J.P., P. Shaver & L.S. Wrightsman (Eds.). *Measures of Personality and Social Psychological Attitudes* (pp. 17-36). San Diego. Academic Press.

Paulhus, D.L. (2002). Socially Desirable Responding. The Evolution of a Construct. In H.I. Braun, D.N. Jackson & D.E. Wiley (Eds.), *The Role of Constructs in Psychological and Educational Measurement* (pp. 49-69). Mahwah. Erlbaume.

Paulhus, D.L. (2003). Self Presentation Measurement. In R. Fernandez-Ballesteros (Ed.), *Encyclopaedia of Psychological Assessment* (pp. 858-861). Thousand Oaks. Sage.

Paulhus, D.L., Fridhandler, B. & Hayes, S. (1997). Psychological Defense. Contemporary Theory and Research. In R. Hogan, J. Johnson & S. Briggs (Eds.), *Handbook Of Personality Psychology* (pp. 543-561). Toronto. Academic Press.

Peele, S. (1981). Reductionism in the Psychology of the Eighties: Can Biochemistry Eliminate Addiction, Mental Illness, and Pain? *American Psychologist., 36*, 807-818.

Petrides, K.V. (2009). Technical manual for the Trait Emotional intelligence Questionnaire (TEIQue) (1st ed.). London. London Psychometric Laboratory.

Puhan, G. & Gierl, M. J. (2006). Evaluating the Effectiveness of Two-Stage Testing on English and French Versions of a Science Achievement Test. *Journal of Cross-Cultural Psychology, 37* (2), 136-154.

Rachlin, H. (2002). Altruism and selfishness: *Behavioral and Brain Sciences, 25* (2), 239-296.

Rao, C.R. & Sinharay, S. (2007). *Psychometrics: handbook of statistics 26*. New York. Elsevier.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen. Danish Institute for Educational Research.

Rasch, G. (1966a). An individualistic approach to item analysis. In P.F. Lazarsfeld & H N.V. (Eds.), *Readings in Mathematical social science* (pp. 89-107). Chicago. Science Research Association.

Rasch, G. (1966b). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*, 49-57.

Rees, C.J. & Metcalfe, B. (2003). The faking of personality questionnaire results: who's kidding whom? *Journal of Managerial Psychology, 18* (2), 156-165.

Rentz, R.R. & Bashaw, W.L. (1975). *Equating reading tests with Rasch model, Volume I final report, Volume II technical reference tables*. Athens. University of Georgia, Educational Research Laboratory.

Richardson, M.W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika, 1* (33-49).

Richardson, M.W. (1951). Note on Travers' critical review of the forced-choice technique. *Psychological Bulletin., 48* (5), 435-437.

Rizopoulos, D. (2006). Item. An R Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software, 17* (5).

Rouse, B.A., Kozel, N.J. & Richards, L.G. (1985). *Self-Report Methods of Estimating Drug Use. Meeting Current Challenges to Validity* (Vol. 57). Washington. US Department of Health and Human Services.

Rupinski, M.T. & Dunlap, W.P. (1996). Approximating Pearson product-moment correlations from Kendall's tau and Spearman's rho. *Educational and Psychological Measurement, 56* (3), 419-429.

Sakeim, H.A. & Cur, R.C. (1978). Self-deception, other-deception and consciousness. In G.E. Schwartz & S.D. (Eds.), *Consciousness and self-regulation. Advances in research* (Vol. 2, pp. 139-197). New York. Plenum Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, 17*.

Samejima, F. (1972). A general model for free-response data. *Psychometric Monograph, 18*.

Samejima, F. (1973a). A comment on Birnbuam's Three parameter logistic model in the latent trait theory. *Psychometrika, 38*, 221-223.

Samejima, F. (1973b). Homogeneous case of the continuous response model. *Psychometrika, 38*, 203-219.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39* (1), 111-121.

Samejima, F. (1977a). A method of estimating item ahcracteristic functions using the maximum Likelihood estimate of ability. *Psychometrika, 42*, 163-191.

Samejima, F. (1977b). A Use of the Information Function in Tailored Testing. [Journal Article]. *Applied Psychological Measurement, 1* (2), 233-247.

Samejima, F. (1996). *Polychotomous Response and the Test Score*. Paper presented at the 1996 Annual Meeting of NCME.

Samejima, F. (1999). *General Graded Response Model*. Paper presented at the The1999 Annual NCME Meeting.

