Towards a culture-free model of the Big Five - a cross-cultural investigation of the Orpheus in four different language families

Lina Daouk-Oyry

Thesis submitted in fulfilment

of the requirements for the degree of

Doctor of Philosophy

Department of Psychology
City University, London
September 2008

Volume 2

Table of Contents

List of Tables

List of figures

Acknowledgments

"Hi lulu, teta told me that on wednsday somebody I don't know who will ask you some questions and then if they are all correct you will be a doctor, right?"
My 9 year old niece, Yasmine Daouk

I would like to thank my supervisor Dr Almuth McDowall. Almuth offered me great support, encouragement, and extremely constructive feedback throughout my PhD, and has been extremely inspirational on an academic, practitioner, and personal levels. Almuth contributed hugely to my personal development, by helping me develop my writing, organisational and research skills. I would also like to thank my previous supervisor Professor John Rust for believing in my abilities and for offering me to use Orpheus for my PhD. John supervised my MSc thesis too and was instrumental in raising my interest in Psychometrics and cross-cultural studies. He also opened many doors for me that lead to my current professional position.

I am most thankful to my husband Toni, to whom I dedicate this PhD. Toni had to learn about Psychometrics through listening to me practicing my conference talks, hearing my complaints near every obstacle I faced, and celebrating with me near every achievement I made throughout the PhD. Toni's unconditional love, support, and belief in me alleviated the weight of the PhD and made this journey a much more enjoyable one. When I was overwhelmed by the work involved, Toni used rock climbing to teach me how to focus on smaller goals in order to achieve the bigger ones, and it worked! I am also indebted to Toni for reading through the whole PhD and helping me edit it.

I would also like to thank my parents Oussama and Afif whose efforts, encouragement and trust lead me to London to study for my MSc. I would also like to thank the department of Psychology at City University for awarding me a full scholarship for the PhD and the Psychometrics Centre for awarding me further funding to support my PhD. I am also thankful to my sister Rania for going out of her way to help me collect data and for assisting me with her critical thinking in interpreting some of the culturally related findings. I am also thankful to her husband Mazen, who believed I would end up with a PhD before I even started my MSc; I will never forget that conversation we had in New York. My brothers Mazen and Samer, and his wife Nancy, were also great support in helping me with my data collection and being so proud of me.

Several friends and colleagues offered to dedicate some of their time to read through my chapters and give me their comments: Richard Davies, Peter Martin, Byron, Cristina Pittelli, Nicky Schlatter, and Vicky Ellam-Dyson. Their feedback was invaluable and greatly appreciated at the time where I needed it the most. Toni Oyry and Maha Taki did a wonderful job in transcribing my interviews. Maha and Souraya Karami also participated in the interviews, which provided great input to my PhD. The international aspect of my PhD also introduced me to wonderful people, who continued to answer my questions even after they moved away from London: Ou Lan from China, and Carmen Munoz from Spain. I would also like to thank Shining Li for going all the way back home to help me with my Chinese data collection and also for participating in interviews.

I would also like to thank Professor David Marks and Professor James Hampton for acting as my internal supervisors and offering their help when needed. I extend my thanks to other members of staff in the Psychology department for their advice and support.

A big thank you to Daniel Heussen, Silvio Aldrovandi and Sebastian Gaigg, for listening to my statistically loaded questions and helping me solve them, but also for being great colleagues and friends during and after the PhD. I also thank all the other colleagues in rooms 509 and 415, with whom I shared the stress and joy of the PhD.

Last but not least, I would like to than all my participants, friends, colleagues and family members for spending time filling out my questionnaires and taking part in my interviews and forwarding it to their networks of friends and colleagues.

"The librarian protects the books not only against mankind but also against nature and devotes his life to this war with the forces of oblivion."

Umberto Eco

Declaration

I grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Abstract

This thesis concerns the development of a practical and theoretical framework for adapting of questionnaires building on van de Vijver and Leung's (1997) Theory of Equivalence and Bias. In contrast to extant research which has largely concentrated on the adaptation of ability measures the present research was operationalised through adapting and translating Orpheus, a work-based Big Five personality questionnaire, into English, Arabic, Chinese (Mandarin) and Spanish.

The first phase, 'Quality Control', used a mixed method technique in two studies. Study 1 (Translation and Monitoring) was qualitative and used *forward and back translation* followed by *dyads and triads*. Results from this study (n = 10) reflected the importance of qualitative judgment techniques in test adaptation and showed the emergence of three main types of bias (linguistic, psychological, and conceptual), which were discussed in the literature review but do not constitute part of the Theory of Equivalence and Bias (van de Vijver & Leung, 1997). Study 2 (n = 185) (Pre-Testing) and Study 3 (n = 12) (Cognitive interview) combined quantitative (pre-test) and qualitative techniques (cognitive interviews). Results were inconclusive as to what extent p values or Cohen's d is better at detecting potential problems in adaptation of items. Cognitive interviews were shown to be effective for interpreting statistically significant results as they unravelled many linguistic, psychological, and cultural problems that went unnoticed in back translation dyads/triads.

The second phase ('Field Pilot') was laid out over two studies that used the same data but focused on different statistical investigations. Study 4 (n=815) centred on item bias analysis using Logistic Regression as well as ANOVA and showed that 12 items in Arabic, 11 in Chinese and 3 in Spanish were functioning differently than the English version of the items. Study 4 examined the metric equivalence between the four groups using EFA and MG-CFA. Results showed that no model fits the data as it was. Intrinsic test problems and using criterion-related validity as a sole method of validation were identified as two potential causes of model failure.

To Beirut and to my husband who adores Beirut…

إلى بيروت ست الدنيا وإلى زوجي الذي يعشق بيروت...

Chapter 8: Reliability and Differential Item Functioning Analysis

8.1. Chapter overview

The first two studies in chapters 6 and 7 represented the *quality control phase* that was aimed at maximising the linguistic, psychological and cultural equivalence of the adapted versions. This was achieved through the use of mixed method techniques, relying on several bilingual speakers and psychometrics experts, and piloting the adapted versions. The following two chapters outline the second phase of the adaptation process, *piloting*. Both chapters are based on the same methodology, which involves administering the four parallel versions, including the original English one, to approximately 200 participants in each culture. However, they differ with regards to the focus of the statistical analysis employed.

This chapter investigates the reliability of Orpheus scales, in addition to item difficulty and discrimination across the four cultures as recommended by van de Vijver and Leung (1997). The second part of this chapter focuses on the examination of differential item functioning (DIF) across cultures.

We will begin this chapter by reviewing the concepts of bias and fairness and the techniques that are usually implemented to assess DIF. While there are many such techniques, we will focus mainly on Logistic Regression (Zumbo, 2003) and ANOVA (van de Vijver & Leung, 1997) as two techniques that can be applied on polytomous data. Although Logistic Regression will be used as the main analysis in this study, its results will be compared to ANOVA to investigate any differences they produce.

We will conclude this chapter with a discussion of the findings and their implication for future adaptation of work based personality tests into Arabic, Chinese, and Spanish.

8.2. Fairness as a social concept

Fairness is a fundamental element in assessment in organisations from a business, ethical, and legal point of view (Gilliland, 1993). While the definition of fairness is context dependent and can vary accordingly, it is reasonable to say that fairness is a social concept that assumes equality in treatment between different groups of people (SIOP, 2003). Following up on earlier discussion in chapter 4, fairness is not a psychometric principle as such, but rather relates to the inferences that are drawn from any assessment method and the equal treatment of candidates that may or may not result from this assessment (SIOP, 2003). An assessment method that does not treat participants equally is unfair, and therefore biased. Nevertheless, the definition and implementation of the concept of fairness in assessment is not that simple. Fairness in assessment is multifaceted because it requires that 1) all the attributes necessary for a job are assessed, 2) the method of assessment used is empirically proven to be valid, reliable and free from bias, and also that 3) these methods need to be perceived as fair and reasonable by all parties involved (Bartram, 2005).

8.2.1 Fairness in multi-cultural environments

Employment laws in different countries, such as the US and the UK, also define how fairness is perceived and implemented in selection and recruitment contexts (Baron, & Janman, 1996). These laws are increasingly moving towards encouraging diversity through the introduction of equal opportunities legislations (Trubek & Mosher, 2003). In the UK for example, the groups that employers need to ensure fairness against include gender (sex discrimination act 1975), race (race relations act, 1976 and 2000), disability, (disability act, 1998) and more recently age (age discrimination act, 2006). This adds to the complexity of assessing fairness because it adds new groups, in addition to gender, that the assessment needs to be ensure fairness to.

This issue becomes more challenging when dealing with multi national organisations. Fairness in assessment needs to be insured within each country but also between the countries where the organisation is operating. This renders the implementation of fairness even more complex by adding one more group of interest, that is, culture. New sources of bias are likely to emerge in such context, hence threatening the fairness in assessing employees.

8.2.2 Fairness in psychometrics

From a psychometrics perspective, ensuring fairness is achieved through minimising bias, which becomes more challenging to achieve when the number of comparison groups increases and test adaptation is implemented. Jenson (1980, in Kline, 1993) explains that some items in psychometric tests are based on aspects that might be more common or familiar to one culture group than another, which he referred to as culture bound fallacy. In personality assessment, for example, each item pertains to a specific psychological construct based on behaviours associated with it. Some of these behaviours might have alternative interpretation in different cultures, which could make them culturally bound and likely to result in bias. For example, standing up to your seniors might be a behaviour indicative of tough mindedness in the workplace in the UK, but it is unlikely to denote that in a culture where standing up to seniors is against cultural values such as in China.

8.2.3 Fairness and bias in psychometrics

Fairness encompasses the psychometric properties of a test to cover, as mentioned earlier, the appropriateness of using a certain method of assessment and how the recipient views it. Fairness of psychometric tests can be assessed by investigating its validity and reliability, but most importantly its freedom from bias against groups of interest. Fairness and bias are not parallel concepts but closely related ones. When an item is biased, it is undeniably unfair against certain groups of people. However, it is possible for an item to be unbiased but be unfair at the same time. To clarify, an unfair item is an item that discriminates against a certain groups of examinees whether this affects their responding or not (Hambleton & Rodgers, 1995). However, an item is biased when "examinees from one group are less likely to answer an item correctly than examinees of another group because of characteristics of the test item or the measurement situation that are not relevant to the testing situation" (Slocum, Gelin, & Zumbo, 2003, p3). For example, the item in figure 8.1 below was part of the 3rd edition of Stanford Binet Intelligence Test (Terman & Merrill, 1937) and asks: which one of the two women is more attractive? Equally knowledgeable test takers might have equal chance of getting this item right. Therefore the item is not necessarily biased; rather it is more likely to be offensive or unfair than biased.

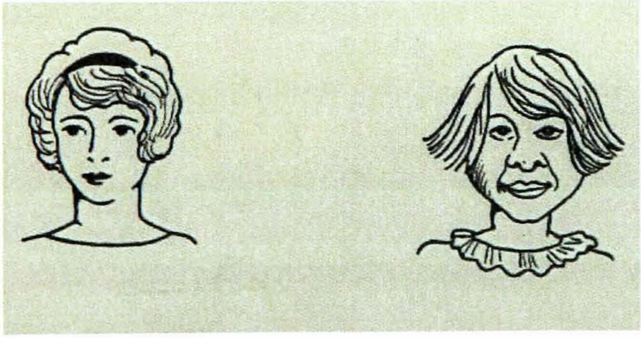


Figure 8.1: Item from the 3rd edition of Stanford Binet Intelligence Test (1937)

8.3. Item bias and item impact

Item bias is usually marked by differences in performance on a specific question by equally knowledgeable individuals from different groups of interest (Hambleton & Rodgers, 1995). In ability testing, equal knowledge refers to matching test takers on their *total score* on the test (Solcum, Gelin, & Zumbo 2003). In personality testing this could be the participants' score on the scale or the construct being measured. For example, females with a specific cognitive ability should be as likely as males with the same level of cognitive ability to answer an item correctly. Should the probability of getting the item correctly be significantly different between these two groups, the item should be inspected for potential item bias. In terms of personality assessment, the matching could be done on the probability of a certain group to endorse an item or not. As a result, participants could be matched on their scale score, which is computed as the product of endorsing a set of items that make up the scale.

Item bias exists when the differences in performance between groups result from a characteristic unrelated to the latent variable being measured, such as age, familiarity with item format, mistranslation or many other reasons discussed in chapter 4. Item bias is therefore an anomaly at item level that affects the fairness of the inferences drawn from psychometric tests (Hambleton & Rodgers, 1995). When an item is identified as biased, it is unfairly discriminating against a certain group of individuals while advantaging another group.

When an item is biased, it functions differently between the two groups of interest, which results in differences in their mean score on the item. Differential item functioning (DIF) is a statistical term used to describe the existence of a discrepancy in performance between two groups (Slocum, Gelin,

& Zumbo, 2003), and it "is necessary, but not sufficient, for item bias" (Slocum, Gelin & Zumbo, 2003, p3). An observed discrepancy in performance between groups on the variable of interest is not always a reflection of item bias as it could be the result of real differences between them (Zumbo, 2006). Therefore assuming that an item is biased based on mean group differences alone is false and insufficient (Kline, 1993).

When DIF is detected, items need to be scrutinised in order to unveil the origins of the difference in performance between groups (SIOP, 2003; Zumbo, 1999 and 2006; Slocum, Gelin & Zumbo, 2003). So is it caused by characteristics of the item itself or by existing and real differences between groups? When the observed difference on an item is the result of real differences between the groups, it is referred to as *item impact* (Zumbo, 2006). Otherwise, the difference represents and measurement artefact and is referred to as *item bias* (Zumbo, 1999). For example, analysis of BarOn Eqi emotional intelligence questionnaire revealed that women tend to score higher on most items assessing empathy than men (BarOn, 2002). This is a case of *item impact* whereby the difference detected in the mean score of males and females on items measuring empathy is a real one whether it is due to genetic predispositions or to culturally acceptable norms. To conclude, item bias is a case of unfairness towards a certain group of assesses whereas item impact is a fair and realistic difference between two groups of interest.

8.3.1 Uniform and non-uniform bias

van de Vijver and Leung (1997) distinguish between two forms of bias: uniform and non-uniform. Uniform bias affects scores consistently in one group of participants, whereas non-uniform bias affects these scores inconsistently. In cross cultural research, the main groups of interest for comparison are culture groups. When uniform bias occurs in this case, it implies that one of the groups is consistently endorsing extreme items or in other words performing consistently better or consistently worse than the other groups on a specific item (van de Vijver & Leung, 1997). In contrast, non-uniformly biased items indicate that in a certain culture A, participants with higher score levels are likely to perform differently, say better, than cultures B and C whereas participants with lower score levels are likely to perform in the opposite direction, in this case lower, than cultures B and C. Therefore differences between groups of interest are inconsistent across score levels, whether these are total scores (i.e. ability tests) or scale score (i.e. personality) (Mungas et al, 2000).

van de Vijver and Leung (1997) illustrate the distinction between uniform and non-uniform bias using the weighing scale as an example. If a scale adds 1Kg to every measure, all the measures will be biased but consistently since the error is always equal to one. So a person that weighs 60 Kg will come out as weighing 61Kg on this scale, and a person weighing 80 Kg will weigh 81Kg on this scale. However, if each 1Kg is being wrongly measured as 1.1Kg, then a person weighing 60 Kg will be 66kg and the person weighing 80 Kg will come out as 88Kg. In uniform bias, all the scores have a consistent error of 1Kg, whereas in the case of non-uniform bias one score has an error of 6 Kg while the other one has an error of 8Kg.

To clarify the concept of non-uniform bias further, let us consider this hypothetical example. Say a researcher was interested in examining the relationship between strength of family ties and parents' perception of their children's happiness among Lebanese and British parents. Considering that strong family ties are highly correlated with view of happiness among Lebanese nationals, parents with close relationships to their children would therefore be likely to rate them as very happy. In contrast, parents without close relationships with their children would be more likely to rate them as unhappy. However, this will not be apparent in the British sample if the same correlation did not exist between family ties and happiness. This could therefore be interpreted as a non-uniform bias where strong family ties are associated with higher happiness and weak family ties are associated with low happiness. Non-uniform bias is rarely reported in the literature and is much less likely to occur than uniform bias (Van de Vijver & Leung, 1997).

In conclusion, when items are flagged as functioning differentially between two groups, it is essential to investigate the nature of these differences qualitatively to distinguish between item bias and item impact (Zumbo, 1999). However, if the item is judged to be biased, understanding the uniformity of the bias can help decipher some of the group differences that might have lead to this bias.

8.3.2 Statistical methods for assessing DIF

Several statistical techniques, more recently referred to as differential item functioning (DIF) analysis, have been designed to detect item bias in cross-cultural research such as ANOVA, Mantel Haenszel statistic (MH), Item Response Theory (IRT), logistic regression (LogR), Simultaneous Item Bias Test (SIBTEST), log-linear, logistic regression, and ordinal logistic regression (Le, 2006; Zumbo, 2005; Griel, Jodoin & Ackerman, 2000; Zumbo, 1999; Van de

Vijver & Leung 1997; Swaminathan & Rogers, 1990). However, these analyses fall short in that they point to differences in performance without distinguishing between item impact and item bias. Although little emphasis has been put on the development of techniques that distinguish between these two items, many rely on the use of sound qualitative in-depth techniques such as cognitive interviewing previously discussed in chapter 7.

As with most other statistical procedures, some of the statistical techniques aforementioned tend to work better with parametric data and others with nonparametric data. Ability tests tend be dichotomously scored as a wrong or right (nominal or categorical data) and should be analysed using non-parametric tests. Normative personality tests, such as Orpheus, are usually polytomously scored on a Likert scale (ordinal data) and should also be analysed with non-parametric tests. Yet, many academic journal articles treat this type of data as interval and use parametric tests to analyse it (Fife-Shaw, 2006).

Most DIF analysis techniques developed are for dichotomously scored items (Zumbo, 1999). Van de Vijver and Leung (1997; 2005) and Zumbo (1999; 2005) provide a detailed description of running several DIF analyses statistically using SPSS and will be used as the main references for the following sections.

8.3.2.1. DIF for dichotomously scored data

Historically, DIF analysis gained attention in aptitude, achievement, certification and licensing tests for analysing bias against minority groups (Wendt & Worcester, 2000; Zumbo, 2006). These types of tests fall under the umbrella of ability testing and are scored dichotomously. It is therefore not surprising that most DIF analysis methods are designed for dichotomously scored tests.

Mantel Haenszel (MH) is the most popular DIF method used for analysing bias in binary data and produces powerful statistics (Van de Vijver & Leung, 1997; Sireci, Patsula, & Hambleton, 2005, in Hambleton, Merenda, & Spielberger, 2005). MH technique, for example, was assigned by the National Council of State Boards of Nursing as the official method for identifying DIF in licensure examination for nurses in the US (Wendt & Worcester, 2000). Although popular, MH technique suffers from several limitations highlighted by van de Vijver and Leung (1997) and these are as follows:

1) MH only applies to dichotomous data
2) It does not allow for detection of non-uniform bias and
3) It only produces pairwise comparison and does not allow for comparisons of more than two groups.

van de Vijver and Leung (1997) suggest log-linear as an alternative DIF method for dichotomous data. They argue that it outweighs MH since it allows for the detection of non-uniform bias and also for comparison between more than two groups. Nonetheless, log-linear can only be applied to dichotomous data so it cannot be used to detect DIF in personality questionnaires.

Zumbo (1999) argues that the most recommended and effective method for detecting DIF in dichotomous data is logistic regression (LogR). LogR outweighs MH and log-linear analyses because it can detect uniform and non-uniform bias, it allows for comparing more than two groups but most importantly it can be used on both dichotomous and polytomous data sets (binary logistic regression and ordinal logistic regression consecutively). Moreover, a study by Gierl and Jodoin (2001) comparing the use of MH, LogR and SIBTEST for analysing DIF, found that both MH and LogR are as powerful in correctly detecting uniform DIF. Therefore, LogR is not necessarily more powerful than MH in detecting DIF, it only has a wider scope in terms of allowing for the detection of non-uniform bias and the analysis of ordinal data.

The main challenge associated with LogR is the inflated type I error (Jodoin & Gierl, 2001). However, the Zumbo-Thomas effect size was developed by Zumbo and Thomas (1997 in Zumbo, 1999) in order to provide a measure of the magnitude of bias in LogR. This measure was created in order to increase the accuracy of hypothesis testing and reduces Type I error in LogR (Gierl, Jodoin & Ackerman, 2000). Using the Zumbo-Thomas effect size should decrease the probability of flagging items as DIF when they actually are not. Jodoin and Gierl's study (2001) revealed that when flagging DIF items in LogR using the p value of the 2 degrees of freedom chi square ($\Delta\chi^2$), the type I errors increased as sample size increased. However, when Zumbo-Thomas effect size was used as the criterion for flagging DIF, they observed a decrease in type I error as the sample size increased, regardless of how large the proportion of DIF items was. Conversely, when the difference in sample sizes under comparison were very large (1000 vs 250 participants), the results were inconsistent. These findings suggest that when using LogR for flagging DIF, relatively equal sample sizes should be used in addition to Zumbo-Thomas effect size (Jodoin & Gierl, 2001).

8.3.2.2. DIF for polytomously scored data

With the scarcity of DIF techniques for polytomous data, ANOVA is typically employed for this type of analysis (van de Vijver & Leung, 1997). Zumbo (1999) proposed ordinal LogR as an extension of binary logistic regression for detecting DIF in polytomously scored data. We will first begin by exploring ANOVA as a DIF technique before moving to Ordinal LogR, which builds up on concepts that we will discuss in ANOVA.

8.3.2.3. ANOVA

ANOVA can be considered as a widely used DIF method for identifying uniform and non-uniform bias in polytomously scored data (van de Vijver & Leung, 1997). ANOVA uses 3 main variables in the analysis: item score, total score and groups. The item score is the score on the item being analysed and is entered as the dependent variable (DV); the total score is total score on the test or scale and is entered as the first independent variable (IV) in the analysis; and the group, which could be the culture, gender, ethnicity or any other group of interest, is the second independent variable (IV) in the analysis (van de Vijver & Leung, 1997).

As for the interpretation of results, a main effect of culture is interpreted as a uniform bias where people from a certain group or from a certain total score are consistently performing better or worse than other groups. However, the effect of score group is expected to be significant since low scorers by default score differently than high scorers, and is therefore overlooked in item bias analysis (Byrne & Watkins, 2003). A statistically significant interaction between culture and total score is a reflection of non-uniform bias. It is recommended to use Bonferroni adjustment with ANOVA in order to control for the increase in type I error through multiple comparisons (Lee, Falbo, Doh, & Park 2001). In a study examining the identity of Koreans living in China and the US through a questionnaire, ANOVA was used to analyse DIF and 3 items were shown to have a significant interaction (Lee, Falbo, Doh, & Park, 2001). However, after adjusting for Type I error using Bonferroni correction procedure, none of these items showed DIF.

8.3.2.4. Ordinal Logistic Regression

As discussed earlier, Logistic regression is the most recommended method of DIF analysis for binary data (Zumbo, 1999). Ordinal logistic regression is an extension of binary logistic regression and follows the same logic as ANOVA but is more hierarchical by nature. The item score is always the DV but in hierarchical analyses such as regression it is important to specify the order for entering the IVs. The total score is entered first in the LogR analysis (Zumbo, 1999). Naturally, candidates with a higher total score will perform differently than those with lower ones. So by entering the total score as the first step, the amount of the variance explained by the total score will be removed from the equation. The group is entered in the second step to consider uniform bias, and the interaction between group and total score is entered in the last step to test for non-uniform bias (Zumbo, 1999). As discussed earlier with binary logistic regression, Zumbo-Thomas effect size is used as the criterion for flagging DIF items.

8.3.2.5. Comparing ANOVA and LogR

ANOVA and ordinal LogR both have the advantage of flagging the two types of DIF and can be applied to polytomous data. However, one conceptual difference between them is that ANOVA treats Likert scales as interval, whereas LogR rightly treats them as ordinal. This is a controversial issue because Likert type scales are widely treated as interval, although they are not, and therefore their mean is not a suitable measure of central tendency (Fife-Schaw, 2006). For the purpose of this study, we will employ ordinal LogR as the main method for detecting DIF. Nonetheless, we will also provide a comparison of LogR with ANOVA in order to provide an empirical investigation of the statistical power of these two methods in accurately identifying DIF items.

8.3.2.6. Matching and purification

DIF analysis requires matching "equally knowledgeable persons" from each group in order to investigate whether they have equal chance of endorsing the item (Zumbo, 2005). This is done by computing their total score of the test, which could be different from one test to the other depending on the length of the test. van de Vijver and Leung (1997) recommend having at least 50 participants in each score group. That is, if a test comprises of 10 questions and is scored on a 1 to 4 Likert scale, the lowest score possible is 10x1=10 and the largest score is 10x4= 40. Scores of 10 and 40 represent floor and ceiling effect respectively and should therefore be removed from the analysis (van de Vijver & Leung, 1997). Therefore, this test has 39 possible score groups. However, it is unlikely to have at least 50 participants in each of these score groups leading to total of 1950 participants. Therefore these could be grouped into a smaller manageable number of groups[1].

Gierl, Jodoin & Ackerman, (2000) point out that the number of items

[1] See van de Vijver and Leung, 1997 for a detailed description of this procedure

flagged as DIF in cross-cultural studies is usually quite substantial compared to the total number of items in the test. They argue that these DIF items will undeniably contaminate the matching procedure if they are included in the calculation. As an alternative, they suggest a “purification procedure”, originally suggested by Lord (1980 in Gierl, Jodoin, & Ackerman, 2000), which consists of two steps. Initially, all items are included in the analysis to flag DIF items. In the second step however, these items are removed from the test, and the total score is calculated based on non-DIF items only. Nevertheless, it is recommended that the particular item under investigation for DIF should be included in the calculation of the total score (Holland & Thayer 1988 in Zumbo, 1999).

It is arguable that the purification technique does not have an effect on DIF detection when the proportion of DIF items was small (Miller & Oshima. 1992, in Gierl, Jodoin, & Ackerman, 2000). However, when the proportion of DIF items was large (20 to 40%), the purification resulted in MH DIF analysis to be more accurate in flagging DIF.

8.4. Methods

8.4.1 Participants

Participants in this study (N=815; Arab world n=198, China n=222, Spain n=191, UK n=204) were sampled using a snowballing technique where an initial sample of convenience was contacted through email and individuals were asked to forward the invitation email and also though social networking websites (described in the section 8.4.3 below). Age information was collected using age groups: 18-25; 26-30; 31-35; 36-40; 41-45; 46-50; 51-55; 56-60; 61-65; 66 and above. Participants across the four cultures were predominantly between 18 and 35. Gender ratio is approximately equal in all groups as summarised in table 8.1 below.

		% from the UK	% from the Arab world	% from China	% from Spain
n		204	198	222	191
Age bands					
	18-25	12.7	29.3	73.9	38.4
	26-30	17.6	49.5	10.8	30.5
	31-35	19.1	10.1	5.0	17.4
	36-40	9.3	3.5	1.8	5.3
	41-45	9.3	1.5	0	2.1
	46-50	6.9	1.0	0	2.1
	56-60	7.4	.5	0.5	.5
	61-65	4.4	4.5	0	.5
	999	3.4	29.3	8.1	3.2
Gender					
	% Male	51.5	52.5	51.8	52.9
	% Female	48.5	47.5	46.4	47.1

Table 8.1: Gender and age percentage across cultures

All the samples had a relatively equal number of males and females filling out the questionnaires as illustrated in figure 8.2 below.

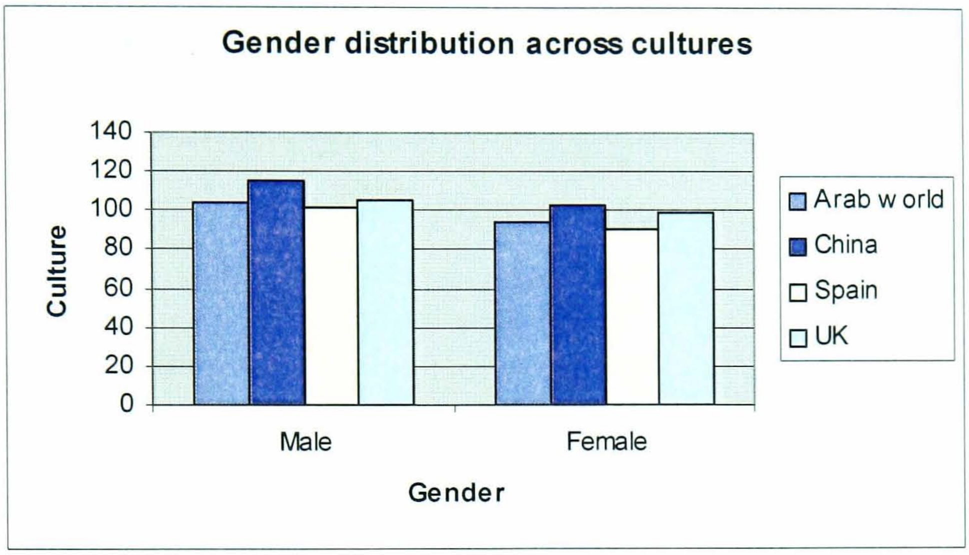


Figure 8.2: Gender distribution across cultures

8.4.2 . Materials

Four multi-lingual electronic versions (V5) of the Orpheus questionnaire in: Arabic, Chinese, English and Spanish comprising of 190 items each (appendix 35, 36, and 37).

8.4.3 Procedure

The electronic versions of Orpheus were circulated with an introductory email explaining confidentiality issues and the purpose of the research (appendix 27) through snowballing technique. Each participant took the test in his or her native language. Additionally, the link to the questionnaire with a brief about the study was posted on social media networks such as Facebook and MySpace to attract more participants since there was no funding available for incentives for participants. However, in China data was partly collected in paper and pencil format. The instructions clearly states that there was no time limit associated with the test and encouraged participants to answer as honestly as possible because the questionnaire contains honesty check. After completing the test, participants received a thank you email with a feedback report describing their personality preferences at work with an opportunity for further feedback.

8.4.4 Analysis

8.4.4.1. Reliability analysis

Scale reliability analysis was computed for each scale and for each culture separately. The analysis was conducted on the raw data to be consistent with the measurement invariance analysis that will follow. Reliability coefficient **α**, item facility and item discrimination are reported in tables 8.4, 8.5, 8.6, 8.7, and 8.8 below.

8.4.4.2. DIF Analysis

Ordinal Logistic Regression was used as the main method for analysing DIF, but ANOVA was also conducted in order to compare the scores that these two DIF analysis techniques provide. For both analyses Total Scores and Score Groups were calculated as outlined in the section below.

Calculating the score groups

A total score for every scale was calculated by adding all the positive items to all the reversed negative items and labelled TSF (for fellowship), TSA (for agreeableness) etc.

The total scores were then grouped into 8 groups comprising of approximately 50 participants each following the procedure suggested by van de Vijver and Leung (1997). This entails combining the total scores (TSF for fellowship, TSA for authority etc.) of UK data with the total scores of the target culture data separately for each scale, in order to determine the cut off points using the frequency option in SPSS. These cut off points were used to group the scores into 8 groups with a relatively equal number of participants in each. The new variables were called score group SGFArabic (fellowship Arabic) SGFChinese (fellowship Chinese) SGFSpanish (fellowship Spanish) SGAArabic (for authority Arabic) and so on. As an example, the cut of points for Arabic and English data are presented in table 8.2 below.

		TSA	TSC	TSE	TSD	TSF
N	Valid	400	397	401	399	397
	Missing	2	5	1	3	5
Percentiles	12.5	37.00	48.00	38.00	44.00	50.00
	25	39.00	51.00	41.00	48.00	53.00
	37.5	41.00	53.00	43.00	51.00	55.00
	50	43.00	55.00	45.00	53.00	56.00
	62.5	45.00	57.00	47.00	56.00	58.00
	75	47.00	59.00	49.00	58.00	59.00
	87.5	50.00	60.00	52.00	61.00	62.00

Table 8.2: cut off points for Arabic and English data combined.

Therefore, Total Score between 32 and 37 on Fellowship (TSF) were recoded as Score Group 1 (SGFArabic), scores between 38 and 39 were recoded as SGFArabic 2 and so on.

Ceiling and floor scores were also calculated by multiplying the number of items in the scale (for example 22 for Fellowship) by the lowest option possible in the Likert scale (1) to get the floor score (22) and by the highest option possible in the Likert scale (4) (van de Vijver & Leung, 1997) to get the ceiling score (88) as shown in the first two columns of table 8.3 below. As no ceiling or floor scores were observed, no participant was disregarded form the analysis.

Scale	Minimum score	Maximum score	Number of possible scores	Number of possible scores without ceiling and floor
F				
A				
C			**Copyrighted information**	
E				
D				

Table 8.3: Ceiling, floor and total scores for the 6 scales

Logistic Regression

As suggested by Zumbo (1999), Score Group was entered in the first level of Ordinal Logistic Regression, Language was entered in the second step and the interaction between Total Score and Culture was entered in the last step. The two-degrees-of-freedom Chi-Square for detecting DIF was calculated by deducting the $\Delta\chi^2$ Chi-square of step 3 χ^2 (3) from Chi-squared of step 1 χ^2 (1) as follows:

$$\Delta\chi^2 = \chi^2(2) = \text{Step } 3\ \chi^2(3) - \text{ Step } 1\ \chi^2(1)$$

DIF items were flagged based on the two-degree-of-freedom Chi-square having p value less or equal to 0.01 **and** a Zumbo-Thomas effect size larger than 0.130.

Neglecting to examine the effect size can lead to trivial effects being statistically significant especially in large sample sizes (Zumbo, 1999). The Zumbo-Thomas effect size is a weighted least squares effect size measure for LogR calculated as follows:

$$\Delta R^2 \text{ (Nigelkerke)} = \text{Step 3 } R^2 \text{ - Step 1 } R^2$$

As discussed earlier, flagging an item as DIF results in a simultaneous test of uniform and non-uniform bias (Swaminathan & Rogers, 1990 in Zumbo, 1999). Therefore, further examination of the difference between R^2 from steps 2 and 3 is necessary for determining whether the DIF is uniform or non-uniform (Zumbo, 1999). These will be presented in the results section below.

ANOVA

Analyses of variance (ANOVA) was carried out with the items as dependent variable and Language (two levels: English and Target language) and Score Groups (eight levels) as independent variables. The analysis for each scale in each culture was performed separately. The results are presented for all the cultures but for each scale independently in tables 8.4, 8.5, 8.6, 8.7, and 8.8 below.

8.5. Results

8.5.1 Reliability Analysis

Table 8.4 below presents a summary of the scale means and standard deviations across the four cultures. Then for each scale, the results internal consistency, item facility and item discrimination are summarized for the four cultures. This will be followed up with tables to present the full result for each scale across the four cultures.

8.5.1.1. Fellowship

Copyrighted information

8.5.1.2. Authority

Copyrighted information

8.5.1.3. Conformity

Copyrighted information

8.5.1.4. Emotion

Copyrighted information

8.5.1.5. Detail

Copyrighted information

		Alpha	Mean	Std. Deviation	N of Items
Authority	Arab world	0.62	43.45	5.078	
	China	0.44	42.41	4.021	
	Spain	0.63	41.71	5.174	
	UK	0.82	43.56	6.992	
Fellowship	Arab world	0.51	56.35	4.759	
	China	0.56	55.70	4.833	
	Spain	0.64	56.43	5.860	
	UK	0.64	55.49	5.563	
Conformity	Arab world	0.47	55.68	4.725	
	China	0.42	55.87	4.255	
	Spain	0.46	55.89	4.663	
	UK	0.67	54.15	5.429	
Emotion	Arab world	0.76	45.34	6.001	
	China	0.72	45.58	5.121	
	Spain	0.74	45.93	5.898	
	UK	0.81	44.74	6.207	
Detail	Arab world	0.72	56.46	5.767	
	China	0.59	52.85	4.684	
	Spain	0.62	53.19	5.217	
	UK	0.85	49.32	7.768	

Table 8.4: Scale means and Standard Deviations for the 4 cultures

Culture (**Coefficient α**)	**Arab World** (0.51)		**China** (0.56)		**Spain** (0.64)		**UK** (0.64)	
Item	**Facility**	**Discrimination**	**Facility**	**Discrimination**	**Facility**	**Discrimination**	**Facility**	**Discrimination**
Copyrighted information	3.41	-.142	3.26	**.078**	3.21	-.087	2.82	-.210
	2.71	.328	2.45	**.190**	2.62	.373	2.30	**.288**
	3.10	**.094**	2.57	**.217**	2.78	**.143**	2.84	**.051**
	2.91	.377	2.97	**.270**	3.10	.330	2.69	**.252**
	2.52	**.113**	2.58	**.231**	2.66	**.098**	2.60	**.046**
	2.94	**.137**	2.72	.329	2.69	.362	2.29	**.217**
	2.86	**.221**	2.21	.368	2.52	.443	2.52	.366
	3.10	**.217**	2.80	**.231**	2.48	**.140**	2.65	**.190**
	2.80	**.260**	2.59	**.118**	2.68	.359	2.57	**.299**
	2.78	**.299**	2.83	.313	2.66	.455	2.48	.375
	2.39	**.034**	2.20	**.019**	2.21	-.004	2.27	**.096**
	2.16	**.140**	2.62	**.013**	2.37	.300	2.19	.351
	2.09	**.026**	1.58	**.065**	2.49	.332	2.59	**.095**
	1.88	**.124**	1.97	-.140	2.29	**.132**	2.54	**.274**
	1.97	**.079**	2.19	**.176**	2.28	**.158**	2.27	.309
	2.37	**.214**	2.52	.312	2.82	.342	2.60	.366
	1.62	**.001**	2.35	.188	2.15	**.158**	2.23	**.043**
	2.10	**.217**	2.90	**.233**	2.25	.405	2.46	.420
	3.03	**.223**	2.56	**.138**	2.52	**.099**	2.72	**.089**
	2.25	-.053	2.54	**.207**	2.27	**.002**	2.44	.327
	2.72	**.191**	2.67	**.283**	2.55	**.185**	2.65	.347
	2.79	**.182**	2.58	**.041**	2.84	**.065**	2.57	**.246**

Table 8.5: Comparing item facility for Fellowship scale across cultures

Discrimination below 0.3; negative discrimination

Culture (Coefficient α)	Arab World (0.62)		China (0.44)		Spain (0.63)		UK (0.84)	
Item	**Facility**	**Discrimination**	**Facility**	**Discrimination**	**Facility**	**Discrimination**	**Facility**	**Discrimination**
Copyrighted information	2.22	.245	2.03	.088	2.57	.135	2.05	.396
	2.96	.199	3.05	.115	2.76	.155	2.54	.326
	3.05	.215	2.74	.292	2.59	.290	2.58	.517
	2.50	.160	2.90	.064	2.33	.187	2.37	.296
	2.55	.474	2.16	.240	1.77	.446	2.21	.664
	2.48	.493	2.45	.285	2.36	.310	2.54	.444
	2.68	.206	2.80	.057	2.61	.068	2.59	.486
	2.56	.485	2.30	.249	2.44	.385	2.45	.499
	2.41	.297	2.55	.268	2.83	.070	2.42	.599
	1.67	.183	2.30	.238	1.85	.341	2.27	.485
	2.92	.218	2.16	.210	2.61	.036	2.72	.496
	1.83	.034	2.28	.016	1.79	.268	2.21	.238
	2.24	.081	2.15	-.121	2.69	.117	2.67	.354
	2.41	.174	2.36	.078	2.21	.330	2.58	.484
	2.61	.306	2.06	.248	2.44	.292	2.79	.548
	2.24	.217	1.77	.117	1.80	.325	2.16	.441
	2.17	.055	2.25	-.002	2.18	.264	2.32	.396
	1.86	.013	2.07	.053	1.87	.127	2.21	.121

Table 8.6: comparing item facility and discrimination for Authority scale across cultures

Discrimination below 0.3; negative discrimination

Culture (**Coefficient α**)	**Arab World** (0.47)		**China** (0.42)		**Spain** (0.46)		**UK** (0.64)	
Item	**Facility**	**Discrimination**	**Facility**	**Discrimination**	**Facility**	**Discrimination**	**Facility**	**Discrimination**
Copyrighted information	2.21	**.224**	1.89	**.143**	1.82	.325	1.62	**.292**
	2.25	**.044**	2.60	**.081**	2.00	**.200**	2.12	.409
	2.88	**.299**	2.58	**.102**	2.65	**.073**	2.27	**.202**
	2.89	-.010	2.44	-.093	2.66	**.028**	2.32	-.016
	2.26	**.091**	2.70	**.185**	2.10	**.103**	2.08	**.118**
	2.11	**.275**	2.24	**.081**	2.30	**.057**	1.98	.363
	2.97	-.022	2.76	**.154**	2.83	-.125	2.60	**.239**
	2.13	**.133**	2.32	**.211**	2.37	**.032**	2.26	**.198**
	2.79	**.208**	2.61	**.167**	2.39	**.022**	2.33	.441
	2.53	**.107**	3.11	**.074**	2.15	**.120**	2.23	**.211**
	3.10	**.086**	2.81	**.119**	2.84	-.139	2.74	**.188**
	2.57	.356	2.79	**.164**	2.46	**.224**	2.22	**.254**
	2.24	-.105	1.86	**.034**	2.10	**.138**	2.35	**.096**
	1.83	**.176**	2.32	**.161**	2.05	.404	2.39	**.192**
	1.99	**.249**	2.03	**.121**	2.47	**.133**	2.45	**.188**
	1.74	**.199**	2.00	**.298**	2.15	.316	2.15	.370
	1.94	**.190**	1.88	**.195**	1.97	**.272**	2.31	**.140**
	1.81	-.040	1.85	-.098	2.01	**.171**	1.81	**.254**
	2.40	**.100**	2.16	**.082**	2.38	**.155**	2.41	**.149**
	2.18	**.214**	2.16	**.186**	2.65	**.036**	2.25	**.169**
	1.83	-.044	1.70	**.089**	1.92	**.156**	2.32	**.195**
	2.11	**.166**	2.40	**.035**	2.28	**.093**	2.12	**.134**
	2.41	.324	2.45	**.175**	2.94	**.207**	2.58	**.136**
	2.27	-.073	2.23	**.018**	2.39	**.004**	2.39	**.098**

Table 8.7: comparing item facility and discrimination for Conformity scale across cultures **Discrimination below 0.3**; negative discrimination

Coefficient α	Arab World (0.71)		China (0.72)		Spain (0.74)		UK (0.81)	
Item	Facility	Discrimination	Facility	Discrimination	Facility	Discrimination	Facility	Discrimination
Copyrighted information	2.93	.348	2.96	.222	2.41	.390	2.55	.259
	2.46	.400	2.35	.320	2.61	.359	2.48	.335
	3.28	.213	2.65	.144	2.82	.066	3.02	.171
	3.08	.443	2.83	.507	2.76	.296	2.47	.617
	2.67	.488	2.76	.393	2.79	.271	2.51	.593
	2.39	.480	2.81	.426	2.60	.502	2.41	.472
	2.47	.401	2.64	.280	2.71	.327	2.51	.375
	2.58	.367	2.50	.358	2.53	.342	2.40	.436
	2.91	.315	2.71	.348	2.72	.281	2.66	.404
	1.99	.396	2.50	.423	2.30	.361	2.44	.604
	2.22	-.014	2.30	.279	2.32	.199	2.34	.134
	2.58	.121	2.28	.138	2.19	.147	2.13	.155
	2.81	.410	2.68	.215	2.86	.451	3.02	.529
	2.01	.259	2.31	.352	2.37	.284	2.05	.362
	2.84	.408	2.49	.025	2.79	.492	2.74	.222
	1.88	.318	2.07	.359	2.25	.358	2.04	.465
	2.20	.232	2.45	.272	2.26	.289	2.42	.302
	1.92	.440	2.31	.387	2.64	.343	2.64	.476

Table 8.8: comparing item facility and discrimination for Emotion scale across cultures

Discrimination below 0.3; negative discrimination

Culture (**Coefficient α**)	**Arab World** (0.72)		**China** (0.58)		**Spain** (0.62)		**UK** (0.85)	
Item	**Facility**	**Discrimination**	**Facility**	**Discrimination**	**Facility**	**Discrimination**	**Facility**	**Discrimination**
Copyrighted information	3.41	.375	3.24	.310	3.21	**.224**	2.83	.539
	3.17	.383	2.79	.438	3.01	.455	2.54	.709
	3.31	**.269**	2.53	**.134**	3.09	.312	2.62	.436
	2.54	**.122**	2.75	**.130**	2.10	-.063	2.41	-.102
	2.89	**.271**	2.18	**.147**	2.65	.351	2.31	.567
	3.05	.472	2.45	.468	2.71	.423	2.57	.615
	2.87	.468	2.78	.354	2.22	.378	2.25	.517
	3.30	.350	3.00	**.035**	2.89	**.220**	2.66	.594
	3.06	.353	2.82	**.168**	2.88	**.297**	2.68	**.284**
	2.68	**.141**	3.04	**.261**	2.86	**.230**	2.45	.526
	2.44	.382	2.15	**.185**	2.32	**.149**	1.88	.500
	2.48	**.203**	2.80	.303	2.21	**.122**	2.08	.578
	3.02	**.286**	2.49	-.003	3.08	**.239**	2.87	**.216**
	2.31	.301	2.31	**.138**	2.15	**.133**	1.99	.402
	2.80	**.217**	2.70	**.110**	2.65	**.176**	2.71	.232
	2.37	**.294**	2.37	**.244**	2.53	**.184**	2.18	.550
	2.59	**.288**	2.72	**.081**	2.67	**.229**	2.54	.484
	2.88	.405	2.58	.377	2.66	.384	2.67	.492
	2.81	.328	2.55	.311	2.80	**.194**	2.89	.077
	2.08	-.052	2.61	-.239	2.47	-.211	2.17	.299

Table 8.9: comparing item facility and discrimination for Detail scale across cultures

Discrimination below 0.3; negative discrimination

8.5.2 Logistic Regression

The results of two-degrees of freedom Chi square $\Delta\chi^2$ (Step3-Step1) and the Zumbo Thomas effect size ΔR^2 flagged DIF for approximately 12% of Arabic items, 11% of Chinese items and 3% of Spanish items. Therefore, the purification technique was not conducted on this data because the proportion of items showing DIF is smaller than 20%, which is the smallest proportion that would affect DIF detection (Miller & Oshima. 1992, in Gierl, Jodoin, & Ackerman, 2000). This will be discussed further in the discussion.

For Arabic, 4 out of 22 items were flagged as DIF for Fellowship; 1 out of 18 items for Authority; 4 out of 24 items for Conformity; 3 out of 18 of Emotion items, but none of the items measuring Detail showed DIF. In the Chinese sample, 3 out of 22 items were flagged as DIF for Fellowship; 2 out of 18 items for Authority; 5 out of 24 items for Conformity; none of the Emotion items showed DIF, whereas one out of the 20 Detail items showed DIF. Finally, for the Spanish sample, 2 out of 18 items for Authority and 1out of 24 items for Conformity showed DIF. Fellowship, Emotion and Detail did not have any DIF items. The results $\Delta\chi^2$ and ΔR^2 are listed in tables 8.10, 8.11, 8.12, 8.13, and 8.14 with the DIF items highlighted in blue.

Items	Arab World (18% DIF)			China (14% DIF)			Spain (0% DIF)		
	$\Delta\chi^2$	p value	ΔR^2	$\Delta\chi^2$	p value	ΔR^2	$\Delta\chi^2$	p value	ΔR^2
Copyrighted information	**64.205**	**<0.001**	**0.169**	46.918	<0.001	0.125	27.77	<0.001	0.081
	26.022	<0.001	0.057	2.703	>0.2	0.007	14.423	<0.001	0.035
	11.302	<0.01	0.031	23.282	<0.001	0.059	2.414	>0.2	0.007
	10.639	<0.01	0.024	19.238	<0.001	0.044	34.91	<0.001	0.085
	4.58	<0.2	0.012	4.474	<0.2	0.011	0.464	>0.2	0.001
	63.511	**<0.001**	**0.146**	37.678	<0.001	0.082	22.961	<0.001	0.057
	14.423	<0.001	0.033	31.078	<0.001	0.066	2.433	>0.2	0.005
	31.844	<0.001	0.075	4.101	<0.2	0.009	7.834	<0.02	0.021
	4.795	<0.1	0.011	1.641	>0.2	0.004	0.773	>0.2	0.001
	13.034	<0.01	0.028	32.575	<0.001	0.069	3.472	<0.2	0.008
	1.094	>0.2	0.003	1.074	>0.2	0.003	1.878	>0.2	0.006
	7.201	<0.05	0.017	33.778	<0.001	0.077	1.959	>0.2	0.005
	35.628	<0.001	0.092	**148.954**	**<0.001**	**0.326**	8.211	<0.02	0.021
	101.563	**<0.001**	**0.242**	**99.153**	**<0.001**	**0.243**	18.142	<0.001	0.052
	22.572	<0.001	0.056	3.45	<0.2	0.008	2.566	>0.2	0.007
	15.685	<0.001	0.036	2.149	>0.2	0.004	6.071	<0.05	0.014
	83.08	**<0.001**	**0.214**	3.468	<0.2	0.009	3.689	<0.2	0.011
	41.061	<0.001	0.094	**64.691**	**<0.001**	**0.143**	17.863	<0.001	0.086
	17.263	<0.001	0.044	6.342	<0.05	0.016	9.688	<0.01	0.028
	16.225	<0.001	0.043	2.811	>0.2	0.007	13.166	<0.01	0.037
	0.134	>0.2	0	0.19	>0.2	0.001	11.367	<0.01	0.031
	6.664	<0.05	0.017	2.079	>0.2	0.006	11.4	<0.01	0.03

Table 8.10: Uniformly and non-uniformly biased items from the Fellowship scale using Ordinal Logistic Regression

Items	Arab World (6% DIF)			China (11% DIF)			Spain (11% DIF)		
	$\Delta\chi^2$	**p value**	ΔR^2	$\Delta\chi^2$	**p value**	ΔR^2	$\Delta\chi^2$	**p value**	ΔR^2
Copyrighted information	6.45	<0.05	0.015	1.512	>0.2	0.004	**54.078**	**<0.001**	**0.136**
	48.404	<0.001	0.118	**74.822**	**<0.001**	**0.174**	15.672	<0.001	0.042
	44.453	<0.001	0.099	11.794	<0.01	0.024	4.114	<0.2	0.01
	4.531	<0.2	0.011	82.511	<0.001	0.2	0.586	>0.2	0.002
	32.328	<0.001	0.056	2.476	>0.2	0.005	20.501	<0.001	0.038
	13.883	<0.001	0.028	1.637	>0.2	0.004	1.924	>0.2	0.005
	1.701	>0.2	0.004	20.624	<0.001	0.05	5.375	<0.1	0.015
	7.258	<0.05	0.014	0.857	>0.2	0.001	2.979	>0.2	0.007
	3.128	>0.2	0.007	11.69	<0.01	0.024	**61.791**	**<0.001**	**0.153**
	66.382	**<0.001**	**0.146**	4.955	<0.1	0.01	16.959	<0.001	0.036
	9.939	<0.01	0.022	53.807	<0.001	0.105	11.309	<0.01	0.028
	29.252	<0.001	0.075	2.929	>0.2	0.007	20.444	<0.001	0.051
	28.091	<0.001	0.068	48.018	<0.001	0.113	3.767	<0.2	0.01
	6.936	<0.05	0.017	8.048	<0.02	0.019	16.489	<0.001	0.038
	6.633	<0.05	0.014	**101.222**	**<0.001**	**0.182**	13.985	<0.001	0.031
	1.341	>0.2	0.003	24.241	<0.001	0.053	14.967	<0.001	0.033
	10.641	<0.01	0.028	3.864	<0.2	0.01	0.633	>0.2	0.001
	25.141	<0.001	0.066	4.303	<0.2	0.01	18.568	<0.001	0.05

Table 8.11: Uniformly and non-uniformly biased items (in bold) from the Authority scale using Ordinal Logistic Regression