SATP. (2003). HPCSA: List of tests classified as being psychological tests. Complied by the Psychometrics Committee of the Professional Board for Psychology. from http.//www.apt.org.za/form207.htm.

Saville, P. & Wilson, E. (1991). The reliability and validity of normative and ipsative approaches in the measurement of personality. *Journal of Occupational Psychology, 64*, 219-238.

Schumacker, R. E. (2005). Classical Test Analysis. Unpublished Class Note. Applied Measurement Associates.

Schutter, G. & Maher, H. (1956). Predicting grade-point average with a forced-choice study activity questionnaire. *Journal of Applied Psychology, 40*(4), 253-257.

Scott, W.A. (1963). Social desirability and individual conceptions of the desirable. *The Journal of Abnormal and Social Psycholog , 67* (6), 574-585.

Shaw, E. (2007). Kendall/Spearman rank correlation. Retrieved 2008, 14th Sep, from http.//sci.tech-archive.net/Archive/sci.stat.math/2007-08/msg00189.html.

Sim, S.-M. & Rasiah, R.I. (2006). Relationship Between Item Difficulty and Discrimination Indices in True/False- Type Multiple Choice Questions of a Para-clinical multidisciplinary Paper. *Annals Academy of Medicine Singapore, 35* (2), 67-71.

Sireci, S.G., Yang, Y., Harter, J. & Ehrlich, E.J. (2006). Evaluating Guidelines For Test Adaptations *Journal of Cross-Cultural Psychology, 37* (5), 557-567.

Sisson, E.D. (1948). Forced choice-the new Army rating. *Personnel Psychology, 1* (2), 365-381.

Soto, C.J., John, O.P., Gosling, S.D. & Potter, J. (2008). The Developmental Psychometrics of Big Five Self-Reports: Acquiescence, Factor Structure, Coherence, and Differentiation From Ages 10 to 20. *Journal of Personality and Social Psychology, Vol. 94* (No. 4), 718–737.

Staff, Personal Research and Procedures Branch, the Adjutant General's Office. (1946). *The Forced-Choice technique and rating scales.* Paper presented at the 54th annual meeting of American Psychological Association, Philadelphia, Pennsylvania.

Szöllösi-Janze, M. (2009). The natural sciences and democratic practices: Albert Einstein, Fritz Haber, and Max Planck. *Bulletin of the ghi, 44*, 10~22.

Szirmák, Z. & Raad, B.D. (1994). Taxonomy and structure of Hungarian personality traits. *European Journal of Personality, 8* (2), 95 - 117.

Tallent, N. (1992). *The practice of psychological assessment*. Englewood Cliffs. Prentice-Hall.

Taylor, J.A. (1953). A personality scale of manifest anxiety. *The Journal of Abnormal and Social Psychology., 48* (2), 285-290.

Tenopyr, M.L. (1988). Artifactual reliability of forced-choice scales. *Journal of Applied Psychology, 73* (4), 749-751.

Thompson, B. (Ed.). (2003). Score Reliability: Contemporary Thinking on Reliability issues. London. Sage Publications, Inc.

Torrance, H. (1981). The Origins and Development of Mental Testing in England and the United States. *British Journal of Sociology of Education, 2* (1), 45-59.

Torrance, H. (1993). Formative Assessment: some theoretical problems and empirical questions. *Cambridge Journal of Education, 23* (3), 333 - 343.

Travers, R.M.W. (1951). A critical review of the validity and rationale of the forced-choice technique. *Psychological Bulletin., 48* (1), 62-70.

Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1-13.

UCLA.ATS. (2006). SPSS FAQ. What does Cronbach's alpha mean? . from UCLA. Academic Technology Services. http://www.ats.ucla.edu/stat/Spss/faq/alpha.html.

Underhill, C.M., Lords, A.O. & Bearden, R.M. (2006). Fake Resistance Of A Forced-Choice Paired-Comparison Personality Measure. Navy Personnel Research, Studies, and Technology (NPRST).

Van De Vijver, A.J.R. & Rothmann, S. (2004). Assessment In Multicultural Groups. The South African Case. *SA Journal of Industrial Psychology, 30* (4), 1-7.

Van der Merwe, R.P. (2002). Psychometric testing and human resource management. *South African Journal of Industrial Psychology, 28* (77-86).

Vasilopoulos, N.L., Cucina, J.M., Dyomina, N.V., Morewitz, C. L. & Reilly, R.R. (2006). Forced-Choice Personality Tests. A Measure of Personality and Cognitive Ability? *Human Performance, 19* (3), 175-199.