Items	Arab World (17% DIF)			China (21% DIF)			Spain (4% DIF)		
	$\Delta\chi^2$	p value	ΔR^2	$\Delta\chi^2$	p value	ΔR^2	$\Delta\chi^2$	p value	ΔR^2
Copyrighted information	**65.105**	**<0.001**	**0.145**	10.889	<0.01	0.028	4.221	<0.2	0.012
	9.703	<0.01	0.023	32.261	<0.001	0.071	12.566	<0.01	0.031
	69.218	**<0.001**	**0.153**	10.902	<0.01	0.027	25.487	<0.001	0.07
	39.185	<0.001	0.099	1.108	>0.2	0.003	13.155	<0.01	0.038
	4.05	<0.2	0.011	**72.447**	**<0.001**	**0.166**	1.017	>0.2	0.003
	4.51	<0.2	0.01	15.233	<0.001	0.035	32.192	<0.001	0.083
	26.62	<0.001	0.071	2.851	>0.2	0.007	17.276	<0.001	0.054
	8.615	<0.02	0.022	2.296	>0.2	0.005	1.855	>0.2	0.005
	38.389	<0.001	0.081	12.612	<0.01	0.029	13.172	<0.01	0.035
	18.078	<0.001	0.047	**178.405**	**<0.001**	**0.361**	4.488	<0.2	0.013
	20.2	<0.001	0.051	0.564	>0.2	0.002	9.459	<0.01	0.028
	13.637	<0.001	0.03	**58.488**	**<0.001**	**0.131**	3.814	<0.2	0.01
	4.731	<0.1	0.013	**53.949**	**<0.001**	**0.136**	14.751	<0.001	0.042
	89.291	**<0.001**	**0.223**	4.571	<0.2	0.011	44.648	<0.001	0.118
	47.816	<0.001	0.115	44.62	<0.001	0.109	3.495	<0.2	0.009
	57.055	**<0.001**	**0.134**	21.241	<0.001	0.048	4.366	<0.2	0.01
	37.529	<0.001	0.096	**52.982**	**<0.001**	**0.131**	43.441	<0.001	0.116
	9.988	<0.01	0.028	10.324	<0.01	0.028	2.71	>0.2	0.007
	0.844	<0.02	0.002	21.692	<0.001	0.058	3.539	<0.2	0.011
	5.11	<0.1	0.013	5.471	<0.1	0.014	15.73	<0.001	0.043
	65.048	<0.001	0.17	117.414	<0.001	0.285	**48.206**	**<0.001**	**0.136**
	2.902	>0.2	0.007	11.502	<0.01	0.029	1.726	>0.2	0.005
	21.218	<0.001	0.05	9.091	<0.02	0.024	16.204	<0.001	0.044
	6.246	<0.05	0.019	10.29	<0.01	0.03	0.03	>0.2	0

Table 8.12: Uniformly and non-uniformly biased items (in bold) from the Conformity scale using Ordinal Logistic Regression

Items	**Arab World (12% DIF)**			**China (0% DIF)**			**Spain (0% DIF)**		
	$\Delta\chi^2$	**p value**	ΔR^2	$\Delta\chi^2$	**p value**	ΔR^2	$\Delta\chi^2$	**p value**	ΔR^2
Copyrighted information	26.28	<0.001	0.058	33.136	<0.001	0.076	8.019	<0.02	0.02
	2.846	>0.2	0.006	7.044	<0.05	0.015	2.564	>0.2	0.006
	21.544	<0.001	0.057	36.865	<0.001	0.091	9.325	<0.01	0.027
	93.337	**<0.001**	**0.157**	29.26	<0.001	0.047	20.774	<0.001	0.043
	4.297	<0.2	0.007	20.066	<0.001	0.035	20.723	<0.001	0.041
	2.074	>0.2	0.004	44.75	<0.001	0.086	4.763	<0.1	0.011
	1.875	>0.2	0.004	2.137	>0.2	0.005	4.638	<0.1	0.011
	7.164	<0.05	0.016	2.991	>0.2	0.007	2.581	>0.2	0.007
	16.963	<0.001	0.04	0.643	>0.2	0.001	0.028	>0.2	0
	58.874	<0.001	0.105	6.045	<0.05	0.011	19.959	<0.001	0.04
	4.312	>0.2	0.012	4.038	<0.2	0.01	1.686	>0.2	0.005
	57.409	**<0.001**	**0.148**	11.014	<0.001	0.031	1.126	>0.2	0.003
	12.875	<0.01	0.026	99.335	<0.001	0.081	10.294	<0.01	0.02
	0.773	>0.2	0.002	16.183	<0.001	0.036	15.061	<0.001	0.037
	7.522	<0.05	0.018	19.252	<0.001	0.05	12.792	<0.01	0.03
	13.491	<0.01	0.031	1.854	>0.2	0.005	3.912	<0.2	0.009
	13.495	<0.01	0.033	0.086	>0.2	0	11.874	<0.01	0.03
	122.903	**<0.001**	**0.231**	34.81	<0.001	0.069	2.245	>0.2	0.005

Table 8.13: Uniformly and non-uniformly biased items (in bold) from the Emotion scale using Ordinal Logistic Regression

Items	Arab World (0% DIF)			China (4% DIF)			Spain (0%DIF)		
	$\Delta\chi^2$	p value	ΔR^2	$\Delta\chi^2$	p value	ΔR^2	$\Delta\chi^2$	p value	ΔR^2
Copyrighted information	15.374	<0.001	0.029	20.399	<0.001	0.043	12.611	<0.01	0.029
	19.312	<0.001	0.029	0.316	>0.2	0	5.348	<0.1	0.009
	19.081	<0.001	0.037	11.439	<0.01	0.026	13.618	<0.01	0.029
	6.949	<0.05	0.019	22.86	<0.001	0.057	18.043	<0.001	0.052
	18.088	<0.001	0.033	26.387	<0.001	0.057	4.159	<0.2	0.008
	4.649	<0.1	0.008	23.594	<0.001	0.044	2.178	>0.2	0.005
	4.874	<0.1	0.008	24.872	<0.001	0.046	12.764	<0.01	0.028
	27.92	<0.001	0.048	32.041	<0.001	0.072	4.803	<0.1	0.011
	6.028	<0.05	0.013	0.57	>0.2	0.001	2.21	>0.2	0.006
	16.212	<0.001	0.037	58.574	<0.001	0.113	17.226	<0.001	0.038
	4.193	<0.2	0.008	8.107	<0.02	0.016	13.041	<0.01	0.028
	6.322	<0.05	0.013	54.123	<0.001	0.094	14.647	<0.001	0.032
	3.119	>0.2	0.008	40.842	<0.001	0.104	5.21	<0.1	0.015
	0.071	>0.2	0	16.85	<0.001	0.039	1.76	>0.2	0.005
	5.622	<0.1	0.014	4.439	<0.2	0.011	6.928	<0.05	0.019
	5.147	<0.1	0.011	2.993	>0.2	0.007	10.502	<0.01	0.024
	12.063	<0.01	0.029	6.095	<0.05	0.015	0.832	>0.2	0.002
	5.487	<0.1	0.011	21.479	<0.001	0.045	11.696	<0.01	0.025
	18.018	<0.001	0.046	48.135	<0.001	0.112	7.767	<0.05	0.021
	10.738	<0.01	0.03	**53.581**	**<0.001**	**0.135**	22.26	<0.001	0.063

Table 8.14: Uniformly and non-uniformly biased items (in bold) from the Detail scale using Ordinal Logistic Regression

8.5.3 ANOVA

The main effect of Language and the interaction between Language and Score Group are of main interest for this study (van de Vijver & Leung, 1997). However, a main effect of Score Group is expected in any case because participants from different score groups are expected to score significantly differently on the scale. Therefore, the effects of Language and the interaction between Language and Score Group are reported. Tables 8.15, 8.16, 8.17, 8.18, and 8.19 list all the items in every scale and highlights the biased items across cultures with their means and standard deviations. The purification technique was not implemented due the large number of items that showed DIF between cultures (>40%). This is discussed further in the discussion section.

8.5.3.1. Fellowship

Arabic

Copyrighted information

Chinese

Copyrighted information

Spain

Copyrighted information

8.5.3.2. Authority

Arabic

Copyrighted information

Chinese

Copyrighted information

Spain

Copyrighted information

8.5.3.3. Conformity

Arabic

Copyrighted information

Chinese

Copyrighted information

Spanish

Copyrighted information

8.5.3.4. Emotion

Arabic

Copyrighted information

Chinese

Copyrighted information

Spanish

Copyrighted information

8.5.3.5. Detail

Arabic

Copyrighted information

Chinese

Copyrighted information

Spanish

Copyrighted information

Items	UK		Arab World (86% DIF)		China (64% DIF)		Spain (55% DIF)	
	M	SD	M	SD	M	SD	M	SD
Copyrighted information	2.83	.758	3.41	.614	3.25	.552	3.20	.707
	2.30	.714	2.71	.760	2.44	.759	2.65	.777
	2.84	.708	3.09	.625	2.57	.700	2.80	.785
	2.69	.669	2.91	.741	2.96	.745	3.12	.724
	2.61	.724	2.52	.783	2.58	.680	2.69	.744
	2.29	.779	2.94	.711	2.72	.709	2.70	.741
	2.52	.780	2.85	.751	2.21	.668	2.54	.825
	2.65	.731	3.10	.733	2.81	.757	2.50	.864
	2.58	.763	2.80	.723	2.59	.665	2.68	.800
	2.48	.664	2.78	.672	2.82	.587	2.66	.714
	2.27	.725	2.38	.833	2.20	.699	2.19	.848
	2.19	.839	2.17	.822	2.63	.861	2.37	.910
	2.58	.873	2.09	.782	1.59	.616	2.49	.930
	2.55	.646	1.88	.697	1.97	.612	2.30	.727
	2.28	.828	1.98	.807	2.19	.725	2.26	.822
	2.60	.746	2.38	.804	2.51	.833	2.83	.806
	2.24	.712	1.63	.583	2.34	.731	2.15	.703
	2.47	.622	2.09	.731	2.90	.593	2.23	.764
	2.73	.688	3.03	.689	2.56	.727	2.53	.793
	2.45	.700	2.24	.771	2.54	.703	2.26	.817
	2.66	.743	2.72	.764	2.67	.690	2.54	.709
	2.57	.736	2.80	.744	2.58	.744	2.82	.803

Table 8.15: Uniformly and non-uniformly biased items from the Fellowship scale using ANOVA

Uniform bias, Non-Uniform bias, Uniform and Non-Uniform bias

Items	UK		Arab World (83% DIF)		China (56% DIF)		Spain (61% DIF)	
	M	SD	M	SD	M	SD	M	SD
Copyrighted information	2.05	.851	**2.23**	**.804**	2.03	.811	**2.54**	**.869**
	2.53	.716	**2.96**	**.631**	**3.06**	**.673**	**2.76**	**.822**
	2.59	.762	**3.05**	**.773**	**2.75**	**.747**	2.61	.688
	2.37	.610	**2.50**	**.836**	**2.90**	**.633**	2.35	.737
	2.21	.793	**2.55**	**.828**	2.17	.800	**1.76**	**.796**
	2.54	.759	2.48	.875	2.46	.776	2.37	.789
	2.59	.704	2.67	.846	**2.81**	**.655**	2.61	.806
	2.45	.787	**2.56**	**.799**	2.31	.691	2.44	.811
	2.42	.734	2.42	.667	**2.56**	**.727**	**2.84**	**.737**
	2.28	.841	**1.67**	**.745**	2.29	.686	**1.84**	**.760**
	2.72	.789	**2.92**	**.866**	**2.17**	**.677**	2.61	.780
	2.20	.790	**1.82**	**.745**	2.28	.769	**1.79**	**.782**
	2.67	.719	**2.24**	**.841**	**2.15**	**.769**	2.68	.686
	2.57	.721	**2.41**	**.784**	**2.36**	**.759**	**2.20**	**.745**
	2.79	.800	**2.62**	**.793**	**2.07**	**.616**	**2.43**	**.750**
	2.15	.810	2.24	.807	**1.77**	**.722**	**1.80**	**.878**
	2.32	.663	**2.18**	**.619**	2.25	.692	2.19	.700
	2.22	.721	**1.86**	**.736**	2.06	.764	**1.87**	**.807**

Table 8.16: Uniformly and non-uniformly biased items (in bold) from the Authority scale using ANOVA

Uniform bias, Non-Uniform bias, Uniform and Non-Uniform bia

Items	UK		Arab World (67% DIF)		China (67% DIF)		Spain (67% DIF)	
	M	SD	M	SD	M	SD	M	SD
Copyrighted information	1.62	.558	**2.21**	**.710**	**1.88**	**.627**	**1.84**	**.673**
	2.12	.733	2.26	.775	**2.59**	**.777**	**2.04**	**.767**
	2.27	.630	**2.89**	**.696**	**2.57**	**.756**	**2.63**	**.627**
	2.31	.792	**2.88**	**.889**	2.45	.832	**2.69**	**.823**
	2.07	.603	2.27	.721	**2.70**	**.700**	2.11	.763
	1.98	.740	2.10	.675	**2.23**	**.724**	**2.35**	**.654**
	2.60	.552	**2.97**	**.689**	2.75	.735	2.81	.629
	2.27	.662	**2.13**	**.729**	2.32	.755	2.37	.720
	2.33	.733	**2.79**	**.733**	**2.60**	**.722**	2.44	.765
	2.23	.606	**2.55**	**.649**	3.11	.529	**2.15**	**.614**
	2.74	.729	**3.10**	**.709**	2.81	.686	2.82	.708
	2.23	.753	**2.57**	**.702**	**2.78**	**.596**	**2.47**	**.766**
	2.35	.812	2.23	.870	**1.86**	**.622**	**2.13**	**.824**
	2.39	.709	**1.83**	**.513**	2.32	.705	**2.07**	**.678**
	2.45	.818	**2.00**	**.775**	**2.03**	**.685**	2.46	.850
	2.15	.754	**1.75**	**.639**	1.99	.586	2.14	.690
	2.31	.722	**1.93**	**.762**	**1.87**	**.661**	**1.97**	**.784**
	1.81	.680	1.80	.688	1.85	.627	2.01	.707
	2.40	.664	2.40	.807	**2.17**	**.615**	2.36	.698
	2.26	.720	2.18	.723	2.17	.655	**2.65**	**.837**
	2.32	.647	**1.82**	**.644**	**1.69**	**.535**	**1.92**	**.643**
	2.11	.683	2.11	.782	**2.39**	**.676**	2.27	.796
	2.58	.749	**2.41**	**.853**	2.45	.696	**2.92**	**.774**
	2.38	.595	**2.27**	**.594**	**2.22**	**.547**	2.39	.607

Table 8.17: Uniformly and non-uniformly biased items (in bold) from the Conformity scale using ANOVA

Uniform bias, Non-Uniform bias, **Uniform and Non-Uniform bias**

Items	UK		Arab World (67% DIF)		China (72% DIF)		Spain (61% DIF)	
	M	SD	M	SD	M	SD	M	SD
Copyrighted information	2.55	.805	**2.94**	**.781**	**2.96**	**.654**	2.41	.828
	2.48	.832	2.46	.901	**2.34**	**.802**	2.61	.881
	3.02	.594	**3.28**	**.671**	**2.64**	**.794**	**2.82**	**.747**
	2.47	.753	**3.08**	**.729**	**2.82**	**.727**	**2.76**	**.734**
	2.51	.882	**2.67**	**.869**	**2.77**	**.698**	2.79	.773
	2.41	.681	2.39	.751	**2.81**	**.659**	2.60	.649
	2.51	.746	2.47	.789	2.63	.672	**2.72**	**.728**
	2.40	.680	**2.58**	**.797**	2.50	.678	2.52	.648
	2.66	.676	**2.92**	**.682**	2.70	.580	2.73	.762
	2.44	.818	2.00	.796	2.49	.710	**2.30**	**.803**
	2.34	.707	2.22	.710	2.29	.691	2.32	.761
	2.13	.508	**2.58**	**.709**	**2.29**	**.577**	2.19	.604
	3.02	.725	**2.82**	**.755**	**2.68**	**.692**	**2.87**	**.851**
	2.05	.647	2.01	.743	**2.31**	**.670**	**2.36**	**.740**
	2.74	.722	2.84	.691	**2.50**	**.650**	2.78	.804
	2.04	.714	**1.88**	**.628**	2.07	.588	**2.25**	**.761**
	2.42	.659	**2.19**	**.673**	2.45	.655	**2.26**	**.722**
	2.64	.746	**1.92**	**.713**	**2.31**	**.723**	2.64	.767

Table 8.18: Uniformly and non-uniformly biased items (in bold) from the Emotion scale using ANOVA

Uniform bias, Non-Uniform bias, Uniform and Non-Uniform bias

Items	UK		Arab World (65% DIF)		China (75% DIF)		Spain (55% DIF)	
	M	SD	M	SD	M	SD	M	SD
Copyrighted information	2.83	.758	**3.41**	**.614**	**3.25**	**.552**	**3.20**	**.707**
	2.54	.916	**3.18**	**.708**	2.79	.838	3.01	.833
	2.62	.843	**3.31**	**.784**	**2.53**	**.845**	**3.12**	**.759**
	2.41	.761	2.55	.868	**2.74**	**.786**	**2.11**	**.842**
	2.31	.856	**2.89**	**.732**	**2.18**	**.675**	2.65	.792
	2.57	.885	3.05	.693	**2.45**	**.683**	2.71	.779
	2.25	.832	**2.87**	**.806**	**2.78**	**.693**	**2.24**	**.864**
	2.66	.714	**3.29**	**.616**	**3.01**	**.662**	2.88	.640
	2.68	.672	**3.06**	**.570**	2.83	.648	2.88	.647
	2.45	.744	**2.69**	**.681**	**3.05**	**.585**	**2.84**	**.654**
	1.88	.853	2.45	.830	**2.15**	**.773**	**2.28**	**.910**
	2.08	.867	**2.49**	**.858**	**2.80**	**.657**	**2.18**	**.772**
	2.87	.664	3.02	.698	**2.47**	**.759**	3.09	.625
	1.99	.788	2.31	.790	**2.31**	**.600**	2.16	.758
	2.71	.702	2.80	.811	2.70	.769	**2.65**	**.738**
	2.18	.728	2.37	.814	2.36	.704	**2.53**	**.689**
	2.54	.671	**2.60**	**.713**	2.72	.619	2.67	.658
	2.67	.769	**2.87**	**.806**	**2.57**	**.625**	**2.67**	**.809**
	2.89	.721	**2.82**	**.768**	**2.57**	**.792**	**2.80**	**.778**
	2.17	.691	**2.07**	**.768**	**2.61**	**.627**	2.48	.807

Table 8.19: Uniformly and non-uniformly biased items (in bold) from the Detail scale using ANOVA

Uniform bias, Non-Uniform bias, Uniform and Non-Uniform bias

8.6. Discussion

8.6.1 Reliability analysis

Reliability coefficients for Fellowship and Conformity were below 0.7 for all the cultures including UK, indicating a low reliability for these two scales. The reliability coefficients for the UK sample for Authority, Emotion and Detail were all above 0.7. a closer investigation of the means and standards deviations scales in each culture shows that low reliabilities are clearly the result of small standard deviation relative to their means. For example, China showed the lowest reliability values across most of the scales, and as evident in table 8.4 the SD in the Chinese sample was consistently lower than in other cultures. This is not surprising considering the number of students who took part in this study, especially in China, and the fact that Orpheus is a work-based questionnaire, which might have items that are out of context for students. Moreover, this was mainly affected by the number of items with low discrimination index. For example, for Fellowship and Conformity, the percentages of English items with low discrimination were 64% and 79% respectively. On the other hand, the percentage of items with low discrimination on Authority, Emotion and Detail were 17%, 28%, and 30% respectively.

Copyrighted information

These findings reflect the importance of the simultaneous development of questionnaires across cultures. In the case of Orpheus, and in most cases in practice, questionnaires are developed in English and then adapted into other languages and cultures. Some items come out as functioning different across several languages, which facilitates the decision of dropping this item from the questionnaire. Additionally, it is important that items tap on the constructs of interest, but as discussed earlier, culture makes this task more challenging during the adaptation. In the case of Orpheus, there is an additional dimension that makes the wording of items more difficult to tap on the same constructs across all cultures and that is the fact that Orpheus is work-based rather than a generic personality questionnaire. Work ethics, values, and behaviours are different across cultures, which makes the design of cross-cultural work-based personality questionnaires more challenging than generic personality questionnaires.

8.6.2 DIF analysis

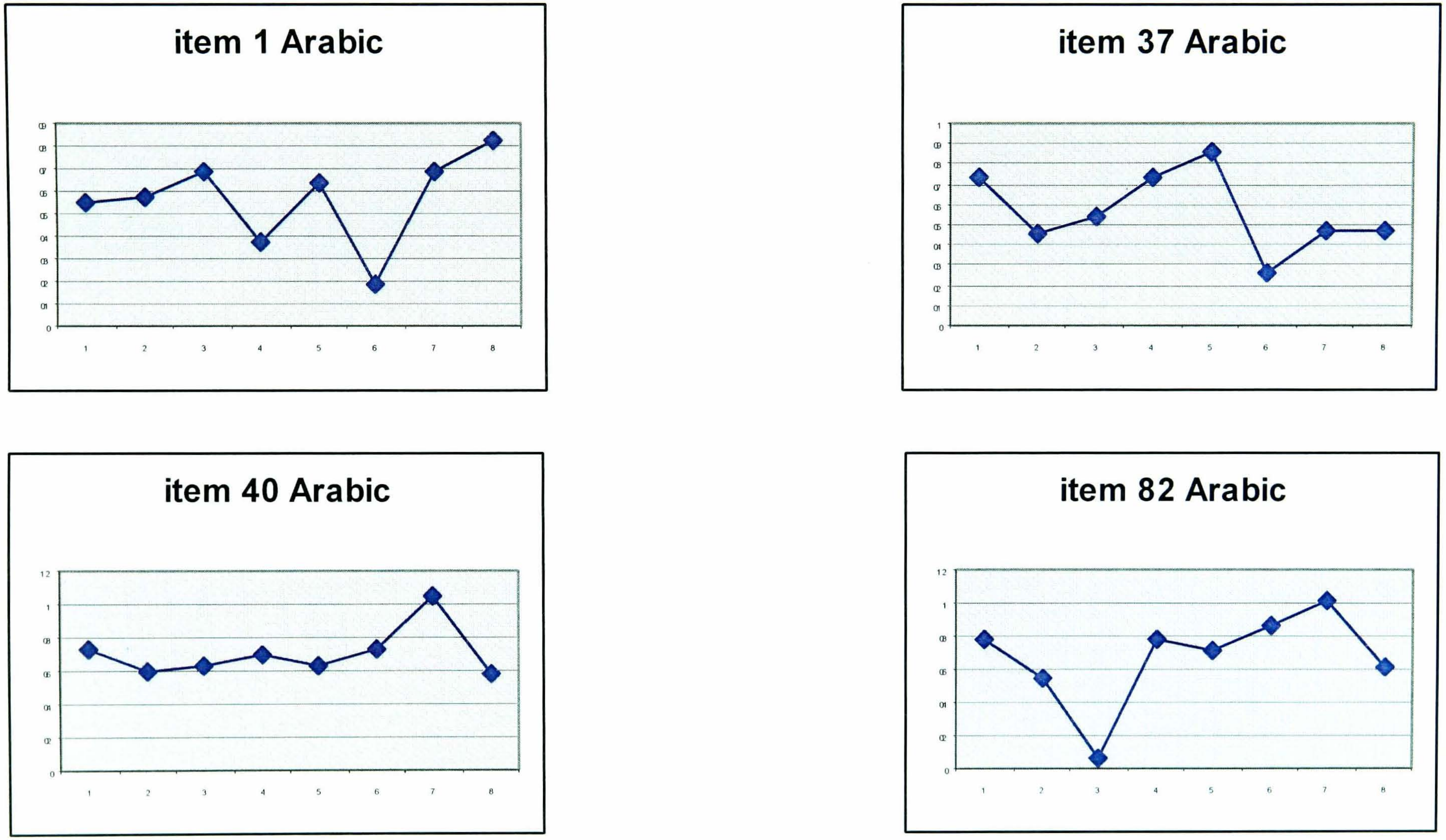
The results showed that, out of the 102 items of Orpheus that assess the Big Five model, 12 items were flagged as DIF in the Arab sample (12%), 11 from the Chinese sample (11%), and 3 from the Spanish sample(3%) (tables 8.9, 8.10, 8.11, 8.12, and 8.13). Only one of the items, item 40, was flagged as DIF across more than one culture (Arab world and China). Conformity scale had the highest number of DIF items across the three cultures (4 for the Arab world, 5 for China and 1 for Spain), followed by Fellowship (4 for the Arab world, 3 for China and 0 for Spain), Authority (1 for the Arab world, 2 for China and 2 for Spain), Emotion (3 for the Arab world, 0 for China and 0 for Spain), and finally Detail (0 for the Arab world, 1 for China and 0 for Spain).

The purification technique explained earlier in section 8.3.2.6 was not applied on this data because the percentage of items that showed DIF was less than 20%, whereas Miller and Oshima (1992, in Gierl, Jodoin, & Ackerman, 2000) suggest that purification should be applied when the proportion of DIF items is between 20 to 40%. When the proportion of DIF is smaller than 20%, the other 80% of the items can still produce a good estimate of the participants' score of the scale. However, if the proportion of DIF items is larger than 40%, then removing these items from the calculation of the total score might not give a good estimate of the participants' score on the scale, thus rendering the matching procedure useless.

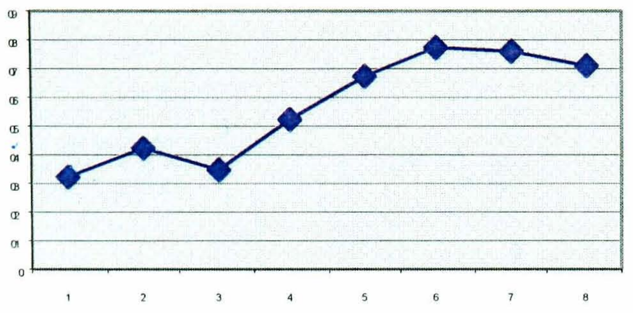
For the Arab sample, Item 1 was flagged as uniformly exhibiting DIF when it was assessed against the Score Group of Fellowship. The ΔR^2 between Step3 and Step 1 is equal to 0.169 whereas the ΔR^2 between Step 3 and Step 2 is equal to 0.002. Therefore, we can assume that DIF is uniform and that the item is behaving systematically differently between the UK and the Arab world (figure8.3).

Copyrighted information

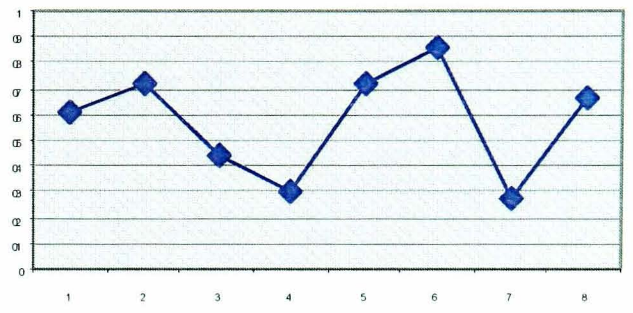
Figure 8.3: Uniformly biased item from in Arabic



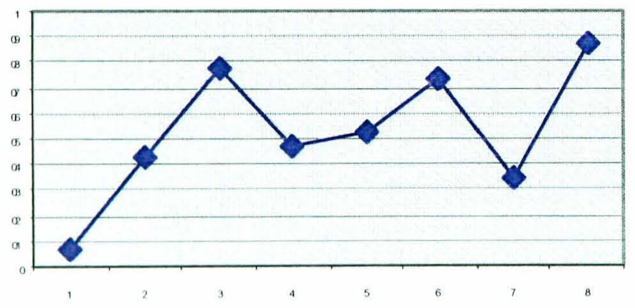
item 12 Arabic



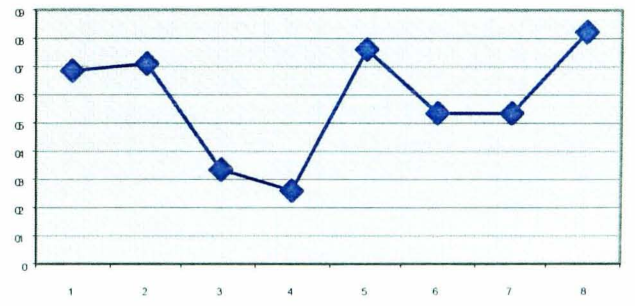
item 13 Arabic

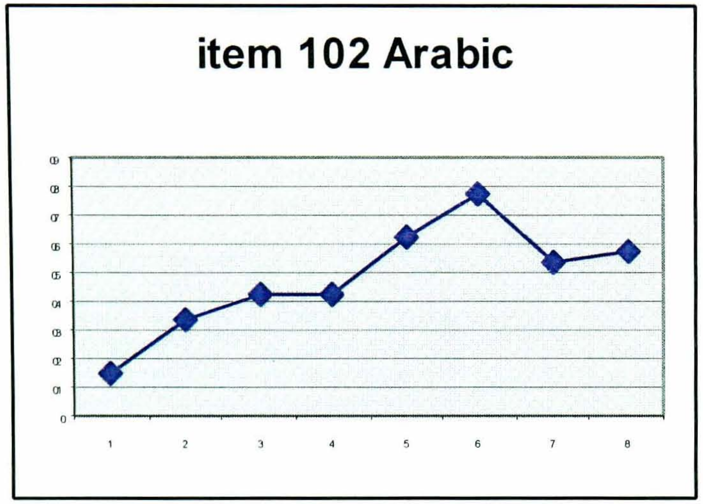


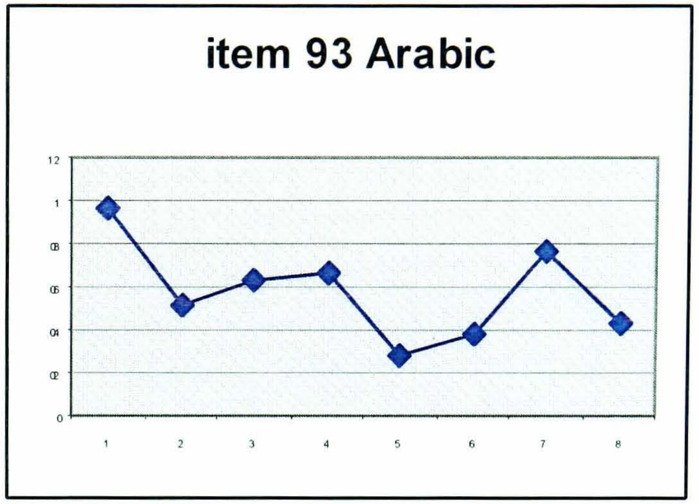
item 62 Arabic

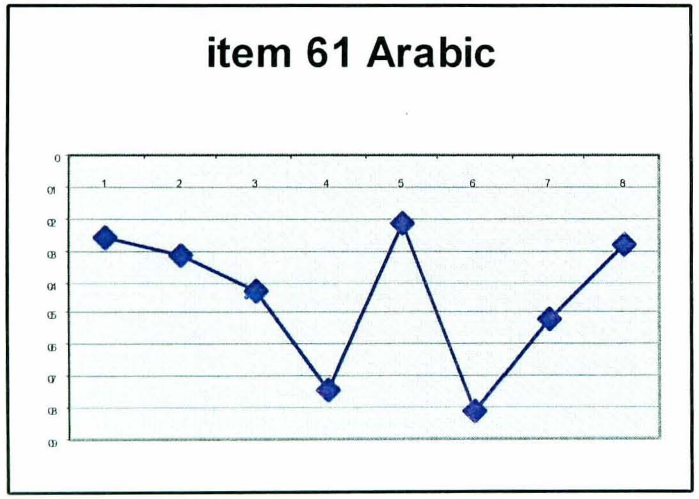


item 67 Arabic




item 102 Arabic
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8


item 93 Arabic
1.2
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
7
8


item 61 Arabic
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
2
3
4
5
6
7
8

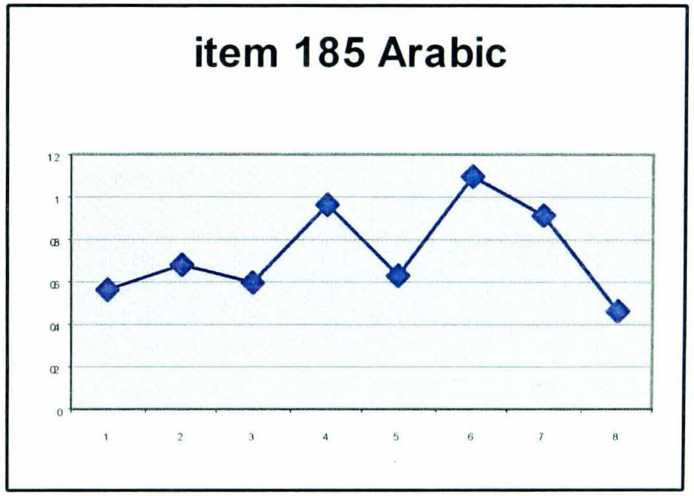
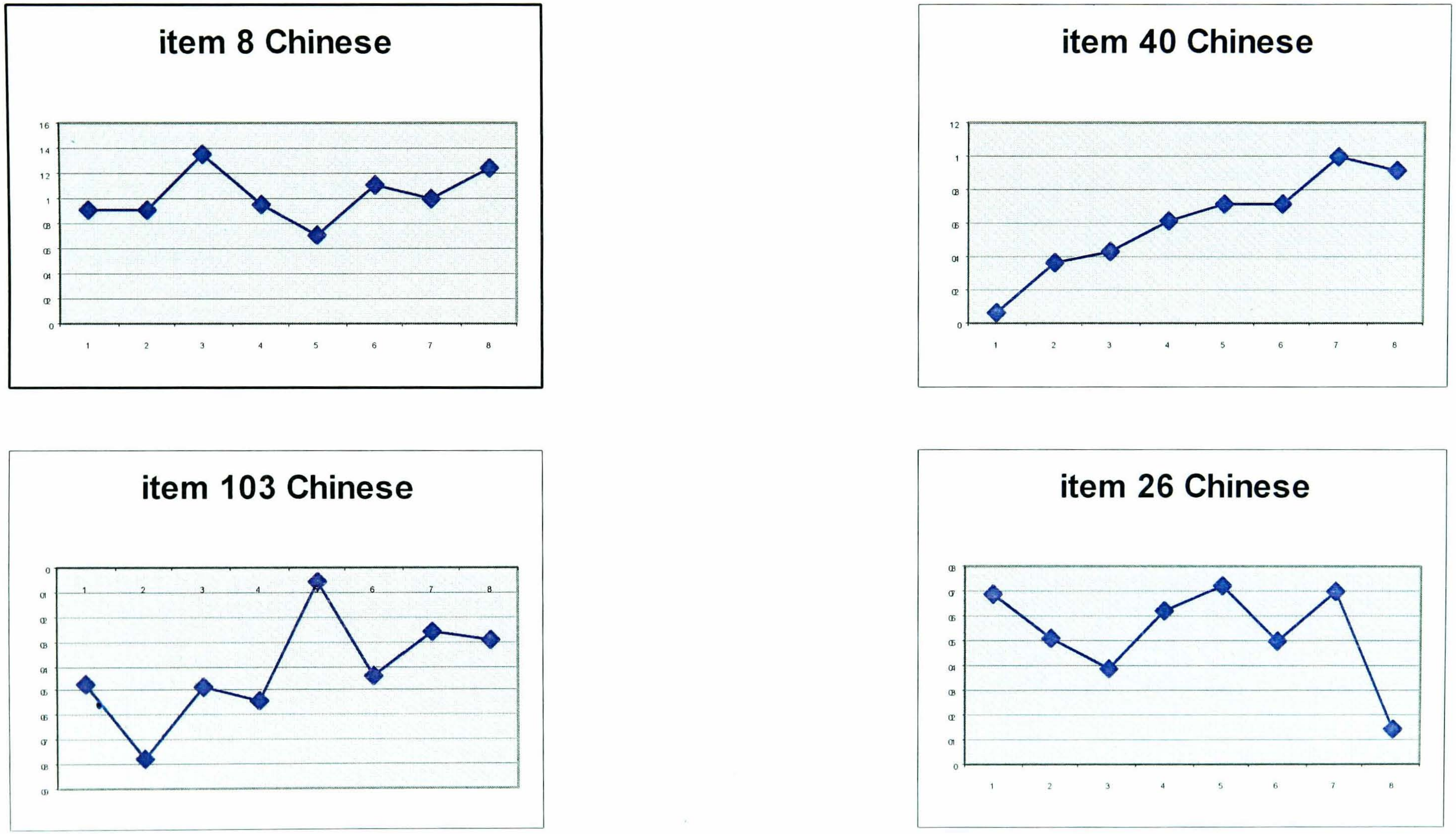
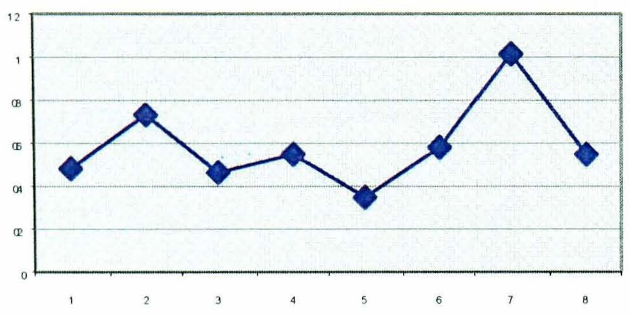

item 185 Arabic
1.2
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
7
8

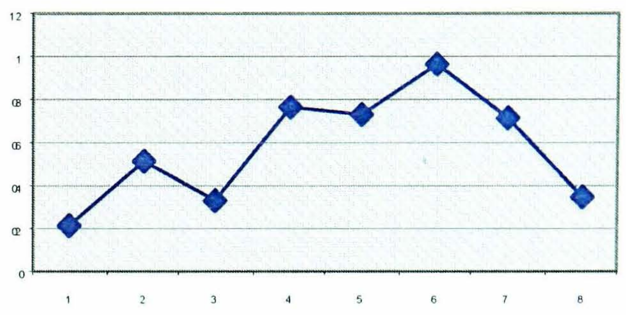
Figure 8.4: Uniformly biased items in the Chinese



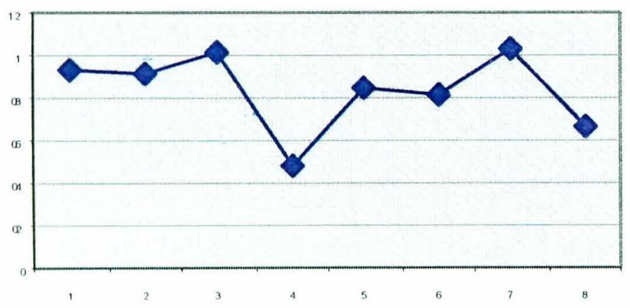
item 92 Chinese



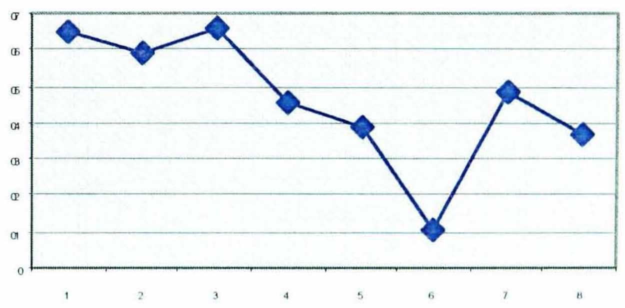
item 91 Chinese

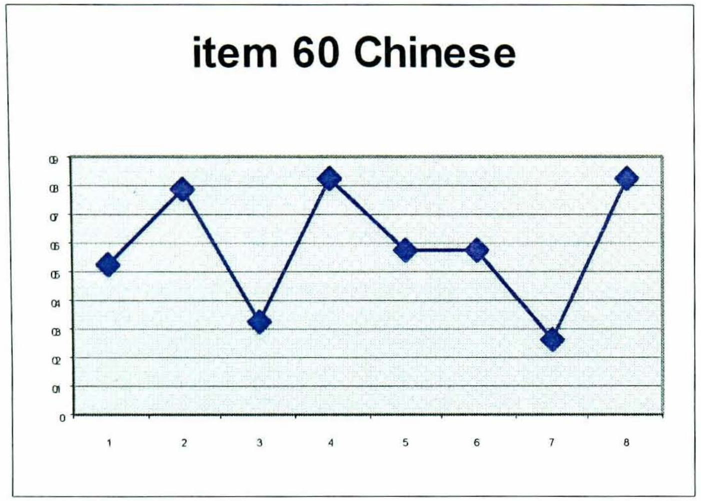


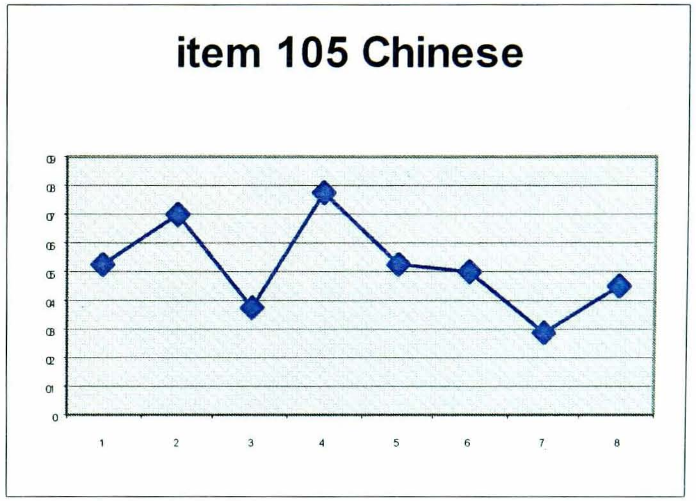
item 175 Chinese



item 183 Chinese




item 60 Chinese


item 105 Chinese

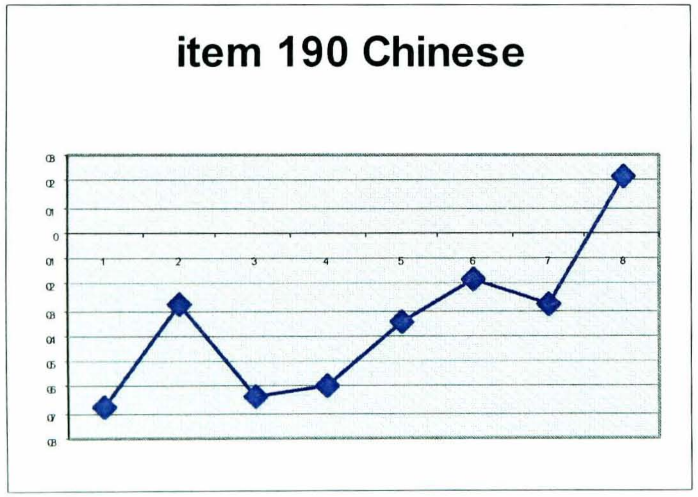
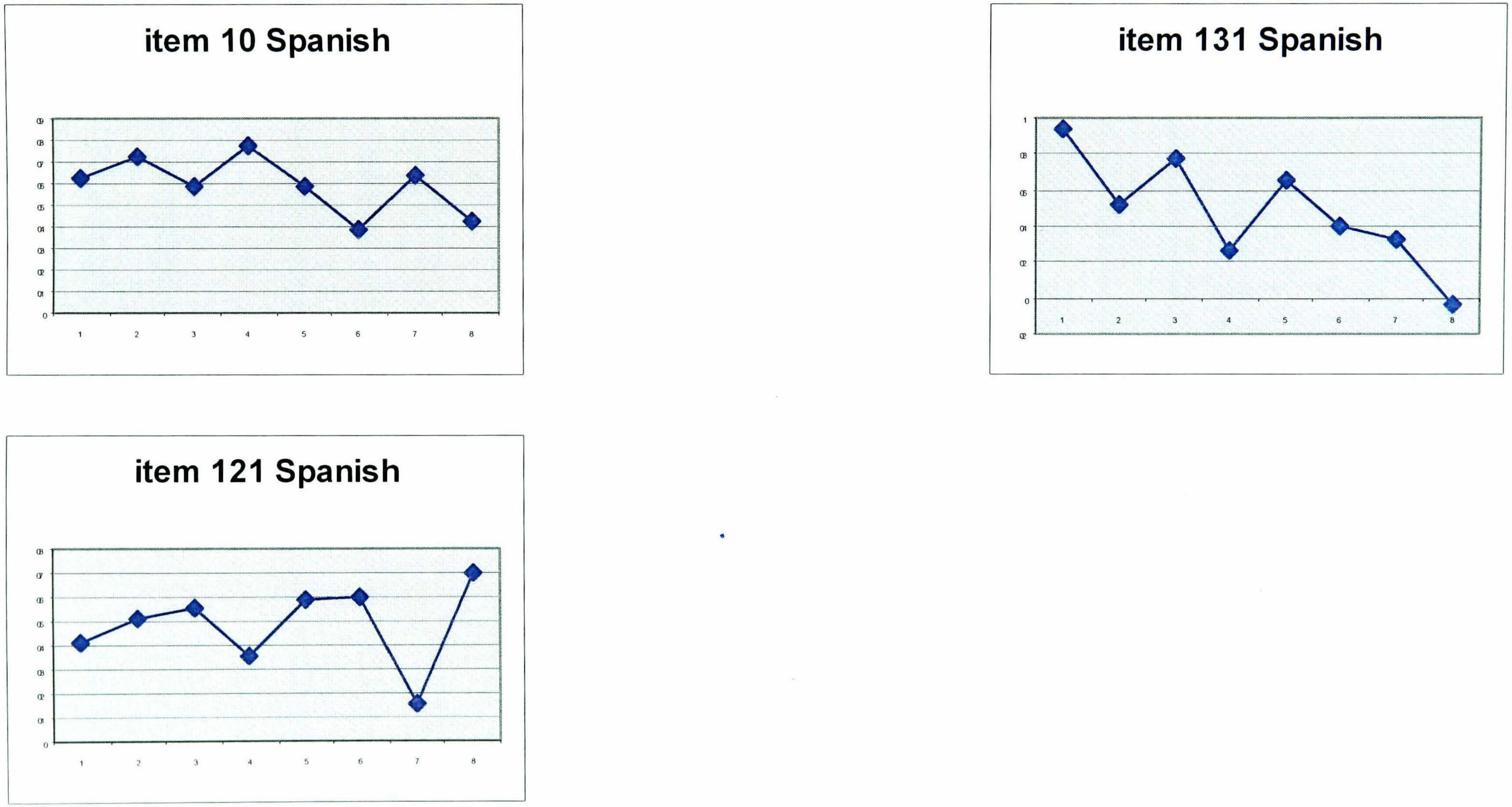

item 190 Chinese

Figure 8.5: Uniformly biased item in Spanish sample



8.7. Conclusion

In conclusion, the wording of items and the simplicity of the language it employs are two factors that affect questionnaire development in general but most importantly test adaptation. Certain words are difficult to translate to other languages and some others might have a differential meaning or psychological impact in the target language as discussed in chapters 6 and 7. Therefore item writing in the original language should be the first step in test adaptation process. Kline (1986) outlines a comprehensive list of guidelines for writing items for personality testing to assist test developers in minimizing error associated with wording. Careful choice of wording can increase accuracy of the information inferred from personality tests (Kline, 1986), but can also facilitate the adaptation into other languages. As an example, *behaviour verbs* such as "I *play* sports" are more specific in their definition and more likely to have a direct equivalent in the target language than *feeling verbs* such as "I *enjoy* sports" which tend to have a more subjective definition from the respondents' point of view (Kline, 1986). Whilst the ITC guidelines discussed in chapter 4 are very useful for adapting tests and Kline's guidelines are very useful for writing items, there is still a need for amalgamating these to produce guidelines specific for writing items intended to be adapted into other languages and cultures. Until these are in place, we encourage the use of "international English" that does not reply on idioms and simple sentence structures in order to minimize linguistic bias during test adaptation.

Moreover, the functioning of items is also dependent of the scale it is scrutinised against, such as item 1 in the Arabic sample, which exhibited DIF under Fellowship but not under Detail. This also adds value to the simultaneous development of multi-lingual versions of the same questionnaire because, on one hand, wording can be carefully chosen based on the structure of the sentence in the target cultures **and** also on how much it taps onto the intended construct in each culture. On the other hand, items that do not work well under a certain scale in a specific culture can be dropped out, rephrased or replaced by other items without risking decreasing the reliability of the scale.

8.8. Comparing ANOVA and Logistic Regression

The proportion of items that ANOVA flagged as DIF was much higher than those flagged by Logistic Regression (74% vs. 12% for Arab world; 67% vs. 11% for China; and 60% vs. 3% for Spain). Moreover, none of the items was flagged as non-uniformly biased using LogR, which is a less common type of bias to occur (van de Vijver & Leung, 1997). Conversely, ANOVA flagged 34 items as non-uniformly DIF. There are several factors that might be contributing to this difference in results. First, ANOVA as a statistical technique has more assumptions that need to be met in order to produce accurate results (Field, 2005). Most importantly, however, test score intervals or groups (see section 8.4.4.2 for the full procedure) were computed in order to match participants in way that the sample sizes per test score group are not too small (around 50). This procedure, known as think matching, can sometimes result in significant differences between the score groups where differences do not actually exist (Sireci, Patsula, & Hambleton, 2005). Additionally, ANOVA has been used extensively in the literature for analysing ordinal data, such as Likert, although one of its assumptions is for the data to be at least interval. Finally, the lack of effect size associated with this technique for DIF detection also contributes to inflating the Type I error and accepting items as DIF whereby they actually are not. As suggested by Zumbo (1999), both significant and not significant results, with effect sizes, should be published in order to be able to establish more accurate criteria than the ones that are being used nowadays. This can lead to determining an effect size value to be associated with ANOVA and that can be used in the future for detecting DIF using this technique.

In conclusion, whether LorR or ANOVA are used for DIF detection, these analyses are problematic when conditioning on test score because of the modest levels of scale reliability across the 4 cultures.

Chapter 9:Measurement Invariance Analysis

9.1. Chapter Overview

This chapter focuses on assessing Measurement Invariance (MI) (Meredith, 1993) between the different language versions of Orpheus using Exploratory and Confirmatory factor analysis with Mplus software to examine the theoretical aspects of equivalence, discussed in Chapter 4, statistically.

We will begin this chapter by defining the statistical concept of measurement invariance (Meredith, 1993) in relation to the Theory of Equivalence and Bias (van de Vijver & Leung, 1997). This discussion will include an in-depth review of the four types of statistical equivalence (invariance): Configural, Weak, Strong, and Strict Invariance.

We then present a detailed description of the analysis conducted, which involved a MG-CFA followed by an EFA for every group separately. This is followed by the results from each step of invariance analysis, which did not yield a good model fit. This might be the result of the small sample size (around 200 from each culture), which could make parameter estimation and interpretations very difficult.

9.2. Defining Measurement Invariance

Measurement Invariance (MI) (Meredith, 1993) is the statistical counterpart of the theoretical concept of *equivalence* discussed in chapter 4. The Theory of Equivalence and Bias falls short in that it has not been directly mapped out on the statistical and practical concepts of MI. This increases the gap between statistical theory and practice, which is already quite vast (Muthen & Muthen, 2008). *Measurement invariance* was developed in a very technical way that rendered it inaccessible to social, behavioural and cross-cultural researchers

(Wu, Li & Zumbo, 2007). In his paper, Meredith (1993) presented the concept of MI for a statistical audience, which made it unattainable by others whose research would benefit from implementing MI. This has lead many advances in psychometrics to stay out of reach of psychologists and other practitioners in the field on measurement (Millsap, 2007). For example, it became prevalent in the literature that the replication of the factor structure between groups is sufficient evidence of measurement invariance, although this is certainly not the case (Byrne &Watkins, 2003). Wu, Li, and Zumbo's (2007) paper has the merit of presenting Meredith's concepts in a very accessible manner, and will be referred to throughout this chapter to illustrate certain concepts of MI. We also aim to illustrate the theoretical and practical sides of equivalence using data from Orpheus collected in four cultures.

Measurement Invariance is a term that refers to statistical hierarchical evidence of equality in *factors* (configural invariance), *factor loading* (weak invariance), *intercept* (strong invariance), and *residual variance* (strict invariance) of two sets of data (Wu, Li, and Zumbo, 2007). Equality on each level, gives statistical evidence for certain levels of comparability between the groups of interest. For example equality in *factors* assumes that the same number of constructs is being measured in each culture but no further evidence is available to allow comparability between them. The four levels of MI will be explained theoretically and statistically under the section labelled MG-CFA on MACS.

MI requires that the same variable(s) be measured using the same metric in order to allow cross group comparisons. MI could apply to one item from a test, a group of items, subtests or whole tests (Millsap, 2007) and requires that the model that links between the *observable* and the *latent variables* be identical across groups (Wu, Li, and Zumbo, 2007). The *latent variable* is the mathematical or statistical variable that we intend to measure through the questionnaire. The *latent variable* is measured through responses to the items, which are referred to as *observable variables*. The total of the responses intended to measure a certain latent variable are referred to as *observable scores*. As explained earlier in chapter 4, according to the theory of True Score, an observed score Y is a combination of an individual's true score and some random error as shown in the formula below (Cronbach, 1990; Kline, 1993; Rust & Golombok, 1999; Fife-Schaw, 2006):

Observed score (Y) =True score (X) + Error (E)

When MI holds, the probability of an individual, with true score X, to attain a certain observed score Y should be independent of the group he or she belongs to. The true score could be the overall score on an ability test, or in the case of personality it could be the score on a particular scale such as extraversion or neuroticism (Wu, Li, & Zumbo, 2007). Therefore, when two versions of a test are fully measurement invariant, respondents with the same true score will have roughly the same observed score regardless of culture, age, gender or any other group that *measurement invariance* is assessed against.

As mentioned earlier, MI is hierarchical so earlier steps should be achieved first before attempting to assess whether a higher order one is tenable. Nevertheless, if the higher order one is not achieved, the earlier step will constitute the highest level of metric invariance possible. For example, configural invariance does not allow any direct comparison between two groups because it only means that the same construct is being measured but not necessarily on the same scale or metric. The level of *equivalence* that can be achieved between two tests (construct, measurement unit and scalar) depends on the level of *measurement invariance* achieved statistically (configural, weak, strong, and strict invariance).