Vernon, P.E. (1934). The Attitude of the Subject in Personality Testing. *18* (2), 165-177.

Wagner, T.A. & Harvey, R.J. (2003). *Developing A New Critical Thinking Test Using Item Response Theory*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology.

Walsh, B. & Betz, N.E. (1990). *Tests and assessment* (2nd ed.). Englewood Cliffs. Prentice Hall.

Walsh, W.B. & Betz, N.E. (1995). *Tests and assessment* (3rd ed.). Englewood Cliffs. Prentice Hall.

Wang, W.-L. & Lee, H.-L. (2006). Challenges and Strategies of Instrument Translation *Western Journal of Nursing Research, 28* (3), 310-321.

Weiss, D.J. (1976). Adaptive testing research at Minnesota: Overview, recent result, and future directions. In C.C.L. (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing*. Washington. United States Civil Service Commission.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing *Applied Psychological Measurement, 6*, 379-396.

Weiss, D.J. (Ed.). (1978). *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis. University of Minnesota.

Weiss, D.J. (Ed.). (1980). *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis. University of Minnesota.

Weiss, D.J. (Ed.). (1983). *New horizons in testing*. New York. Academic Press.

Wherry, R.J. (1959). An Evaluative and Diagnostic Forced-Choice Rating Scale for Serviceman. *Personnel Psychology, 12* (2), 227-236.

Wigdor, A. & Garner, W. (1982). *Ability testing. Uses, consequences and controversies, Part I.* Washington. National Academy Press.

Williams, R.H., Zimmerman, D.W., Zumbo, B.D. & Ross, D. (2003a). Charles Spearman. British Behavioural Scientist. *Human Nature Review, 3*, 114-118

Williams, R.H., Zimmerman, D.W., Zumbo, B.D. & Ross, D. (2003b). Charles Spearman. British Behavioural Scientist. *The Human Nature Review, 3*, 114-118.

Willson, V.L. (1977). Robinson's Measure of Agreement as a Parallel Forms Reliability Coefficient. (No. ED201667).

Wright, B.D. & Douglas, G. A. (1977). Best Procedures For Sample-Free Item Analysis. *Applied Psychological Measurement, 1*(2), 281-295.

Wright, B.D. & Stone, M. H. (1979). *Best test design*. Chicago. MESA.

Yasuda, T., Lei, P.-W. & Suen, H. K. (2007). Detecting Differential Item Functioning in the Japanese Version of the Multiple Affect Adjective Check List—Revised. *Journal of Psychoeducational Assessment, 25* (4), 373-384.

Yu, A. (2008). Cronbach Alpha--Educational Assessment course by Dr. Alex Yu. *Using SAS for Item Analysis and Test Construction* Retrieved September 1, 2008, from http://www.creative-wisdom.com/teaching/assessment/alpha.html.

Yu, D.S.F., Lee, D.T.F. & Woo, J. (2004). Issues and Challenges of Instrument Translation *Western Journal of Nursing Research, 26* (3), 307-320.

Zavala, A. (1965). Development of the forced-choice rating scale technique. *Psychological Bulletin, 63* (2), 117-124.

Zechmeister, E.B. & Posavac., E.J. (2003). *Data analysis and interpretation in the behavioral sciences*. Belmont. Thomson/Wadsworth.

Zimmerman, D.W.W., Richard H. (1980). Is classical test theory "robust" under violation of the assumption of uncorrelated errors? *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 34* (3), 227-237.

Zumbo, B.D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly, 4* (2), 223-233.

Wong, W. J. (2007). *Fuzzy, Understanding Fuzzy Logic* (3rd ed.). Taipei. Chan-Wha Inc.

Yu, M.N. (2002). Educational testing and assessment. Achievement test and educational appraisal (2nd ed.). Taipei. Psych Taipei Inc.

Yu, M.N. (2007). Introduction of Item Response Theory. Retrieved 4th of June, 2007, from http.//www.irt.org.tw/index.php?mod=b4.

Jen, M.F. (2001). *Psychological Testing and Statistics* (3rd ed.). Taipei.