The first level of equivalence is between the constructs being measured and is fulfilled with evidence of *configural invariance*. That is, equivalence of constructs is theoretical and the configural is statistical. The only assumptions that configural invariance assessess statistically is that the *number of factors* is the same across groups. This level of invariance does not provide evidence for any comparison between the groups. Once this level of invariance is achieved, the second level of invariance, weak invariance, is assessed by increasing the stringency of the statistical model. For example, equality of *factor loading* is added to the equality in *number of factors* to test for weak invariance. More statistical constraints continue to be added to the model until reaching the highest level of invariance, *strict invariance.* Theoretically, *scalar equivalence* is considered as the highest level of equivalence achievable and is usually reached when statistical evidence supports *strict invariance*. Nevertheless, it is argued that *strong invariance* can also be considered as evidence of *scalar invariance* (Little, 1997). This will be discussed in detail later in this chapter in MG-CFA on MACS.

Statistical invariance	**Theoretical equivalence**
Configural	Construct
Weak	Measurement Unit
Strong	
Strict	Scalar

Table 9.1: Relationship between Meredith's (1993) statistical model of invariance and van de Vijver and Leung's (1997) Theory of Equivalence and Bias.

9.3. Measurement and Structural equivalence

In following sections, we will illustrate the difference between *measurement* and *structural models*, and then discuss the techniques commonly used for analysing measurement invariance between groups. We will mainly focus on multi group confirmatory factor analysis (MG-CFA) as the main method for assessing measurement invariance. We will also focus on exploratory factor analysis (EFA) as a complementary analysis to CFA (Muthen & Muthen, 2008). We will first distinguish between the terms “construct”, “latent variable”, and “factor”, which are usually used interchangeably in the context of assessing measurement invariance (Byrne, 1998; Zumbo, 2007).

9.3.1 Relationship between construct, latent variable and factor

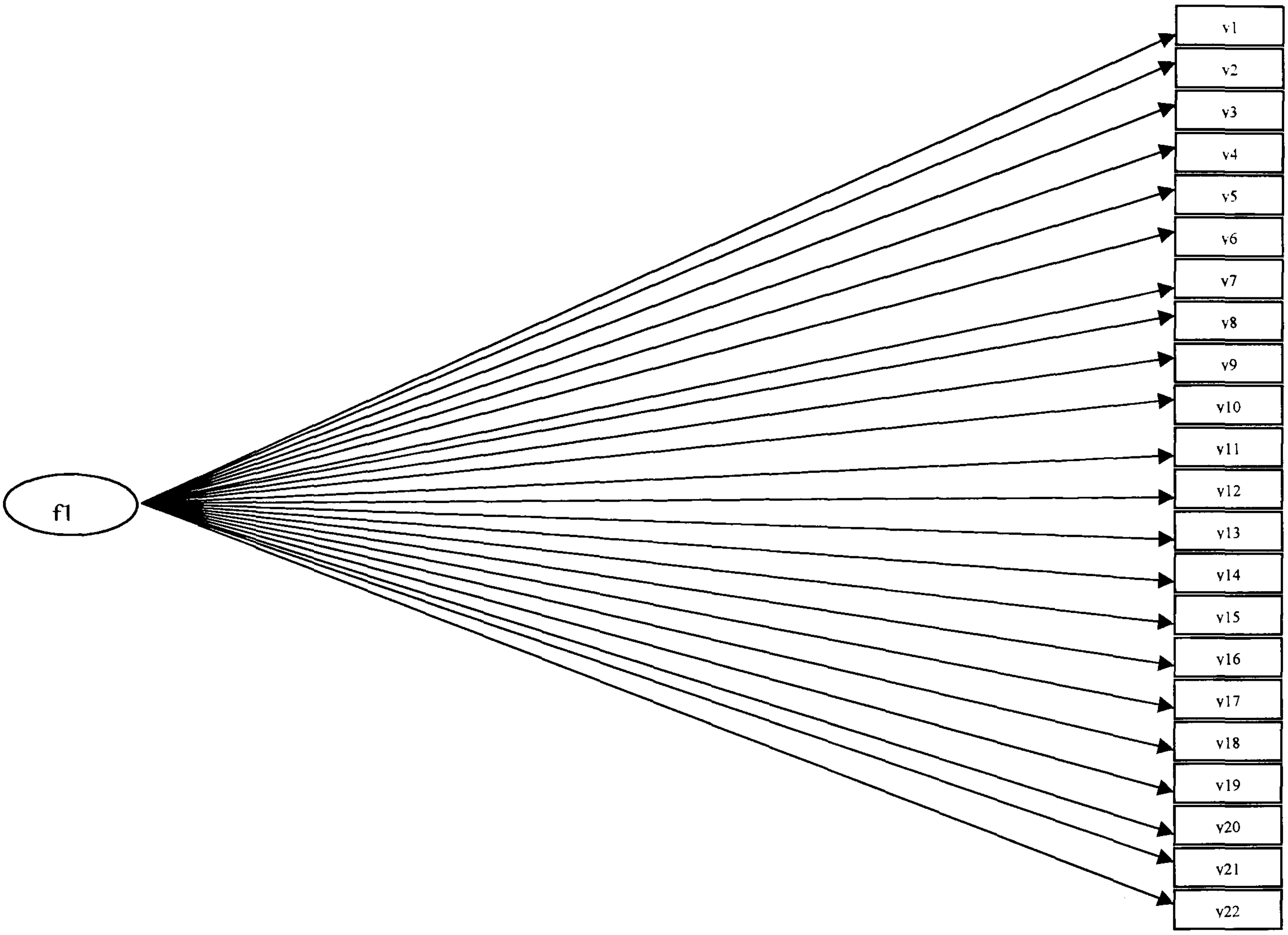
Zumbo (2007) explains that a construct is the theoretical and abstract variable that a researcher is interested in measuring. The latent variable is the mathematical or statistical variable measured through responses to the items. The latent variable mediates between the item responses and the construct of interest to give inferences about this theoretical concept. Therefore, empirical evidence can never be generated directly about the construct, but rather through the statistical mediator the latent variable. Yet, the terms construct and latent variable are used interchangeably in the literature.

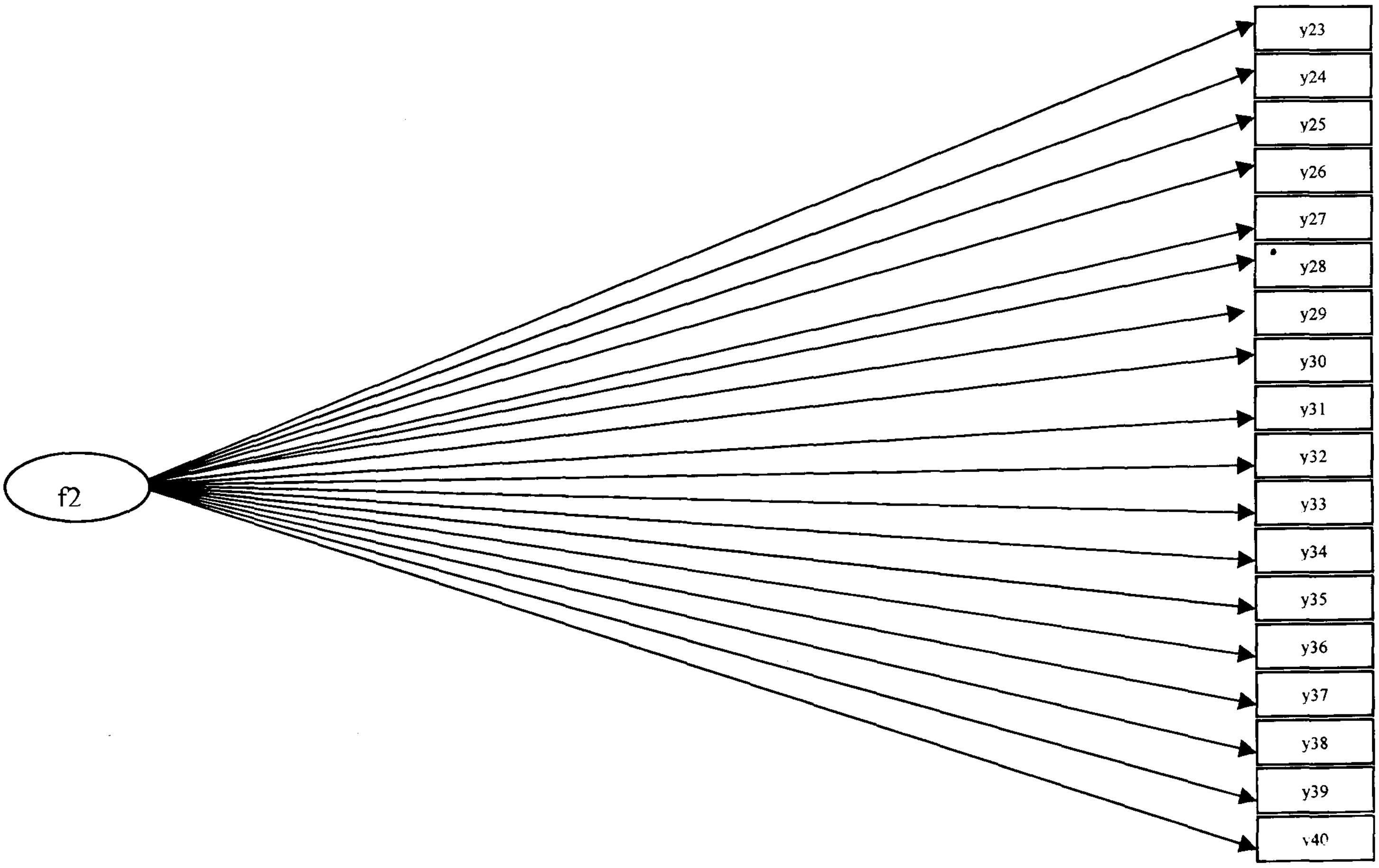
The definition of MI applies to factor analysis and the factor is usually the latent variable, then factor and latent variable can be used interchangeably (Byrne, 1998; Wu, Li, and Zumbo, 2007).

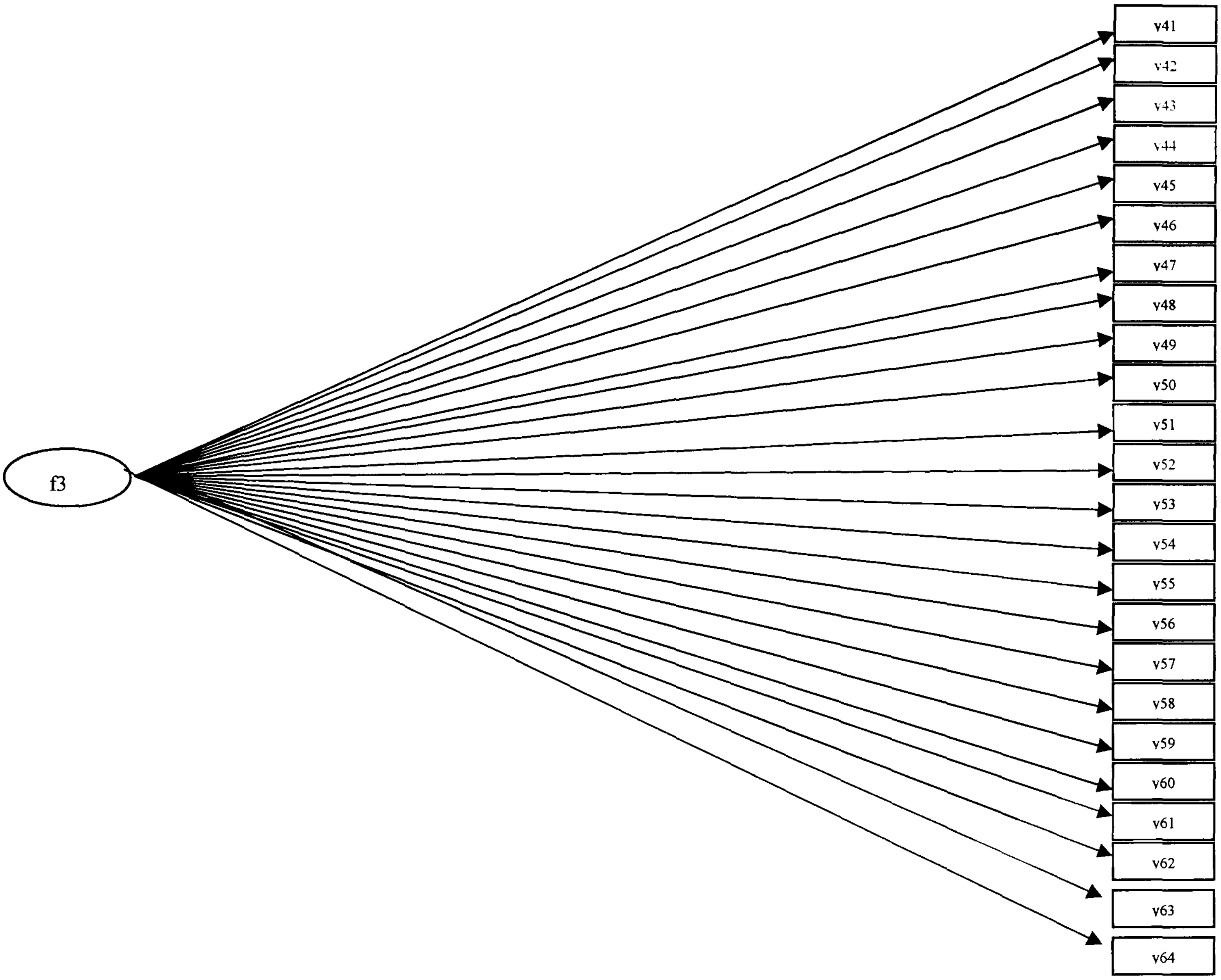
9.3.2 Commonly used methods for analysing MI

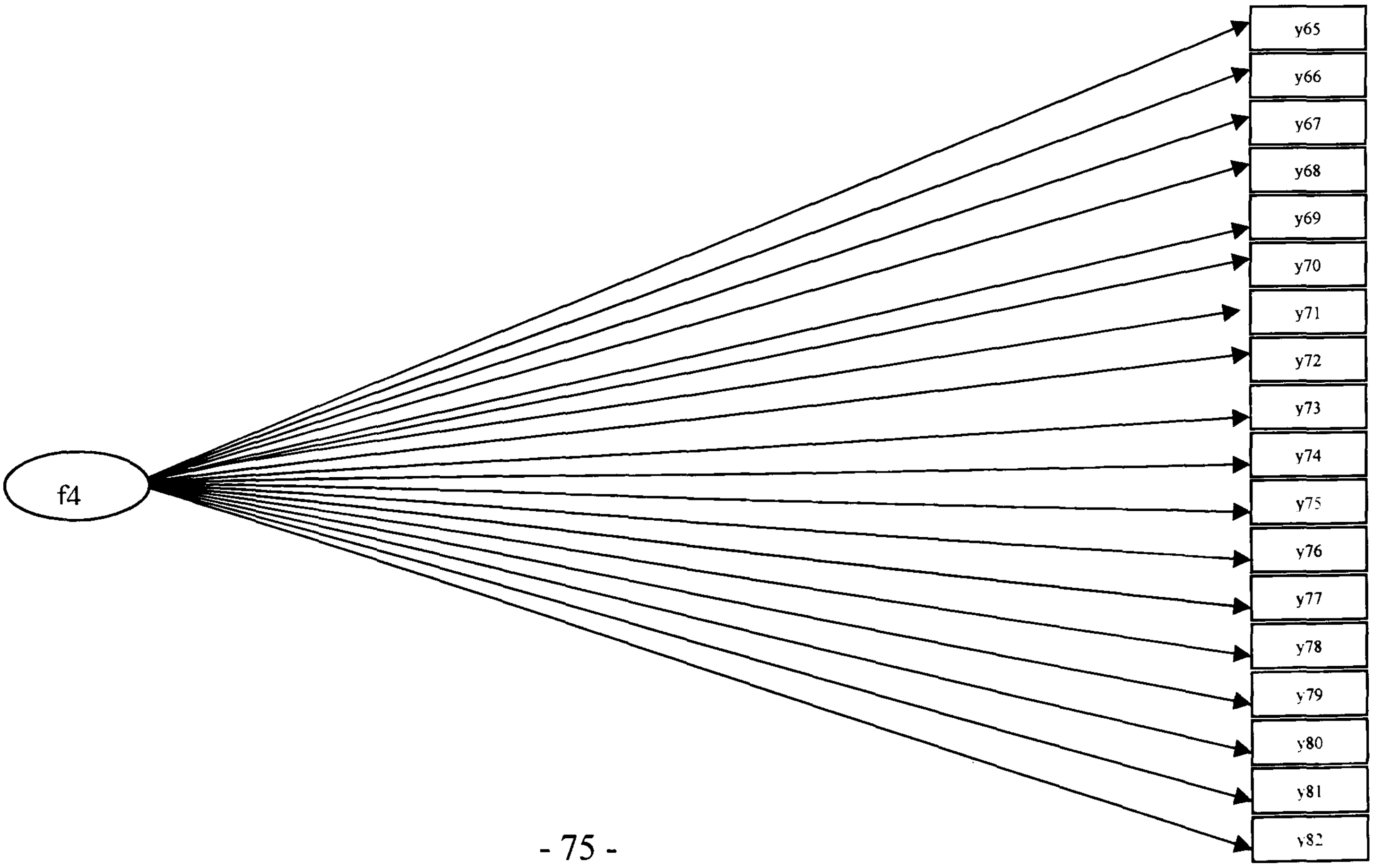
Statistical modelling is essential for describing "the latent structure underlying a set of observed variables" (Byrne, 1993, p 7). Full Latent Variable Models (FLVM) allow for the specification of relationships between observed and latent variables, and also between the latent variables themselves. Statistical models that investigate the relationship between *latent* variables are referred to at *Structural Models*. In contrast, *measurement models* focus on the association between *latent* and *observed variables* (Byrne, 1993; Muthen, & Muthen, 2007).

In this study, the Big Five Factors represent the latent variables being assessed and the relationship between them in Orpheus is less that 0.3, which is considered negligible (Rust and Golombok, 1999). Therefore, there is no structural model associated with this data and only the measurement model needs to be assessed. That is, the model to be implemented on this data should focus on the relationship between the latent variables on one hand (Big Five) and the observed ones on the other (responses to items) as illustrated in figure9.1, which spans over three pages. The variables in the ellipses represent the latent variables and the ones in the rectangles represent the observed ones (items), and the unidirectional arrows represent regressions. There are no bidirectional arrows, which usually represent correlations, between the latent variables because it is assumed that there is no relationship between them (Rust and Golombok, 2001).


f1
y1
y2
y3
y4
y5
y6
y7
y8
y9
y10
y11
y12
y13
y14
y15
y16
y17
y18
y19
y20
y21
y22


f2
y23
y24
y25
y26
y27
y28
y29
y30
y31
y32
y33
y34
y35
y36
y37
y38
y39
y40


f3
y41
y42
y43
y44
y45
y46
y47
y48
y49
y50
y51
y52
y53
y54
y55
y56
y57
y58
y59
y60
y61
y62
y63
y64


f4
y65
y66
y67
y68
y69
y70
y71
y72
y73
y74
y75
y76
y77
y78
y79
y80
y81
y82

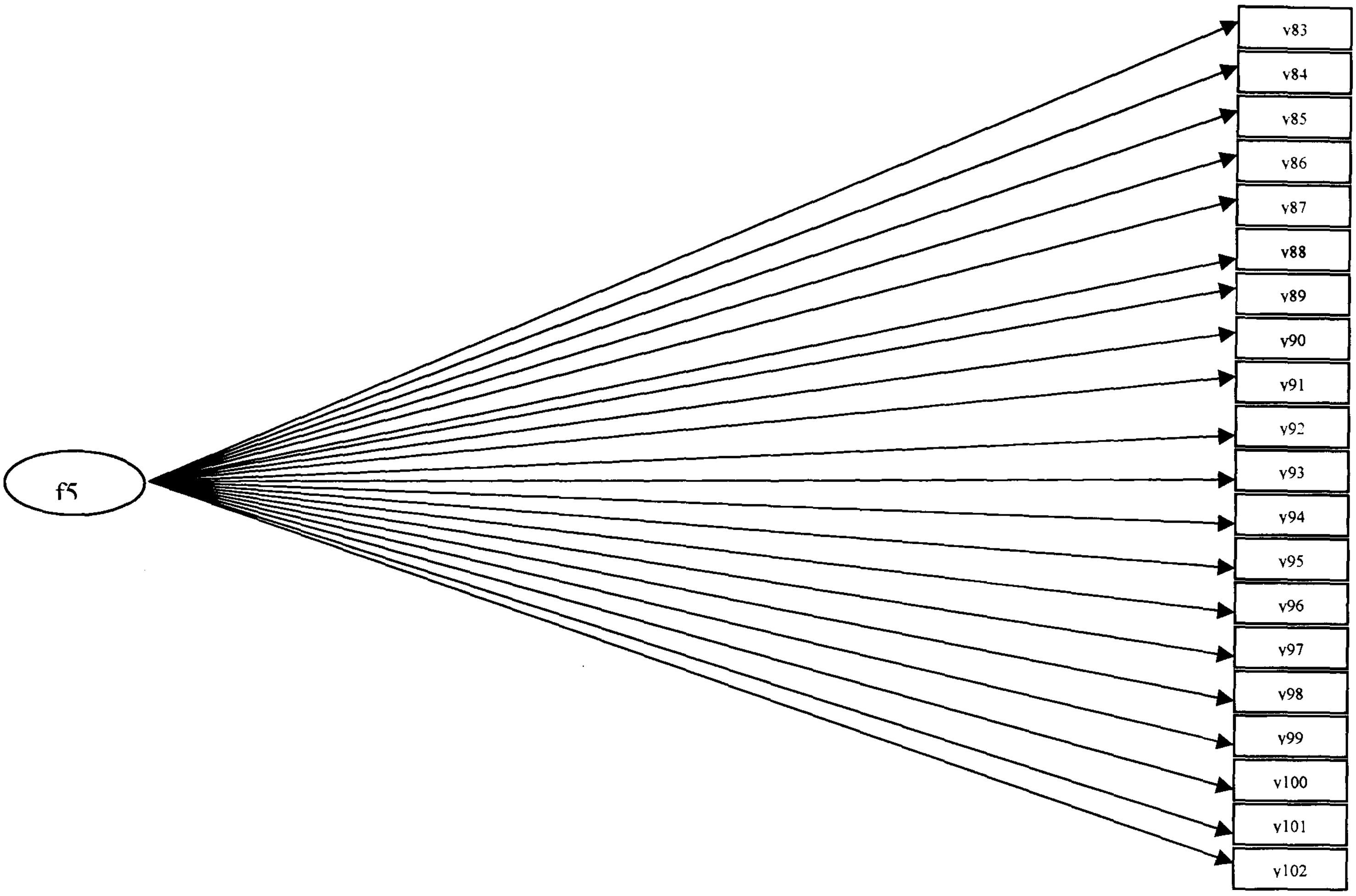


Figure 9.1: Orpheus confirmatory model

The most prominent method for investigating the relationships between latent and observed variables is Factor Analysis (FA), which comprises of two basic types, Exploratory and Confirmatory Factor Analysis (Byrne, 1998). The exploratory approach (EFA) is typically applied in situations where the relationship between the observed and latent variables is unknown whereas CFA is designed to confirm whether the relationship between these two that the researcher assumes is tenable (Kline, 1993). EFA is arguably the most commonly used method for studying construct equivalence (van de Vijver & Leung, 1997). Exploratory factor analysis is used to find the smallest number of factors necessary for defining the relationships between a set of variables without placing any structure on the relationships between observed and latent variables (Muthen & Muthen, 2008). It can be applied to each group separately, and the similarity in the factor-analytic solution will hint to the comparability of the constructs between them. However, it is extremely important that the factor solutions of each group are rotated to each other in order to avoid underestimating the similarity between factors (van de Vijver & Leung, 1997).

Byrne, Shavelson and Muthen (1989) argue that Confirmatory Factor Analysis (CFA), an extension of EFA, is methodologically more sophisticated and is usually used to examine whether a hypothesized model fits another data. CFA has many advantages over EFA, namely "CFA provides a Chi-square test and a goodness-of-fit indicator of the ability of the same factor solution to fit data from different samples" (Marsh & Hocevar, 1985, p 565). Additionally, CFA allows for the factor solution to be specified a priori therefore specific hypotheses can be generated and tested (van de Vijver & Leung, 1997). Nonetheless, the EFA and CFA serve different purposes and could be used to complement each other. Muthen and Muthen (2008) propose a five step process for test development that employs EFA first in order to build a CFA model, which then could be tested on other populations as follows (slide 57)

1) Pilot study 1
    1- Small n,
    2- run EFA,
    3- revise, delete, add items
2) Pilot study 2
    1- Small n,
    2- EFA Formulate tentative CFA model

3) Pilot study 3

   1- Larger n,

   2- CFA Test model from Pilot study 2 using random half of the sample Revise into new CFA model Cross-validate new CFA model using other half of data

4) Large scale study,

   1- CFA

5) Investigate other populations

In cross-cultural test adaptation, it is assumed that the hypothesised model from step 4 has already been established during test development, and therefore CFA is applied to examine its applicability on new samples. This process will be explored further in the discussion.

In this study, we will begin with CFA approach assuming that the hypothesised model that Orpheus has been established as illustrated in figure 9.1. Based on the findings from this analysis, further steps of analysis will be determined.

9.3.2.1. MG-CFA on MACS

Multi-Group Confirmatory Factor Analysis (MG-CFA) based on mean and covariance structures (MACS) is fundamental to the investigation of Strict MI (Wu, Li, & Zumbo, 2007). Technically speaking, MACS models do not only include variance and covariance but also the means of the observed variables. By doing so, the intercept becomes incorporated in the factor analytic model. MG-CFA will be used hereafter to refer to MG-CFA on MACS data.

MG-CFA is the most widely used for this purpose because of its reliance on formal hypothesis testing using likelihood ratio test (Wu, Li, & Zumbo, 2007). When testing for MI, MG-CFA consists of a series of hypotheses and begins with the least restricted model referred to as the configural model. This assumes that the number of factors is equal between the groups, without making any further assumptions. Once one level of equality is satisfied and confirmed, the restrictions become increasingly constrained by adding more assumptions of equality between groups such as equality of factor loading (weak invariance), equality of the intercept (strong invariance), and finally equality of residual variance (strict invariance). Although Meredith (1993) argues that comparability of constructs cannot be established if this level of equality is not established, Little (1997) maintains that strong invariance is sufficient for demonstrating measurement invariance. These arguments can be best described and challenged while referring to the regression equation.

9.3.3 Configural, Weak, Strong and Strict Invariance

The factor analytic model incorporating MACS is represented in the following equation (Wu, Li, & Zumbo, 2007, p 3):

$$y_{ij} = \tau_j + \lambda_{j1}\eta_{1i} + \lambda_{j2}\eta_{2i} + \ldots.\lambda_{jp}\eta_{pi} + r_{ij}$$

Where τ = the intercept which is the factor score y when factor score is equal to 0; λ_{jp} = regression coefficient (slope) which is the loading for item j on factor p;

η_{pi} = factor p; and

r = normally distributed random residual due to random fluctuation in the response process.

These four parts constitute the measurement model of the equation, which denotes how the observed variables relate to the latent common factors. The equality in $\tau, \lambda, and \eta$, indicates that the observed variables in the different groups have the same relationship with the latent variable.

Configural invariance is examined by holding the number of factors constant across the groups (Wu, Li, & Zumbo, 2007). If this type of equivalence is not demonstrated, then the tests are measuring different construct in each group of interest. Theoretically, this leads to construct inequivalence, which renders any comparison between the groups inadequate and no further tests of invariance should be undertaken. On the other hand, if this level of invariance is achieved, it is reasonable to check whether the subsequent level of invariance holds too.

Weak invariance assumes that the same measurement unit (one unit of change) is equal between two groups (Wu, Li, &Zumbo, 2007). In addition to the previous constraints, the factor loading (or slope) is held constant between the groups. If weak measurement invariance fails, we are justified to assert comparison between the groups is inappropriate (Meredith, 1993). When weak invariance is established, this indicates that the measurement units are equal between the groups but does not guarantee that direct comparison between groups is reasonable. The only comparison possible at this stage is difference between groups. That is, a difference between males and females for group 1 can be compared to the difference between males and females for group 2. Figure 9.2 below illustrates weak invariance (equal factor loading or slopes) whereas figure 9.3 illustrates weak non-invariance (unequal slopes). Figure 9.2 shows two parallel slopes (equal slopes) though the intercept is different. When y=3, x_{1red}=1.75 on the red line (group1) whereas x_{1blue}=2.75 for the same y on the blue line (group 2). Moreover, when y=3.5, x_{2red}=3 for group1 and x_{2blue}=4 for group2. Therefore the groups are not directly comparable because the same observable score y refers to different x on group1 than in group2. Nevertheless, the difference on the latent variable between x_{1red} - x_{2red}=1.25 is equal to the difference on the latent variable between x_{1blue} - x_{2blue}=1.25. Since configural equivalence has already been established in the previous step, then the latent variable is assumed to be the same and the differences on the latent variable in one group are comparable to differences on the same latent variable in the other group.

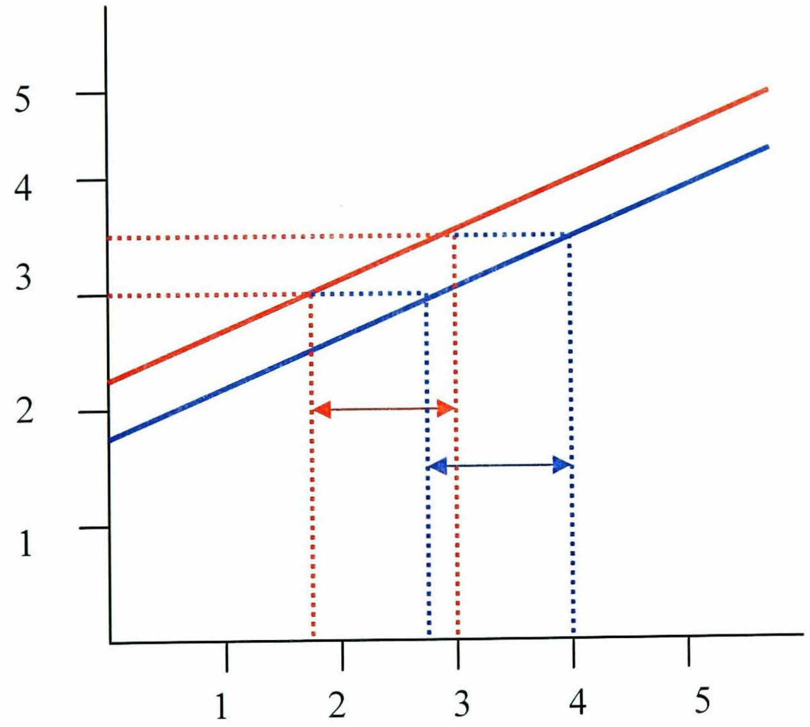


Figure 9.2: Equal loading (slopes)

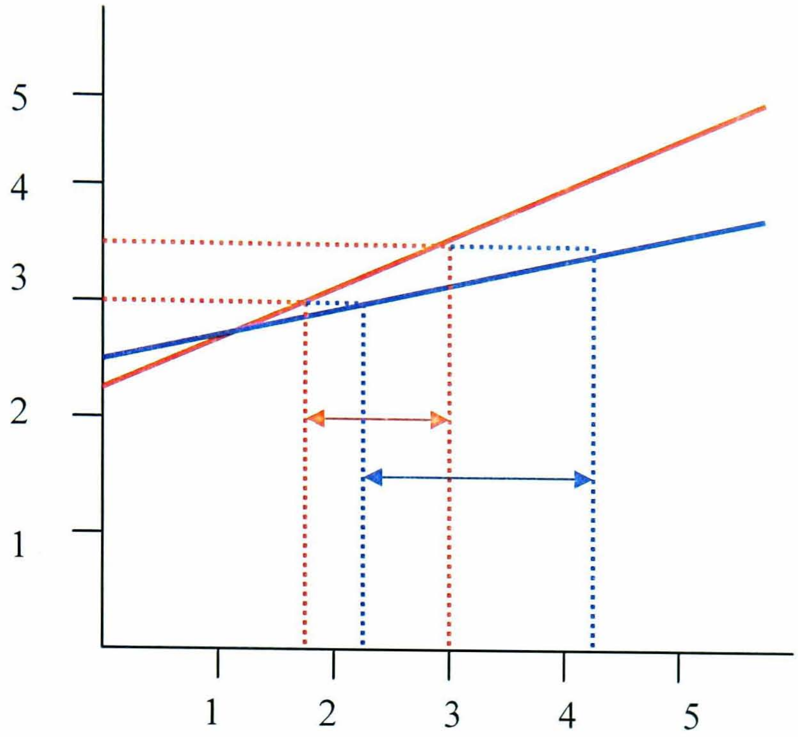


Figure 9.3: Unequal loading (slopes)

On the other hand, figure 9.3 illustrates two weak non-invariance because the slopes are not equal (lines are not parallel). As with the previous example, the same observable score y does not represent the same x in the different groups and is therefore not directly comparable. However, although the latent variable is the same in both groups, the difference between $x_{1red} - x_{2red}$=1.25 is not equal to the difference between $x_{1blue} - x_{2blue}$=1.75. This implies that the same observable scores do not represent the same latent variable and the difference between observable score in one group lead to differences on the latent variable that are not comparable to the second group. Theoretically, this is the case of measurement unit inequivalence discussed earlier in Chapter 4.

In addition to equality of factors and factor loading, strong invariance requires that the intercepts (same starting point) are also equal as illustrated in firgure 9.4 below.

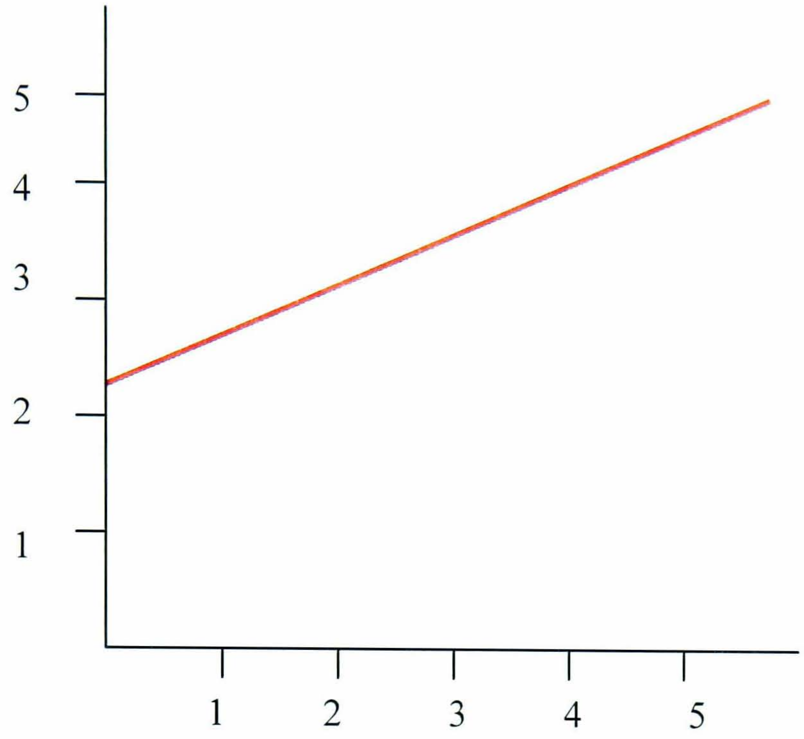


Figure 9.4: Equal slopes and intercepts

The lines are overlapping because the slope is the same since weak invariance must have already been established, and now the intercepts are also equivalent. The two lines are the same and direct comparison between the two groups is possible. This is the case of scalar equivalence discussed earlier in the Theory of Equivalence and Bias in Chapter 4.

Finally, the highest level of invariance that can be assessed is strict invariance, whereby slope and intercept are equal in addition to the regression residual variances (Wu, Li, & Zumbo, 2007).

The variation in people's responses on a certain questionnaire depends partly on their "true score" on the latent variable and partly on random noise or error, which is manifested in the residual variance. However, some argue that this error is not actually random and might affect item responses consistently (Cronbach, 1947, in Deshon, 2003). Depending on the interpretation of the residual variance, strict or strong invariance can be considered as evidence of MI. That is, if the residual is considered as random, then strong invariance is sufficient and equality in residual variance is unnecessary. If residual is random, then there is no need for it to be equivalent across cultures so ensure comparability. Conversely, if the residual is attributed to other causes of variation unrelated to the latent variable then it is important to establish the equality in residuals between the groups.

Wu, Li and Zumbo (2007) explain that researchers force the number of factors, loadings, and intercept to be the same in MG-CFA, which "allows item specific effects to reside only in the residual terms and remain undetected if strict MI is not investigated and consequently disguising possible biases in the test scores" (p18).

Therefore, strict invariance is important to investigate for full measurement invariance because the residual holds differences that are not random. However, strong invariance is sufficient only when the residual variances are low and uncorrelated because this reflects that the residuals in the different groups do not necessarily stem from the same source (see Wu, Li & Zumbo, 2007 for further discussion about this topic).

Type of invariance	Equality of	Comparison	Type of equivalence
Configural invariance	Number of factors	Comparison not possible	Construct equivalence
Weak invariance	Number of Factors + loading	Comparison possible but not ideal	
Strong invariance	Number of factors + loading + intercept	Comparison possible	Measurement Unit equivalence
Strict invariance	Number of factors + loading + intercept + residual	Comparison ultimate	Scalar equivalence

Table 9.2: Relationship between equivalence and invariance

9.3.3.1. Fit indices for accepting MI

Chi-Square is a commonly used index for investigating MI by examining the Chi-Square difference (Δx^2) between two consecutive models, (such as configural and weak or weak and strong). If the difference is significant, then the more restricted model holds and more demanding tests of MI can proceed. However, when Δx^2 is not significant, the less restricted model holds and no further tests of MI need to be done (Byrne, Shavelson, & Muthen, 1989; Cheung & Rensvold, 2002).

Chi-Square difference is dependent on the sample size, which increases the likelihood of rejecting the null hypothesis with large sample sizes (Wu, Li, & Zumbo, 2007). Wu, Li, and Zumbo (2007) reported Chi-Square, RMSEA (Root Mean Square Error of Approximation), and CFI (Comparative Fit Index) on 12 large culture groups ($945<n<1887$) and found that Chi-Square rejected MI at configural invariance level, though CFI and RMSEA demonstrated good invariance at that level. Moreover, the complexity of the model might also affect the accuracy of the inferences drawn from Chi-Square. A Monte Carlo study by Cheung and Rensvold (2002) examined several fit indices and found that RMSEA is not affected by model complexity and therefore recommended RMSEA$\leq$0.05 as an effect size for Chi-Square to determine configural invariance. However, when the model under investigation is not complex, CFI>0.9 for configural invariance and ΔCFI$\leq$-0.01 for model difference can also be employed.

9.4. Method

The same methodological approach from the previous chapter applies.

9.4.1 Analysis

Popular programmes in social sciences such as SPSS15 (SPSS Inc., 2006) cannot perform multi group factor analysis such as MG-CFA. Additionally, these programmes do not allow specification of free and fixed loading necessary for investigating the different levels of measurement invariance (Wu, Li, & Zumbo, 2007). Programmes such as Mplus, LISREL, AMOS, and EQS overpower more commonly used programmes for multi group factor analysis. Mplus (Muthen & Muthen, 2007) is a statistical modelling programme that was chosen for this analysis and which allows multi group factor analysis from different samples and with ordinal, binary, continuous, censored, nominal, counts or any combination of these data.

The first analysis employed was a MG-CFA on MACS, whereby the loading and intercept were set to be free. However, the data did not converge using this model so this was followed by CFA on MACS for every culture independently. This analysis did not converge either suggesting that the model needs reinvestigation. Therefore, EFA for up to 5 factors was applied for every culture separately, followed by varimax, then oblimin rotations.

Within subject standardisation was not used in this case since the data will become ipsative thus obstructing the multilevel factor analysis (van de Vijver & Poortinga, 2002). Moreover, standardisation eliminates differences in response style but might also eliminate actual cross-cultural differences (van de Vijver & Leung, 1997).

9.5. Results

Results of EFA showed that a different number of factors was extracted from each cultures, 4 for UK and China, 5 for the Arab world and 3 for Spain. The values of the fit indices that indicate good model fit are as follows: χ^2 with p>0.05; RMSEA ≤0.05; and CFI>0.9. None of these however showed good model fit with any of the fit indices (table 9.3 below).

The same analysis was run with a Varimax then Oblimin rotations, and the results were exactly the same as the first EFA.

Culture	factors	χ^2 (df), p	RMSEA	CFI
UK	**1**	12147.193 (5049), p=0.000	0.083	0.233
	2	11277.277(4948), p=0.000	0.079	0.316
	3	10508.402(4848), p=0.000	0.076	0.388
	4	10102.110(4749), p=0.000	0.074	0.421
	5	No convergence		
Arab world	**1**	11101.721(5049), p=0.000	0.078	0.146
	2	10038.924(4948), p=0.000	0.072	0.282
	3	9375.177(4848), p=0.000	0.069	0.361
	4	8944.125(4749), p=0.000	0.067	0.408
	5	7996.652(4651), p=0.000	0.060	0.528
China	**1**	12479.114(5151), p=0.000	0.075	0.136
	2	10559.255(4948), p=0.000	0.071	0.234
	3	10079.243(4848), p=0.000	0.070	0.286
	4	9691.175(4651), p=0.000	0.068	0.326
	5	No convergence		
Spain	**1**	11479.883(5049), p=0.000	0.082	0.120
	2	10694(4948), p=0.000	0.078	0.213
	3	10081.162 (4848), p=0.000	0.075	0.284
	4	No convergence		
	5	No convergence		

Table 9.3: EFA

9.6. Discussion

9.6.1 Test development and adaptation

The initial result using multi group and single group confirmatory factor analysis (CFA) revealed no convergence of the data. The suggested model that derives from Orpheus original factor structure did not fit the data and needed further exploration and readjustment. Although CFA is the approach usually used to establish cross-cultural equivalence of multi-lingual versions of an already existing valid and reliable test, Orpheus was not developed based on a factor analytic model so a CFA model can be considered premature. In order to alleviate the restriction on the data, exploratory factor analysis was used next to investigate whether a more appropriate model can be established across the four cultures. The results showed that there is no model that fits the current data as it is.

Several possible explanations for this observation can be offered including linguistic, psychological, or cultural adaptation problems; non-universality of the model at hand; methodological problems; and intrinsic problems with the test. Since there was no good model fit for the UK data, the issue of non-universality of the Big Five model cannot be addressed here because the Big Five model have been shown to emerge in the UK on may occasions (see chapter 2 for further details) but not with the test in this study. There is currently to our knowledge no cross-validation evidence available for the Orpheus either, which should ascertain how the scales correlate with other FFM instruments. Moreover, potential adaptation problems and the thoroughness of the adaptation process cannot be discussed to a full extent, given that the convergence problem was not unique to the adapted versions. Rather, the original English version failed to converge into an appropriate model. However, methodological problems (such as small sample size and reliance on samples of convenience) as well as intrinsic problems with the test may have lead to these results. Small sample sizes habitually create problems with interpretations and parameter estimation as CFA and EFA rely on an adequate participant to parameter ratio (Muthen & Muthen, 2008). Hence, it is possible that a larger sample size might have lead to different results, the present research was for logistic reasons (all data was collated by the researcher herself) to a sample size that could be considered less than adequate. Although this is a valid argument, the technical qualities of the original version of Orpheus are modest, such as relatively low reliabilities and no factor analytic evidence, making it impossible to be more precise about the potential source of the problem in model convergence. Validity and reliability data from Orpheus are not fully supportive of the model it aims to measure. Therefore the failure of model convergence could be the consequence of intrinsic test problems rather than sampling issues. The original Orpheus data may be as inadequately fit to any specific model as the data collected for this study, which highlights the importance of psychometric validity and robustness in any test.

In chapter 3, we explained that validity is concerned with the soundness of the inferences that are made from tests (Cronbach, 1990) and that a test is considered valid when it measures what it purports to measure (Kline, 1993; Rust & Golombok, 1999). There are different approaches to assessing validity (Anastasi, 1988), and whichever technique is used should lead to assumptions about how well the test measures what it claims to measure. Construct validity of Orpheus was examined through criterion related validity. It could be argued, based on the results of this study, that either the criterion used in the validation of Orpheus was inadequate or criterion related validity as a method of test validation is not sufficient on its own. The observations from this study are in contrast to the fact that Orpheus is an accredited test, and has gone through the peer review process by the British Psychological Society which deems it adequate in terms of validity and reliability. A more critical investigation of the criterion used in validity studies is essential for making appropriate conclusion and validity of tests, as is the assumption that one method of testing for validity is sufficient. Additionally, as Orpheus is a test based on a data driven model we posit that it should have been developed using such methods.

9.6.2 Implications for future research and practice

The adaptation of Orpheus into Arabic, Chinese and Spanish was a long and painstaking process that incorporated a great deal of prudence and in-depth investigations of wordings, sentence structures, and psychological effects of items. Yet, a model fit proved impossible to generate, and Orpheus in its present form should not be used cross-culturally due to its technical qualities. One possible follow up for this research could involve removing items that did not load well in any of the cultures, and developing new items that tap on the five main constructs of interest. The steps suggested by Muthen and Muthen (2008; see section 9.3.2) could be followed until a good model with observable variables that accurately assess the latent variable can be achieved. However, a totally different approach might be more suitable for developing tests for cross-cultural use. In practice, tests are usually developed and validated in one language then adapted to other languages and cultures. This approach limits the scope of achieving equality as it treats the original version of the test as static. Therefore, cultural accommodations are bound to the adapted versions only. An alternative approach to *simultaneous test adaptation* could be *simultaneous test development*, which treats all versions as the original ones. Changes can be made to all of the original versions simultaneously in order to increase the likelihood of developing a cross-culturally sensitive tool.

As discussed earlier, Muthen and Muthen (2008) suggested a five-step model for test development that combines EFA and CFA. Following up from this suggested model, simultaneous test development can be preceded by multi group EFA in order to build the hypothesized model across the cultures of interest concurrently. Consequently, items can be removed, added and changed after the first pilot, based on their loading in all culture groups. Then, the data could be explored with another EFA to examine whether these changes improved to model (Muthen & Muthen, 2008). This can instigate a potential CFA model, which should be tested on a small sample first to ensure its applicability before testing it on a larger sample. Although Muthen and Muthen (2008) developed this model for developing a single questionnaire and argue that it is only at this stage that a questionnaire is advised to be investigated in other populations, the same could be done simultaneously to all versions. Simultaneous test development might be a more cross-culturally appropriate approach to adopt rather than rely on the uni-culture centric approach traditionally used for producing trans-linguistic versions of tests.

9.7. Final conclusion

This thesis took a multi-disciplinary approach by synthesizing literature and findings from cross-cultural assessment, organisational psychology, personality, psychometrics, and test adaptation with a main focus on the last topic. The main aim of the thesis was to develop a practical framework for adapting personality questionnaires into different languages while ensuring they continue to measure the intended construct. To investigate this, Orpheus Big Five work-based personality questionnaire was used as the main instrument. Consequently, establishing the equivalence between the different language versions became another focus of this PhD. Since the literature around the universality of the Big Five model is inconclusive and Orpheus is based on the Big Five, the generalisability of this model was also considered central to the thesis. We will first highlight the contributions of this Thesis valuable then summarize the key findings and their implications.

9.7.1 Relevance of the Thesis to academia and practice

Academics across many disciplines are becoming increasingly interested and involved in cross-cultural research, as demonstrated by the large increase it the number of publications in the field (van de Vijver & Leung, 2000; Casillas & Robins, 2005). Academic research usually involves using different types of data collection tools, one of which is very commonly used for its time and cost effectiveness: psychometric tests. Therefore, many published academic studies involve translating tests and comparing findings from different cultures based on them, without sufficient evidence of equivalence between the versions. In parallel, practitioners and test publishers are putting a lot of investment in internationalising their test instruments due to the increasingly globalised economy (Daouk, Rust, and McDowall, 2005). Nevertheless, not all of them are using the appropriate approaches to do so. Many rely on translators to adapt their tests and assume that if the test works in the original language, then it will do the same when translated. The reason for these 'fast tracks' to adaptation, which lack methodological rigour, is partly due to the inaccessibility of the technical language used in the academic field of statistics. To date, knowledge exchange between statisticians, psychometricians, psychologists, cross-cultural researchers, other academics and practitioners is limited. Statistical theory that focuses on assessing equivalence between tests has been established, yet the language employed in different disciplines makes relevant knowledge inaccessible. This thesis was concerned with, among other things, bridging the gap between some of these disciplines, by translating statistical information into a practical and a theoretical framework of test adaptation that can be used in cross-cultural research. The value of such academic research lies in its implication for practice. Many high stakes decisions are being made based on inferences drawn from multi lingual versions of psychometric tests, though their validity has not been necessarily established properly (Daouk, McDowall & Rust, 2005). Test adaptation is multi disciplinary and needs to be disseminated to practitioners to ensure that academic knowledge and findings are being implemented across all academic disciples and in practice.

As a result, a practical framework of test adaptation developed for this Thesis as an easy to follow procedure that academics and practitioners can apply in their work. The framework was divided into two main categories: Quality Control followed by the Pilot. Each part of these categories contained a series of practical steps, the details and strengths of which will be discussed in the next section. The practical framework provides a methodological process that can be adopted, in part or in full, to facilitate the adaptation of the instruments that researchers wish to use in different languages. The theoretical framework, on the other hand, builds on van de Vijver and Leung's (1997) Theory of Equivalence and Bias and condenses the considerations that need to be taken into account before and during the adaptation of psychometric tools into other languages. The Theory of Equivalence and Bias lists the types of bias and equivalence that need to be dealt with during test adaptation, but relies only of few examples of the sources of these biases. This theoretical framework facilitates the understanding of the Theory of Equivalence and Bias as it incorporates many sources of bias that have been discussed in different publications (such as van de Vijver & Leung, 1997; van de Vijver & Poortinga, 2001)

9.7.2 Triangulation of adaptation methods

The practical framework of test adaptation triangulates quantitative and qualitative data collection methods (dyads/ triads, pilot, cognitive interview), while relying on the experience and expertise of native speakers of the target and original languages, professional translators, and psychometricians. This triangulation acts as a cross examination for controlling the quality of the translation process efficiently.

This framework is the product of amalgamation of methods that have been used in various test adaptation studies, however, some of these techniques have been applied in certain fields but not in others. For example, cognitive interviews have been prominently used in survey development (Willis, 2004).

Similar approaches that attempt to understand the thinking process of test takers might have also been used in adapting educational tests. However, cognitive interviews are not recognised as one of the common techniques that can help underpin many of the issues faced in test adaptation. During the adaptation of Orpheus, some items that functioned differently when presented to participants from the same culture but in different languages. Cognitive interviews proved to be an extremely useful technique in identifying linguistic, psychological, and cultural problems at item level. Cognitive interviews unravelled problems deeply embedded in items and helped salvage many of them and are therefore highly recommended as an integral tool for test adaptation.

9.7.3 Personality, culture, and the Big Five Model

Personality tests are common methods of assessment that are increasingly being used in the workplace. Personality tests rely on questions that highlight certain behaviours in order to make inferences about key psychological constructs relating to test takers. Considering the complex relationship between personality and culture, and the fact that they are both manifested behaviourally, it is difficult to assess personality using psychometric tests without carefully considering culture. In cases where full measurement equivalence between multilingual versions of a test exists, it does not mean that the tests are culture-free. It is very likely that the behaviours and wording of the questions in that test are less culturally sensitive than other tests.

When dealing with work-based personality questionnaire, the effect of culture becomes increasingly challenging. All personality questionnaires are somehow affected by culture because of their reliance of behaviours. Even nonverbal tests, which were previously thought to be culture-free, can exhibit culturally dependent stimuli such as familiarity with certain shapes (van de Vijver, 2005). Work-based questionnaires are more challenging because they include another layer of culture, namely organisational culture. This creates an added complexity to the adaptation of work-based personality tests, making them more culturally bound than other personality tests. Cultural accommodations relating to general life as well as organisational behaviour can be done more effectively if changes can be done across all languages, including the original version. Therefore, the simultaneous test development suggested in the discussion of section 9.6 might be particularly more appropriate for developing multi-lingual versions of work-based personality tests.

Conclusions about the universality of the Big Five model could not be made based on the studies in this Thesis due to the technical problems in Orpheus outlined earlier. However, the review of the literature presented in chapter 2 clearly indicates that conclusion about the universality of this model are premature. Cross-cultural investigations presented in the literature have lead to conflicting results, with cross loading between certain constructs (e.g. Cheung et al, 2001). The replication of Tupes and Cristal's (1961) study in non-western cultures have lead to the emergence of models of personality different than the Big Five, such as a Six Factor model in China (Cheung et al. 2001). Similar indigenous replication studies in other non-western cultures are needed as they may produce new factor structures and inform us about the universality of the Big Five.