APPENDIX A

Mathematical expression (refer to section 4.9)

Definition of axis

$$X = \frac{cf_l + .5(f_i)}{N} \times 100\% \;\; ; \; Y = P(R_X)$$

cf_l =Cumulative frequency of all scores lower than the score of interest

f_i =Frequency of the score of interest

N=Number of examines in the sample

P(R_X)=Probability of target response type of a particular point of percentile rank

Parameter estimation (refer to section 4.9)

$$\rho_{P_1(\theta)\theta} = \frac{\sigma P_1(\theta)\theta}{\sigma P_1(\theta)\sigma\theta}$$

For testing the existence of such relatioship

$\hat{P}_1(\theta) = b\theta + a$

for slope perdiction and intercept

with Ogive value 1.7 of slope

C=minimum probability (y intercept or minimal data)

Plotting formula (3PL) (refer to section 4.9)

$P_i(\theta) = C_i + (1 - C_i)\frac{e^{\mathrm{D}a_i(\theta - b_i)}}{1 + e^{\mathrm{D}a_i(\theta - b_i)}}$ (Formula 1.1) (Baker, 1992; Lord, 1980a)

i = item

a_i = Item Discrimination Parameter. (Item's abitliy to discriminate high-Low DISC individuals)

C_i = Pseudo-chance Parameter (Chance of response M, L, and no respose without DISC construct score)

b_i = Difficulty Parameter (Strength of DISC item Parameter, when item pass 0.5

$D = 1.7$ (constant value, a scaling factor introduced to make the logistic function as close as possible to the normal ogive function.)

$P_i(\theta)$ = Probability of correct response
(Probability of Response, response can be M, L, and no response)

$\theta =$ Ability (Strength of DISC construct)

IRT modification of Chi Square Goodness of Fit (refer to section 4.9)

$\sum_{i=1}^{k}\frac{\left(O_i - E_i\right)^2}{E_i}$ (4-8)

O_i = Observed response probability of selected response

E_i = Expected response probability of selected response for target ability (θ_j) via current parameter estimation method

df= (r-1) x (c-1)

According to Baker (1992), the ability value can be roughly estimated from the Z score. The formula can be expressed as following.

$$\because \mathrm{Z}_{ij} = \left(\frac{\theta_j - \mu_i}{\sigma_i} \right) = \alpha_i(\theta_j - \beta_i)$$

$$\therefore \mathrm{Z}_{ij} = \alpha_i(\theta_j - \beta_i) \qquad \text{(4-9)}$$

$$\therefore \theta_j = \left(\frac{\mathrm{Z}_{ij}}{\alpha_i} \right) + \beta_i$$

This value is used as an indication of ability (θ_j) in the current study.

Mathematical expression of GRM in R (Rizopoulos, 2006)**. (refer to section 4.10)**

$$P(x_{im} = k \mid z_m) = g(\eta_{ik}) - g(\eta_{i,k+1}),$$
$$\eta_{ik} = \alpha_i(z_m - \beta_{ik}),\ k = 1.....,K_i, \qquad \text{(4-10)}$$

$\beta_{i1} < ... < \beta_{ik} < ... < \beta_{ik-1}$; $\beta_{iK_i} = \infty$

x_{im} = Ordinal manifest variable with K_i possible response categories.

K_i = Number of Response Categories (i.e. For PPA, 3 response categories, M, blank, and L)

z_m = *m*th respondent (candidate) in the latent trait continuum, represent ability θ_j

α_i = Discrimination parameter

β_{ik} = Extremity parameter with $\beta_{i1} < ... < \beta_{ik} < ... < \beta_{ik-1}$; $\beta_{iK_i} = \infty$. The cut-off point in the cumulative probabilities scale

Above mathematic is expressed as “grm( )” in R syntax language.

MMLE in general: (refer to section 4.10)

The formula for the *m*th sample unit is:

$$\ell_m(\theta) = \log p(x_m;\theta) = \log \int p(x_m|z_m;\theta)p(z_m)dz_m \qquad \text{(4-11)}$$

$p(\cdot)$ = Probability density function

x_m = Vector of responses for the *m*th sample unit.

z_m = Latent ability that is assumed to follow a standard normal distribution and $\theta = (\alpha_i, \beta_i)$

MMLE GRM specification in R. (refer to section 4.10)

$$\log\left(\frac{\gamma_{ik}}{1-\gamma_{ik}}\right) = \beta_{ik} - \beta_i z \qquad \text{(4-12)}$$

γ_{ik} = Cumulative probability of a response in category *k*th or lower to the *i*th item, given the latent ability *z*.

K = *k*th options within an item

i = *i*th item on the list

z = latent ability