9.7.4 ITC guidelines

The ITC guidelines directed much of the adaptation process developed and applied in this study. Since cross-cultural research in international by nature, the ITC guidelines provide “neutral grounds” that cross-cultural researchers can rely on in their work. Although the guidelines tap on to the different areas that need to be considered during the process, it is difficult to judge how they could be achieved. For example, guideline D.1 states that

> “Instrument developers/publishers should insure that the translation/adaptation process takes full account of linguistic and cultural differences among the populations for whom the translated/adapted versions of the instrument are intended.”

This is absolutely essential for achieving equivalence between trans-linguistic versions of tests, however, there are many methods that researchers can use and assume that it has been achieved. Some for example might argue that professional translators are trained to take into account linguistic and cultural aspects in their work. However, as discussed in the section 6.2.1 and 6.2.2, translation of a sentence depends on the context it is in (Newmark, 1996). But in personality tests, it is difficult to rely on the context for achieving equivalence in meaning, as the sentences that make up a questionnaire are independent from each other.

The presence of papers that discuss the implications of using different methods for adapting tests (Hambleton, 1993) and illustrate the ITC guidelines with examples (such as Hambleton and van de Vijver, 1996; Hambleton, 2001) make it possible to understand an apply these guidelines.

9.7.5 Final conclusion and future directions

Dangers of relying on a simple translations and the gravity of the implications of using adapted tests without sufficiently rigorous adaptation have been highlighted throughout this Thesis. Evidence from this set of studies demonstrated that rigorous test adaptation, although essential, does not always guarantee the reproduction of cultural versions of a test. Orpheus was adapted using a very thorough approach and a triangulation of techniques. Yet it was not possible to obtain comparable versions in different languages due to technical problems related to Orpheus as a test.

For future studies, it is important for researchers to be able to scrutinise the test of interest before adapting it into other languages and cultures. There are several resources that can be used to do so, such as visiting certain websites, checking the accreditation of questionnaires and consulting the technical manual available from the test publishers (McDowall, Rust, & Daouk, 2005). However, accreditation processes and criteria used for establishing the validity and reliability of tests in one culture need to be revised, unified, and internationalised. Cross-cultural research will continue to involve different languages, cultures, and countries. Common standards for accepting validity and reliability of tests and equivalence between their trans-linguistic versions are a natural follow up to the guidelines that have being developed for using tests internationally (such as the ITC test adaptation and computer-based testing guidelines).

Although there are many advances in test adaption, in practice, we are ahead of ourselves because the knowledge and expertise in assessing the comparability between tests has not been fully disseminated across disciplines and between academia and practice. Nevertheless, our need for tests is only increasing as discussed in chapter 1. Developing indigenous tests to serve in one culture is ideal; however, it is not sufficient if our aim is to compare individuals across cultures. As previously discussed, parallel adaptation of tests is not always sufficient and more focus needs to be diverted towards parallel development of tests across languages and cultures. This shift has recently started to take place in academia and practice, but it needs to be directing the future of research across cultures.

Bibliography

Anastasi, A. (1988). *Psychological testing (6th ed.).* New York: MacMillan.

Anderson, N., & Shackleton, V.J. (1993). *Successful selection interviewing.* Oxford: Blackwell.

Arffman, I. (2007). *The problem of equivalence in translating texts in international reading literacy studies: a text analytic study of three English and Finnish texts used in the PISA 200 reading test.* Juvaskyla: University of Jyvaskyla, Institute for Educational Research

Azuma, H., & Kashiwagi, K. (1987). Descriptions for an intelligent person: a Japanese study. *Japanese Psychological Research, 29*, 17-26.

Ballenger, J. F., Caldwell-Andrews, A., & Baer, R. A. (2001). Effects of positive impression management on the NEO PI-R in a clinical population. *Psychological Assessment, 13*, 254-260.

Baron, H and Janman, K. (1996). Fairness in Assessment Centre. *International Review of Industrial and Organisational Psychology, 11*, 61-114

Baron, H., & Janman, K. (1996). Fairness in assessment centre. *International Review of Industrial and Organisational Psychology, 11*, 61-114.

BarOn, R. (2002). *The BarOn emotional quotient inventory technical manual.* Canada: MHS Inc.

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

Barrick, M.R., & Mount, M.K. (1993). Autonomy as a moderator of the relationship between the big five personality dimensions and job performance. *Journal of Applied Psychology, 78*, 11-118.

Barrick, M.R., & Mount, M.K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81*, 261-272.

Bartram, D. (2001). Predicting competency dimensions from components: A validation of the two-step process. *Internal Technical Research Report.* Thames Ditton: SHL group plc.

Bartram, D. (2005). The changing face of testing. *The Psychologist, 18*, 11, 666-668.

Bartram, D., & Brown, A. (2004). Online testing: mode of administration and the stability of OPQ32i. *Internationàl Journal of Selection and Assessment, 12,* 278–284.

Berry, J.W., Poortinga,Y.H., & Pandey, J. (Eds) (1997). *Handbook of cross-cultural psychology (2nd ed.)*. London: Allyn & Bacon.

Bertua, C., Anderson, N., & Salgado, J. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78,* 3, 387-409.

Brislin, R. W. (1980). Translation and content analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology, 1*, 389-444. Boston: Allyn & Bacon.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.). *Field methods in cross-cultural research*, 37-164. Newbury Park, CA: Sage.

Brislin, R.W., Looner, W.J., & Thorndike, R.M. (1973). *Cross-Cultural Methods.* New York: Wiley.

Brown, P., Green, A., and Lauder, H. (2001). *High skills: globalisation, competitiveness, and skill formation.* Oxford: Oxford University Press.

Butcher, J. N. (2004). Personality assessment without borders: Adaptation of the MMPI-2 across cultures. *Journal of Personality Assessment, 83*, 90-104.

Butcher, J. N., Cheung, F. M., & Lim, J. (2003). Use of comprehensive personality inventories with Asian populations. *Psychological Assessment, 15,* 248-256.

Bureau of European and Euroasian studies (2004). Retrieved in December 2004 from: http://www.state.gov/p/eur/

Byrne, B.M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34,* 2, 155-175.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Byrne, B.M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models*. New York: Springer.

Byrne, B.M. (1993). The Maslach Burnout Inventory: Testing for factorial validity and invariance across elementary, intermediate, and secondary teachers. *Journal of Occupational and Organisational Psychology, 66,*°197-212.

Casillas, A., and Robbins, S.B. (2005). Test adaptation and cross-cultural assessment from a business perspective: issues and recommendations. *International Journal of Testing, 5,* 5-21.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response set in cross-cultural research using structural equation modelling. *Journal of Cross-Cultural Psychology, 31*, 187–212.

Cheung, G. W., & Rensvold, R. B. (2000). Testing measurement invariance using critical values of fit indices: A Monte Carlo study. *Research Methods Forum.* Retrieved in June 2005 from http://www.aom.pace.edu/rmd/cheung_files/cheung.htm.

Cheung, F. M., Leung, K., Fan, R. M., Song, W. Z., Zhang, J. X., & Zhang, J. P. (1996). Development of the Chinese personality assessment inventory. *Journal of Cross-Cultural Psychology, 27,* 143-164.

Cheung, F. M., Leung, K., Zhang, J. X., Sun, H. F., Gan, Y. Q., Song, W. Z., & Xie, D. (2001). Indigenous Chinese personality construct: Is the Five Factor Model complete? *Journal of Cross-Cultural Psychology, 32*, 407-433.

Cheung, F. M., Cheung, S. F., & Zhang, J. X. (2004). Convergent validity of the Chinese personality assessment inventory and the Minnesota multiphasic personality inventory-2: Preliminary findings with a normative sample. *Journal of Personality Assessment, 82*, 92-103.

Cheung, F. M. (2004 a). Use of Western- and indigenously- developed personality tests in Asia. *Applied Psychology: An International Review, 53*, 173-191.

Cheung, F. M. (2004 b).*Transaltion East and West.* Paper presented at the ICP conference in Beijing.

Cober, R.T., Brown, D.J, Blumental, A. J., Doverspike, D. and Levy, P.E. (2000). The quest for the qualified job surfer: It's time the public sector catches the wave. *Public Personnel Management. 29*, 479-494

Cohen, Y., Gafni, N., & Hanani, P. (2007). *Translating and adapting a test, yet another source of variance; the standard error of translation.* A paper submitted to the annual meeting of the IAEA Baku, Azerbaijan.

Costa, P.T., Jr. and McCrae., R. R. (1992). *The Revised NEO personality inventory and NEO Five Factor inventory: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Crabtree BF, Miller WL, eds. (1999). *Doing qualitative research (2nd ed.).* Newbury Park, California: Sage Publications.

Cronbach, L.J. (1990). *Essential of psychological testing (5th ed.).* New York: Harper and Row Publishing.

Daouk-Oyry, L. & Rosinski, P. (in press). Understanding cultural differences. In S. Palmer & A. McDowall (Eds), *Putting people first: the interpersonal aspects of coaching.*

Daouk, L., Rust, J., & McDowall, A. (2005). Testing across languages and cultures: challenges for the development and administration of tests in the internet era. *Selection and Development Review, 21*, 4, 11-1.

DataMonitor (2004). Retrieved December 2004 form: http://www.datamonitor.com/.

Den Hartog D.N., House R.J., Hanges, P J., .Ruiz-Quintanilla, S.A., & Dorfman, P.W. (1999). Culture specific and cross-culturally generalizable implicit leadership theories Are attributes of charismatic/transformational leadership universally endorsed? *The Leadership Quarterly, 10, 2*, 219-256

Deslisle, J., & Woodsworth, J. (eds) (1995). *Translators through history.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

DeMaio, T.J., Rothgeb,J., & Hess, J. (1998). *Improving survey methods through pretesting*. Washington, DC: US Bureau of the Census.

Dorfman, P.W. (1996). International and cross-cultural leadership research. In B.J. Punnet & O. Shenkar (Eds). *Handbook for international management research*, 267-349. Oxford: Blackwell.

Downing.S.M., (2002).Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education, 37*, 235-241.

Downing, B.T., Bogoslaw, L.H., and Juntos, H. (2002). Effective patient-provider communication across language barriers: A focus on methods of translation. Claremont, CA: Hablamos Juntos.

Ferguson, D. (1997). In as a prefix. *Random House.* Retrieved February 15, 2007, from: http://www.randomhouse.com/wotd/index.pperl?date=19970604

Field, A.P. (2005). *Discovering Statistics Using SPSS (2nd Ed).* Sage: London.

Fife-Schaw (2006). Levels of measurement. In G.M., Breakwell, S. Hammond, C, Fife-Schaw, & J.A., Smith, Research methods in psychology (3rd ed). London: Sage..

Fife-Schaw (2006). Questionnaire design. In G.M., Breakwell, S. Hammond, C, Fife-Schaw, & J.A., Smith, Research methods in psychology (3rd ed). London: Sage.

Furnham, A. (1997). Knowing and faking one's Five-Factor personality score. *Journal of Personality Assessment, 69, 1,* 229-243.

Furnham, A., Forde, L., Ferrari, K. (1999). Personality and work motivation. *Personality and Individual Difference, 26,* 1035-1043.

Furnham, A., Steele, H., and Pendleton, D. (1993). A psychometric assessment of the Belbin Team-Role Self-Perception Inventory. *Journal of Occupational and Organizational Psychology, 66,* 245-257

Geisinger, K.F. (1994). Cross-cultural normative assessment: Transition and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*, 304–312.

Grammar. (n.d.). Dictionary.com Unabridged (v 1.1). Retrieved October 17, 2007, from Dictionary.com website: http://dictionary.reference.com/browse/grammar

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational perspective. *Academy of Management Review, 18*, 4, 694-734.

Glaser, B.G., & Strauss, A.L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Chicago: Aldine.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000, April). *Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is Large*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international Assessment. *Language Testing, 20,* 2, 225-240.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9*, 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*, 229-244.

Hambleton, R. K. (2001). The next generation of the ITC test translation and application guidelines. *European Journal of Psychological Assessment, 17*, 3, 164-172.

Hambleton, R.K. (2002). Adapting achievement tests into multiple languages for international assessments. In A.C. Porter, & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 58–79). Washington: National Academy Press.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.). *Adapting educational and psychological tests for cultural assessment* (pp. 3–39). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Li, S. (2004). Effective implementation of the International Test Commission guidelines for adapting tests. Paper presented at the ICP conference, Beijing, China.

Hambleton, R. K., Merenda, P. F., Spielberger, C. D. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah: Lawrence Erlbaum Associates.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*, 1-30.

Hambleton, R. Rodgers, J (1995). Item bias review. *Practical Assessment, Research & Evaluation, 4,* 6. Retrieved September 1, 2005 from http://PAREpnline.net/getvn.asp?v=4&n=6.

Hambleton, R.K., Yu, J., & Slater, S.C. (1999). Field-test for the ITC guidelines for adapting educational and psychological tests. *European Journal of Psychological Assessment, 15, 3*, 270-276..

Harris, P. & Moran, R. (1996). European leadership in globalization. *European Business Review, 96*, 2, 32-41.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*, 238-247.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*, 9-24.

Ho, D. Y. F. (1996). Filial piety and its psychological consequences. In M. H. Bond (Ed.), *Handbook of Chinese psychology* (155-165). Hong Kong: Oxford University Press.

Hofstede, G. (1983). The Cultural Relativity of Organizational Practices and Theories. *Journal of International Business Studies, 14*, 75-89.

Hofstede, G. (2003). *Geert Hofstede cultural dimensions*. Retrieved November, 2006, from http://www.geert-hofstede.com/

Hofstede, G.J. (2006). What did GLOBE really measure? Researchers' minds versus respondents' minds. *Journal of International Business Studies, 37*, 6, 882-896.

Hogan, R. (2008). *What is personality*. Retrieved February, 2008, from: http://www.hoganassessment.com/personality_research/personality_what_is.aspx

Holland, P. W., & Thayer, D. T. (1988). *Differential item performance and the Mantel-Haenszel procedure.* In H. Wainer & H. I. Braun (Eds.). *Test validity* (129-145). Hillsdale, NJ: Erlbaum.

Hollensen S. (2nd eds) (2001). *Global marketing-A market-responsive approach.* Harlow: Financial Times/ Prentice Hall.

House, R., Javidan, M. & Dorfman, P. (2001). Project GLOBE: An introduction. *Applied Psychology: An International Review, 50*, 4, 489–505.

House, R., Javidan, M., Hanges, P., & Dorfman, P. (2002). Understanding cultures and implicit leadership theories across the globe: An introduction to project GLOBE. *Journal of World Business*, 37, 1, 3-11.

Hughes, K.A., & DeMaio, T.J. (2002). Hughes, K., & DeMaio, T. (2001). *Does this question work? Evaluating cognitive interview results using respondent debriefing questions.* Paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, Florida.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20,* 296-309.

International Telecommunication Union (2005). http://www.itu.int/ITU-D/icteye/Reporting/ShowReportFrame.aspx?ReportName=/WTI/InformationTechnologyPublic&RP_intYear=2005&RP_intLanguageID=1 Extracted 2/14/2007

Jeanrie, C, & Bertrand, R. (1999). Translating tests with the International Test Commission's Guidelines: Keeping validity in mind. *European Journal of Psychological Assessment, 15*, 277-283.

Jensen, A (1980). Bias in Mental Testing. New York: Free Press in Kline, P. (ed) (1993). *The Handbook of Psychological Testing.* New York: Routledge

Jodoin, M. G., & Gierl, M. J. (2001). Type-one error and power rates using an effect size measure with the logistic regression for DIF detection. *Applied Measurement in Education, 14,* 329-49.

Kabasakal, H., & Bodur, M. (2002). Arabic cluster: A bridge between east and west. *Journal of World Business, 37*, 49-54.

Kanji, G. K. (2006). *100 statistical tests (3rd ed.).* London, UK: Sage.

King, N., Thomas, K. & Bell, D. (2003). An out-of-hours protocol for community palliative care: practitioners' perspectives. *International Journal of Palliative Nursing 9,* 7, 277-282.

Kleinman, A. (2004). Culture and depression. *The New England Journal of Medicine, 351,* 10, 951-953.

Kline, P. (1986). *A handbook of test construction: introduction to psychometric design*. London: Methuen.

Kline, P. (1993). *Personality: The psychometric view.* New York: Routledge.

Krishnakumar, A., C. Buehler, & Barber, B.K. (2004). Cross-ethnic equivalence of socialisation measures in European American and African American families. *Journal of Marriage and Family, 66*, 3, 809-820.

Lalwani, A. K., Shavitt, S., & Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding? *Journal of Personality and Social Psychology, 90*, 1, 165-178.

Landis, J. R. & Koch.G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Lee, R. M., Falbo, T., Doh, H. S., & Park, S. Y. (2001). The Korean diasporic experience: Measuring ethnic identity in the United States and China. *Cultural Diversity & Ethnic Minority Psychology, 7*, 207–216.

Lievens, F.,& Harris, M. M. (2003). Research on internet recruitment and testing: Current status and future directions. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology,* Vol 18, (131–165). Chichester: Wiley.

Little, T.D. (1997). Mean and covariance structures (MACS) analyses of cross cultural data: Practical and theoretical issues. Multivariate Behavioural Research, 32, 53-76

Jöreskog, K., & Sörbom, D. (1999). *LISREL 8.30.* Chicago: Scientific Software International Inc.

Maamouri, M. (1998). Language education and human development: Arabic diglossia and its impact on the quality of education in the Arab region. Paper presented at The World Bank Mediterranean Development Forum, Marrakesh. Philadelphia: University of Pennsylvania International Literacy Institute.

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. *Psychological Bulletin, 97*, 562-582.

Martin, M.O., Mullis, I.V.S. and Chrostowski, S.J.(2004). *TIMSS 2003 Technical Report: Findings From IEA's Trends in International Mathematics and Science Study at the Eighth and Fourth Grades.* Chestnut Hill, MA: Boston College.

McCrae, R.R., & Costa, P.T. Jr. (2003). *Personality in adulthood: A five-factor theory perspective.* New York: Guilford Publications.

McCrae, R. R., & Costa, P. T. Jr. (1997). Conceptions and correlates of openness to experience. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 825-847). San Diego, CA: Academic Press.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599–616.

McDowall, A., Rust, J., &. Daouk, L (2005). Navigating the Test Maze with Confidence. *Selection and Development Review, 21*, 4, 11-13.

Meckler, M., & Mullen, M. (1997). Enhancing the validity of comparative research: testing for metric equivalence across populations. In L.N. Dosier & J. B. Keys (Eds.), *Academy of Management Best Papers Proceedings, 57* (p. 566). Georgia Southern University: Faculty Research Services, College of Business.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525–543.

Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational Measurement (3rd Ed)* (13-104). New York: Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*, 5–8.

Millsap, R.E. (2007). Structural equation modelling made difficult. *Personality and Individual Differences*, 42, 5, 875-881

Morphology. (n.d.). dictionary.com. retrieved October 17, 2007. from dictionary.com website:http://dictionary.reference.com/browse/morphology

Muthén, L. K., & Muthén, B. O. (1998). *Mplus: The comprehensive modelling program for applied researchers*. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2008). *EFA and CFA lectures.* Retrieved June2nd, 2008, from: http://www.ats.ucla.edu/videos/stats/2008/stats_part_1_2008_small.mov

Newmark, P (1996). Paragraphs in translation. In G.G. Anderman & M.A. Rogers. *Words words words: the translator and language learner*. Bristol, UK: Multilingual Matters.

Nida, E.A., & Taber, C.R. (1982). *The theory and practice of translation.* Leiden: EJ Brill

Nord, C. (1997). *Translating as a purposeful activity. functionalist approaches explained.* Manchester: St. Jerome

Nunally, J C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Oakland, T. (2004). Use of Educational and Psychological Tests Internationally. *Applied Psychology*, *53, 2,* 157-172

Okasha, A., el Akabawi, A.S., Snyder, K.S., Wilson, A.K., Youssef, I., &

elDawla, A.S. (1994). Expressed emotion, perceived criticism, and relapse in depression: a replication in an Egyptian community. *American Journal of Psychiatry, 151, 7,* 1001-1005.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11, 2,* 245-269.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81,* 660-679.

Organisation of Economic Co-operation and Development (OECD) (2006). *The programme of international student Assessment (PISA) executive summary.* Retrieved March 3, 2008, from http://www.oecd.org/ dataoecd/15/13/39725224.pdf

Osterlind, S. J., Miao, D.M., Sheng, Y.Y., & Chia, R. C. (2004). Adapting item format for cultural effects in translated tests: Cultural effects on construct validity of the Chinese versions of the MBTI. *International journal of testing, 4, 1,* 61-73.

O'Sullivan, D. (2006). Meta-analysis. In G.M., Breakwell, S. Hammond, C, Fife-Schaw, & J.A., Smith, *Research methods in psychology* (3rd ed). London: Sage.

Pelham, B.W. (1999). *Conducting Research in Psychology: Measuring the*

Weight of Smoke. Pacific Grove: Brooks/Cole Pub.

Peterson, D.B. (2007). Executive coaching in cross-cultural settings. *Consulting Psychology Journal: Practice and Research. 59, 4*, 261-271.

Petrides, K. V., & Furnham, A. (2003). Trait emotional intelligence: Behavioural validation in two studies of emotion recognition and reactivity to mood induction. *European Journal of Personality, 17*, 39–57.

Petrides, K.V., Pita, R., & Kokkinaki, F (2007). The location of trait emotional intelligence in personality factor space. *British Journal of Psychology, 98, 2*, 273-289

Redline, C., Smiley, R., Lee, M., DeMaio, T., & Dillman, D. (1998). *Beyond concurrent interviews: an evaluation of cognitive interviewing techniques for self- administered questionnaires.* Retrieved September 15, 2007, from US Census Bureau Website: http://www.census.gov/srd/papers/pdf/sm98-06.pdf.

Reiss, K. and Vermeer, H.J. (1984). Grundlegung einer allgemeinen Translationstheorie [*Groundwork of a general Theory of Translation*].Tübingen: Niemeyer.

Robertson, I. T., & Smith, M. (2001). Personnel selection. *Journal of*

Occupational and Organisational Psychology, 74, 4, 441-472.

Rosinksi, P. (2003). *Coaching across cultures: New tools for leveraging national, corporate and professional differences*. London: Nicholas Brealey Publishing.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634–644.

Rothgeb, J., Willis, G., & Forsyth, B. (2001). *Questionnaire pretesting methods: do different techniques and different organizations produce similar results.* Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Montreal, Canada.

Rust, R., & Golombok, S. (2001). *Modern psychometrics: The science of psychological assessment (2nd ed).* London: Routledge.

Rust, R., & Golombok, S. (1999). *Modern psychometrics: The science of psychological assessment.* London: Routledge.

Rust, J. (1996). Orpheus Handbook. *The Psychological Corporation*, London and San Antonio.

Ryan, A. M., McFarland, L., Baron, H. & Page, R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology, 52*, 359-391.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C. & de Fruyt, F. (2003). International Validity Generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology, 56,* 573-605.

Shackleton, V. & Newell, S. (1997). International Recruitment & Selection. In. N. Anderson, & P. Herriot, (Eds).. *International Handbook of Selection & Assessment.* Chichester: Wiley.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmit, M. J. & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78,* 966-974.

Society for Industrial and Organisational Psychology (SIOP), Inc (2003). *Principles for the validation and use of personnel selection procedures.* Retrieved July 6, 2007, from http://www.siop.org/_Principles/principles.pdf

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117-138). Hillsdale, NJ: Lawrence Erlbaum.

Sireci, S.G., Patsula, L., & Hambleton, R.K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R.K. Hambleton, P.F. Merenda, & C.D. Spielberger (Eds), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-116). Hillsdale, NJ: Erlbaum.

Slocum, S. L., Gelin, M. N., & Zumbo, B. D. (2003). Statistical and graphical modelling to investigate differential item functioning for rating scale and Likert item formats. In B. D. Zumbo (Ed.), *Developments in the Theories and Applications of Measurement, Evaluation, and Research Methodology Across the Disciplines,* Vol. 1. Vancouver: Edgeworth Laboratory, University of British Columbia.

Smith, M. & Robertson, I.T. (1993). *The Theory & Practice of Systematic Personnel Selection*. London: The Macmillan Press Ltd.

Snell-Hornby, M. (1988). *Translation Studies. An Integrated Approach.* Amsterdam: Benjamins.

Snijkers, G. (2003). Cognitive Laboratory Experiences and Beyond: Some ideas for future research. ZUMA-Nachrichten Spezial Band 9, *Questionnaire Evaluation Standards*,190-216.

Sternberg, R. J., Nokes, K., Geissler, P. W., Prince, R., Okatcha, F., Bundy, D. A., & Grigorenko, E. L. (2001). The relationship between academic and practical intelligence: A case study in Kenya. *Intelligence, 29*, 401–418.

Sulaiman, S.O.Y., Bhugra, D., & De Silva, P. (2001). The development of a culturally sensitive symptoms checklist for depression in Dubai. *Journal of transcultural psychiatry, 38, 2,* 219-229.

Sulaiman, S.O.Y., Bhugra, D., & De Silva, P. (2001). Perceptions of depressions in community sample in Dubai. *Journal of transcultural psychiatry, 38, 2,* 219-229.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement,* 27, 361-370.

syntax. (n.d.). Dictionary.com Unabridged (v 1.1). Retrieved October 17, 2007, from dictionary.com. website:http://dictionary.reference.com/browse/syntax

Terman, L. M. & Merrill, M. A. (1937) *Stanford Binet Intelligence Scale Manual for the 3rd Edition Form L-M,* Harrap: London.

Tourangeau, R. (1984). Cognitive science and survey methods: A cognitive perspective, IN Jabine, T. Straf, M. Tanur, J. and Tourangeau, R. (eds.) *Cognitive aspects of survey methodology: Building a bridge between disciplines.* Washington, DC: National Academy Press.

Triandis, H. C. (1980). Values, attitudes, and interpersonal behavior. In H. E. Howe & M. M. Page (Eds.), *Nebraska Symposium on Motivation*, 1979 (pp. 195-259). Lincoln: University of Nebraska Press.

Triandis, H.C. (Ed) (1980). Handbook *of cross-cultural psychology*. London : Allyn & Bacon.

Trubek D. and Mosher J. (2003). New Governance, Employment Policy and the European Social Model. IN Zeitlin J., Trubeck D.(eds*). Governing Work and Welfare in a New Economy : European and American Experiments.* Oxford University Press, Oxford.

Tupes, E.C. & Christal, R.C. (1961). *Recurrent personality factors based on trait ratings (Tech. Rep.*). Lackland Air Force Base, TX: USAF.

Urist, J. (1977). The Rorschach test and the assessment of object relations. *Journal of Personality Assessment*, 41, 3-9.

van de Vijver, F. R., & Jeanrie, C. (2004). *Assessing structural and metric* *equivalence: A case study*. Paper presented in the ICP conference Beijing, China.

van de Vijver, F. R., & Hambleton, R. K. (1996). Translating tests: some practical guidelines. *European Psychologist, 1, 2,* 89–99.

van de Vijver, F. and Poortinga, Y. (2002). Structural equivalence in multilevel research. *Journal of Cross Cultural Psychology, 33*, 141-156.

van de Vijver, F.J.R., & Poortinga, Y.H. (2005). Conceptual and methodological issues in adapting tests. In R.K, Hambleton, P.F., Merenda, & C.D. (Eds). *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah: Mawrence Erbaum Associates.

van de Vijver, F J.R. & Leung, Kwok (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31, 1,* 33-51.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Newbury Park, CA: Sage.

van de Vijver, F., & Phalet, K. (2004). Assessment in multicultural groups: The role of acculturation. *Applied Psychology: An International Review, 53, 2,* 215-236

van de Vijver, F., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: *An overview. European Review of Applied Psychology, 47,* 263-279.

van de Vijver, F.J.R., & Tanzer, N.K. (2004, reprint). Bias and equivalence in cross-cultural assessment: an overview. *European Review of Applied Psychology*, 54, 2, 119-135.

Wendt, A., & Worcester, P. (2000). The National Council Licensure Examinations/ Differential Item Functioning Process. *Journal of Nursing Education, 39, 4,* 185-187.

Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61,* 275–290.

Willis, G.B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: Sage Publications.

Woodruffe, C (1991). Competent by any other name. *Personnel Management*, p. 30-33.

Wu, A.D., Li, Z., & Zumbo, B.D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: a demonstration with TIMSS data. *Practical Assessment Research & Evaluation, 12, 3,* 1-26.

Yang, K. S. (2000). Monocultural and cross-cultural indigenous approaches: The royal road to the development of a balanced global psychology. *Asian Journal of Social Psychology*, 3, 241-263.

Yeganeh, H., Su, Z., & Chrysostome, E.V.M. (2004). A Critical Review of Epistemological and Methodological Issues in Cross-Cultural Research. *Journal of comparative international management. 7, 2*, 66-86

Zucker, S., Miska, M., Alaniz,L., & Guzman, L. (2005). *Transadaptation: Publishing assessments in world languages* (Assessment Report). San Antonio, TX: Pearson Education.

Zumbo B.D.. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modelling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defence.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses?: Implications for translating language tests. *Language Testing, 20*, 136-147.

Zumbo, B.D. (July , 2006). *Psychometric Methods for Investigating DIF and Test Bias During Test Adaptation Across Languages and Cultures.* Workshop presented at the 5th International Test Commission Conference, Brussels, Belgium.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao and S. Sinharay (Eds.). *Handbook of statistics, Psychometrics*, 26,:45-79. Elsevier Science B.V.: The Netherlands.

Appendix

Appendix 1: Orpheus major scales

Major Scales	Positive Items	Negative Items
Fellowship		
Authority		
Conformity	**Copyrighted information**	
Emotion		
Detail		

Appendix2: Orpheus minor scales

Minor Scales	Positive Items	Negative Items
Proficiency		
Work-Orientation		
Patience		
Fair-mindedness		**Copyrighted information**
Loyalty		
Disclosure		
Initiative		

Appendix 3: Email detailing the aims of the translations

Dear xxx

Thank you for agreeing to translate Orpheus into xxx. Orpheus is a work-based personality test designed to assess preferences of employees in the workplace.

Each question is designed to serve a certain psychological purpose and the wording of each question has been carefully thought of. Therefore is extremely important for us to make sure that the translated version is as similar as possible in wording to the original one. However, we do understand that this might not be possible for all questions since some connotations cannot be fully captured without adapting the sentence to serve the same meaning in the target culture. Should you require to do so, please ensure that you use wording as close as possible to the original English version and also make sure that the sentence still follows the same grammatical format as the original one. For example, if a sentence is written in the passive tense, please make sure it stays so in the target language. Also, if the sentence is written in a negative way (i.e. we are not uncomfortable) please make sure that you do not reverse it in the translation (i.e. we are comfortable).

Thank you for your effort in advance and I look forward to receiving the transited version from you. In case you have any questions about the meaning of any terminology or expression, please do not hesitate to contact me at any time.

All the best,

Lina

Appendix 4: Version 2 Arabic

Copyrighted information

Appendix 5: Version 2 Chinese

Copyrighted information

Appendix 6: Version 2 Spanish

Copyrighted information

Appendix 7: Potentially problematic items

Copyrighted information

Appendix 8: Brief for judges

Dear judges,

Thank you for taking part of the dyads and triads. As part of my PhD project, we are currently working on developing five multi-lingual versions of a work-based personality questionnaire-Orpheus- into English, Arabic, French, Mandarin, and Spanish. In order for these tests to be comparable/equivalent in all languages, they need to be equivalent linguistically, culturally and metrically.

We have already developed the questionnaire in all those languages and the process has involved several speakers of these five languages in order to increase the accuracy of the translation. However, in order to be even more sure that the questions are well translated, we gave the questionnaire to Mandarin speakers, in English and in Mandarin. After careful statistical analysis, we found that there are some items that are answered differently when presented in different languages. So, in this interview, we would like to explore how each one of these questions is understood in Mandarin so that we can build a better understanding of the culture, language, and comparability between the English and Mandarin versions of each question.

Although your role is central to this exercice, you do not need to worry about what happens next as I will be facilitating it and making sure you are focusing on the essential parts. However, I would like to inform you that the main aim is to ensure that the wording is similar between the two language versions. There are cases where wording cannot be exactly the same as in English, but I would like you to focus on these areas, highlight them to me in English and discuss amongst yourselves whether it is the most appropriate word to be used in this context. It is extremely important that you tell me whever you find words that are not exactly the same so I can explain to you the psychological purpose the questions is trying to chieve so you can judge the equivalence between the two languages. Also, I would encourage you to highlight any wording or issues related to the questions that might be culturally different between your culture and the UK cultures.

This sounds like a lot of information but once we start, the meeting will run smoothly and you will get used to the process. Please ask me about anything that is not clear for you.

Appendix 10: Reasons for Chinese amendments based on V2

4: never ever is not mentioned in the translation and also, the word in red means "I'm not willing" so it was replaced by "I wouldn't" which is closer to the English one.
6: changed the order of two words because it sounds better. The first order means "love" whereas when it's changed, it could describe any type of feeling, not necessarily love.

8: 3 words were split up 能够 and 更 and other words were put between them to explain in detail what is meant by the item. If those words are added, then it makes it easier and quicker to understand.

9: 觉得 and 感觉 both mean "I fell" but the second one is a more formal way of saying the word.

10: the old translation means "income" for the person rather than for the organisation because in mandarin there is no direct word that "profit" can be translated into.

13: 以往 were added because they mean "the traditional way" or "old way". Also the new sentence now means "the way we do things" rather than "the method". Since sentences are made up of a combination of words then sometimes it is better to add a couple of words even if they are not in the original item but only because the sentence will be better understood that way.

14: 如果 means "if" and was added at the beginning of the sentence. Although this is not in the English version, it is essential to add it to the sentence because in mandarin it means that you are asking people to imagine the situation and what decision they would make. Also 影响 which means "difficult to influence" was replaced by 改变 which means "no need to influence" because in this context, the first one doesn't sound natural.

15: the sentence meant that if you follow the formal procedures then things can be done better, but in fact the item in English means that if we by-pass formal procedures things become better so it's the opposite meaning.

17: 鲁莽 was changed into 挑衅 because it means "rude" and the problem is that the sentence means "If I'm treated in a rude way" and in Mandarin this is not the right word to use because it is an adjective that refers to a "person" not a "way". So it was replaced by a word rather than an adjective, which mean "to challenge but not in a friendly way".

18: 惯例 was changed into 行政 the first one means "traditional rules" whereas the replacement means "administration".

19: the 2 statements are similar but were replaced with a more comfortable one that sounds better in mainland china and also because this one is word by word translation but in mandarin you don't really say it this way.

21: 长者 this word refers to: uncle, teacher, people who are older than me and I respect. So it was replaced by 上级 which refers to people higher up in the organisation than you.
22: replaced with a more mandarin friendly one. And added 极度 because it means "really" thus emphasizing hoe attentive to detail one is

24: 发现 changed into 觉得 because the first one means discover or find out, so it was replaced by feel or realise and is intended to replace “I sometimes feel”. The first one is much stronger in this context. Also 被完全赏识到 was replaced by 完全得到赏识 although there isn’t much difference it’s only a matter of changing the order. The first one is a direct translation of the English while maintaining the same structure of the English, so it was rearranged to sound more acceptable in Mandarin.

28: 最高优先权 was replaced by 最应该被优先考虑的 since the first one might mean “top priority” but also “a right” so the replacement explains it in details so the candidate taking the test later will not get confused.

30: change was literally translated as change but here it meant money

33: 事情 was replaced by 工作上的事 because the first means “things” whereas the second one is specifically the things I do at work so in the first one there wasn’t a mention of work.

35: 一次 means “once” and there is no mention for it in the English version so it was removed.

36: 得不到应得到的休息 means that “I never have enough rest” so it was replaced by 有点闲不住 which means that “I cannot stay without doing anything”. Also 感觉 was added and it means “feel” to make easier to understand.

39: 神情 replaced by 样子 although they both mean “look and expression” but the second one also means “manner” and is more commonly used.

41: 在乎 means “care” so it was replaced by 介意 which means “mind” although the English one is care, “mind” was thought to be a better way of expressing this sentence. Also 很爱出风头 was replaced by 太激进 because the first sentence means that the person is so desperate to prove him/herself so the replacement means pushing people to get a job done which explains the English word “pushy” in a more suitable way.

42: need to review that one

43: 高招 was replaced by 妙计 although both mean “good idea” th4 second is more commonly used and sounds more comfortable and formal than the first one

45: 幼稚 replace by 天真. Both of them can mean “naive2 but the first one Is more “childish” whereas the second one is more “naïve”. Also the order of the sentence was changed because it sounds better.

46: this item was translated into “there have been days where I was so upset because I didn’t do any work”. This was changed to “there have been days that I was so excited that I couldn’t sit down and work”.

49: 管理 is management, so it was replaced by 行政 because it means “administration”. Also 无视 which means “ignore” was replaced by 抵制 which means “against”.

51: 被设定了 was added to make the sentence passive like the original English one. 最终期限 and 完成期限 both mean deadline but the second one is the type of deadline that is related to a task or job so even though the word work was already mentioned in the sentence, you need to use the second one because it is more work related and should be used in this context. 很 means “very or extremely” so it was changed to 最 which means “most” so the combination of this word that was replaced and the 2 after it will lead to “essential” whereas if it was kept as it was it would be “important” but nor necessarily “essential” so the strength of the word will decrease.

52: 如果 was taken away because it means “if” and does not exist in the English version and there is no need for it to be in the Chinese one either.

56: 喜欢 means “like” whereas 更愿意 means “prefer” hence why it was chosen.

57: 发脾气 means “angry” so it was replaced by 情绪失控 which means “loose temper”.

58: 内心的 was added to emphasise that it’s “gut feeling” because in the original translation it’s only feeling which doesn’t reflect the English translation properly.

59: 展示 means “demonstrate” to people how to do things but 教 means “teach” people how to do things so the argument is that the meaning intended from the English is not only about showing but also teaching.

62: 它们自己的时间出现 means “they will appear in their own time” was replaced 特定的时间里自然出现 means “they will appear naturally by themselves at some point in the future”. Although the first one seems to be more appropriate when translated to English, however the second is what is common and more normal to use in Chinese and the first seems unnatural to say.

63: the translation means that “I am the kind of person who can fuse in easily in social life” so it was replaced by a sentence that means “I am the kind of person who can easily be the centre of the party and get everybody’s attention”.

69: both sentences are similar but some words were removed to make it simpler to understand. In some situations you need to explain things in detail in order for people get the meaning intended, however in some other situations, when the meaning can be clarified with fewer words, then it would be better to keep things short and simple.

71: typing mistake

72: 他们 “they” and 人 “people”: the order of these words was changed because one should identify the person or subject first and then use “they” to refer to them.

76: change “delighted” into “I don’t mind” because it explains what the English item means in a better way. “I am happy leaving the necessary arrangement” is not about joy, it’s more about not having a problem with something.

78: rearranged the order of words because the structure of the sentence was not changed in the Chinese version although it doesn’t sound right.

79: 一个项目的更广泛含义 means “the general meaning of a task” so it was replaced by 一个项目在其大方向上的工作 which means “the big picture of a job” which reflects the English item “wider implications” better that the first one. Also there is an equivalent for implication which is 包含 but it cannot be used in this context because it will be understood as contain/embody/include

82: 最高水平 means “best level” so it was replaced by 最好的工作成果 which means “best work”.

85: working with my hand was translated to “ I like working by myself” so it was replaced by a sentence that means “I specially enjoy working with my hands”.

86: the word used for reputation here is more movie star kind of famousness, needed to be changed to keep the famousness within a group of friends

88: item made simpler by remove some of the words that are not necessarily important

89: 残忍 means “merciless” so it was replaced by 无情 which means “ruthless”.

90: 狂热 changed into 忙碌 because the first one means crazy which is too strong in this situation so it was replaced by busy which is better because there is no word or combination of words that can describe “hectic” in a better way than 忙碌.

91: 如果就让它去吧 is “let it go” so it was replaced by 如果一切顺其自然 which is similar but it about letting things on their own. But the first one is not formal so needed to be replaced with this one.
94: typing mistake

97: same with reputation

101: the translation means that “I like to go out with my friends after work” so it was replaced by “I like to go out with my friends from work in the evening”.

103: “best work” was translated to “high level”

105: changing the place of the coma

107: 感觉不到压力 is “I just can’t feel the pressure” so replaced by 不容易感受到压力 “it’ not easy for me to feel pressure”

109: 不厌其烦 means “be really patient” the suggested translation is 不择手段 “no matter how, I just want to get what I want”

112: 闯劲 (positive energy that you have in your job) 侵略性 (this one is also energy but can be unfriendly) and it means aggressive in Chinese.

114: 不修边幅 is untidy but it was changed to 凌乱 which also means untidy but the second one is more commonly used. Also the first is used to describe a person lifestyle maybe but not a person's working style.

119: this item was translated wrong into "in a team I put high standards that I cannot achieve" so it was replaced by the proper one.

121: 优良原料 is a direct translation of the English one which in Chinese can be confused with "food used in the farm" so it was replaced by 源泉 "source". Also 活跃 and 积极 are very similar but the last one means active but also positive so seems better in this context.

124: 情绪上涨 this can mean "the atmosphere can become more tense" or "people can get more excited" so it was replaced by 紧张的气氛上升 which only means "the atmosphere gets more tense"

125: the translated item is "I'm always delighted to know what exactly I should do" so "what people expect from me" was not present and was added to this sentence also the translation missed out of "the security of knowing".

126: the translation was "under pressure" so was replaced by "under time pressure".

127: 盲目 "blind/aimless" was replaced by 苛刻 "harsh"

129: the item was translated into "if you work more then you get more reward" so it was changed into "successfully finishing a job is success on its own"

130: the item is not written in a formal way. The written and spoken are different and it's not proper to use the spoken language as written. Also 从来 was added to emphasize that "nobody ever" rather than "nobody" so the addition makes the "ever" stronger".

131: 盲目自信 "unreasonably comfortable" was replaced with 过分自信 which means "over confident

136: the item was translated into "people at who are aggressive at work will always have more trouble that they should have" so it was replaced by "people who are aggressive at work usually make trouble" rather than get trouble form others.

139: 偶然的情况 means "accidental situation" was replaced by 场合 which means "occasion".

142: 狂热 changed into 忙碌 because the first one means crazy which is too strong in this situation so it was replaced by busy which is better because there is no word or combination of words that can describe "hectic" in a better way than 忙碌.

147: needed to be changed because it doesn't sound right and 千钧一发的一刹那 sounds weird.

148: replaced by a better way of saying the same thing

151: 尽力 mean "I will do my best" whereas I would go out of my way may be doing positive or negative stuff so it was replaced by 不择手段 "no matter how, I just want to get what I want"

152: 实话实说 (saying what is on your mind) was replaced by 诚实 (be honest) because it expresses the English one better (truthful)

153: 会让我 was replaced by 意味着我要 because the first one means "something will let me" whereas the second means "something means" which is what is intended.

158: 最高水平 means "best level" so it was replaced by 最好的工作成果 which means "best work".

159: 保持 (keep) which is the literal translation from the English one but here it will mean insist on my own opinions and feelings instead of "keep" o it was replaced by 处理 which means "dealing"

164: 万人瞩目 is usually used in the context where the person is a star so was replaced by a less strong one 被人关注 where it still means that you are the centre of attention but not like 1000000 are staring at you.

166: missed out on special ability.

171: replaced by a more common way of saying this.

172: the old translation means "it's not good to jusdge a person according their influence on organisational profit".

173: no changes, but organisation has 2 meanings, group and company

174: 强调 replaced with 重视 both mean emphasize something but the second one is more "attach importance to" which is closer to "make a point".

176: the translation assumed "people higher up in the organisation" but did not emphasize people who are directly supervising my work.

178: the translation means hide the truth but it should be twist it

180: 做推销 was grammatically missed out from the sentence

184: 无忧无虑 usually used to describe children because they live an easy life without worrying about anything else. So it was changed

Appendix 11: Reasons for Spanish amendments based on V2

1-change "pensar bien" (thinking hard) to "dan consideracion grave" (concider very well) because it's closer to the English version

3- change "enseguida" (quickly) change to "pronto" (immediately) because it's more accurate and "aburrida" (boring) to "tediosa" (tedious) for the same reason.

4- "nunca" (never) was changed to "jamas" which is also (never) but much stronger because the English is "never ever".

6- remove "cuando trabajo en" (when I work in) because it's not needed and it doesn't exist in the English version. Also added "particular" and changed "opinion" (making and opinion) to "tomar juicios" (making judgments). The adjectives had to be changed with that for grammatical reasons.

8- "pienso" (think) was changed to "veces" (wish). "que me gustaría poder" was replaced by "que fuera mas capaz" becasue it's better said that way even thoug they are both about the ability to speak.

10- add "mejor" (best) because it's in the english also slight change of structure because of this change for gramatical purposes.

11- "reflexion" (think) was replaced by "concentrar" (concentrate) becasue it's mre achúrate.

13- added "por lo general" (in general) because there is "usually" at the beginning of the sentence. Also changed the sentence because it meant befote that when we change the way we do things, things get worse. So it was changed to matters get worth.

14-"toman una decisión"(make a decisión) was replaced by "se decide" (make up their mind). Also "gente" replaced "personas" because it's more formal and general. "en intentar influenciarles" trying to influence them) was used to replace "hacerles cambiar de opinión" (change their opinión) because it's closer to the english one.

15- "son mejor abordados" (deal with) becasue befote it was "resolved". Also "no nos aferramos a los procedimientos formales" (when we dont stick to formal procedures) was replaced by "se dejan de lado los procedimientos oficiales" (when formal procedures are put aside) which is an expresión that is closer to by passed.

17- "negativa" (negative) changad to "exagerada" (exagerate) becasue there is no exact Word that explains "over-react" and also the first one was negative but over-react is not necessarily that negative). Also "gente" replaced "personas" because it's more formal and general which lead to a change in the stucture of the sentence.

19- add "debidamente" (properly) becasue this is how it is in the english one.

20- "Se me da muy bien hacer" (i'm very good at) was changad to "Tengo la habilidad particular" (I have a special ability) because it's closer to the english one.

21- "enfrentarme" (confront) wa replaced by "hacer frente" (stand up) because the first one is a bit negative.

22- "Doy mucha importancia al detalle" (i give too much importance for detail) was replaced by "soy muy detallista" (i'm very mucha perfectionnist, in terms of detail).

23- "tradicional" is an andjective so it was replaced by the noun "tradicionalista" because in the english one it's the noun as well.

24- "reconoce como es debido mi aportación en el trabajo" (I dont feel that my contribution to work is recognized" was replaced by "mi contribución a una tarea no ha sido apreciada del todo" (my contribution to a task wasnt fully appreciated)

25- change dbecause the sentence meant "i really enjoy working jointly on a Project that envolves a number of people" where in fact it should have been "there's nothing so enjoyable as working with others in a shared Project"

26- "one need to be shrewd" was changed to "shrewdness" and also "you have to be" was changed into "essential" because it's closer to the English version. It was possible to have exact wording so it was changed to that. Also "personas" changed to "gente" because it's more formal and general.

27- add "gran" (big, important, valuable, significant) was added because "real" was missing but we cannot add "verdad" (real) because it doesn't sound right.

30- the sentence didn't flow well so it needed to be changed especially that it meant "when I am mistakenly given too much in an establishment I always say so" so it was all reaarraged.

36- "a veces" (sometimes) was replaced by "a menudo" (often) because it's closer to the English one.##

37- "me siento" (I feel) was replaced by "estoy" (I am) because the english verisn is I am.

38- "por regal general" (as a general rule) was replaced by "en general" (in general) because it's the same as the English.

39- "cuando la gente no es atenta conmigo" was changad because it meant people are unhelpful towards me rather than " people act in an unhelpful manner".

41- "crean" (they) doesnt refer to anybody here and laso in the original english versionj it's "some people" so it was changad. "pesado" (iiritating) was replaced by "insistente" (puchy). Also, "conseguir mis objetivos" (acheive my objective) is wrong becuase in the english versión it's "get things done".

42-"debriamos" (we) was replaced by "todos deberian" (everybody) because it's like the english one.

43- add "especiales" (special)

44- "mejor" (better) was replaced by "vale la pena" (worth)

45- "inocentes" (innocent) was replaced by "ingenuas" (naive) also "son muy manipulables" (are very manipulable) was replaced by "es facil de manipularles" (very easy to manipulate them).

46- “en los que” (on which) was replaced by “donde” (when) because it's the exact translation of the english and both can fit. “”alterado” (upset) was replaced by (agitado) (agitated).

47- “se me da” (i am good at) was replaced by (soy mucho mas habil) becuase it’s better and ore formal. “razonar” (to reason) was replaced by “pensar lógicamente” (think logically).

48- “no se justo” (it’s not just or fair) was replaced by “es irrazonable” (it;s unreasonable) becuase it;s closer to the english. “jefe (boss) was supposed to be changad to “empresario” (employer) but that would mean having 2 words that are very similar in the same sentence. And “bien todoa los dias” (good every day) was changad to “siempre bien” (always good).

49- “las tareas” (tasks) changad to “trabajo” (work) as this is the original english one.

50-“cuando me estan saliendo mal” (when things arent turning out wel,l) was replaced by “cuando las cosas van mal” (when things go wrong) becuase it’s closer to the english. “me encanta” () was replacd with “alegra” (delighted, make me happy) because it’s more commonly used amd formal in the is context. Also “otros me solucionen” (other solve them) was replaced by “pueden arreglar” (can fix them). “los demas” (other people) replaced “otros” because the english is other people.

51- “un plazo previsto para la finalización” (a time set for the finalisation of) was replaced by “fecha limite” (deadline) because it’s the exact english and also shorter and more porter. “se cumpla” (complete it) was replaced by “atenersa a ella” (stick to it) and “muy importante” (very important) repalced by “siempre esencial” (always essential) because they all translate the english one better.
52- incómodo

54- “el bienestar” (wellbeing) replaced by “los intereses” (the intersts) becuase it’s a more achúrate translation.

55- adding “un tanto” (somewhat) and also changing “aburridas” (boring) to “tedias” (tedious).

56- add “muy” (very)

57- “a veces” (sometimos) replaced “ha habido momentos” (there have bneen times) becaue it’s simpler and more achúrate.

58 “a la hora” (at the time) was replaced by “al” (in) because it’s more achúrate. “los impresiones” (impressions) changad to “intuiciones” (intuition) and also added “siempre” (always).

59- add “mucho”. “perder” (waste) changad to “dedicar” (dedicate) becaue it’s closer to the english one.

60- the whole sentence was changad because it was not achúrate at all.

62- “inútil” (useless) replaced by “imprudente” (foolish, unwise). Also “precipitads” (ruched) was changad to “en el acto” (on the spot). Also taken out “siempre” (always) becasue it doesnt exist in the english version. And also changad “aparecerán” (appear) to "surjiran" (emerge) and "tarde o temprano" (sooner or later) to "a su proprio tiempo" (in their own time) because it's more achúrate and reflects the english one exactly.

65- remove "profesional" because it's not in the english version.

66- "pienso" (think) replace by "concidero" (consider) as it's more achúrate. Also change "lo que voy a decir" (what i'm going tos ay) to "mis palabras" (my words and add "antemano" (beforehand).

67- change "personas" to "gente" because of reason put befote. Also "dirian" (say) changad to (considerar) (consider because it's more like the original one. Also, added "algo" (somewhat).

68- remove "sin tener" (without having) becuase it's not in the english. And "motivo" (motive) changed to "razon" (reason).

69- changing "alguien" (a person) changad to "gente" (people) and "no se que decir" (i dont know what tos ay) to "me quedo de una pieza" (toungue tied. This is an expresión that is usually used to Express this, so it is more coomonly used and more natural.

70- "lo cuento mas facil" (I find it much easier) replaced "es mucho mas facil para mi" (it is much easier for me to) because it;s more achúrate and lso becasue the first one doesnt sounf right.

71- meaning is there but it sounds very colloquial and not formal enough so the sentence was restructured in a better way.

72- "gente" again and also "me están haciendo" (causing me to) changad to "se que estan" (i know that they are) because it's closer tot he english one. "perder" was replaced by "desperdediciando" which both mean (wate)but with the other changas in the sentence, "perder" doesnt sem. Right anymore.

73- "de vez en cuando" (from time to time) was replaced by "a menudo" (often). Also "algo" (something) replaced by "cosas) (things). All problems with tranlsation.

74- "molesta" (I'm bothered, I mind) replaced by "tengo resentimientos" (have resentment). "solucionar" (solve) changad to "ayutar" (help) and "los problemas a las personas que se lo han buscado" (problems for people who looked for them) repaced by "gente que ha creado sus proprios problemas" (people who created their own problems) because they are closer tot he englissh version but also the sentence sounds bad/slang.

75- "con machismo interes" (with a loto f interest) was replaced by "avidemente" (avidly) because it's the exact equivalent of the english. "todo articulo" (any article) replaced by "articulos de periodico" (newspaper articles) and "relacionados" (in relation to) was replaced by "conciernan" (concersning) because both of them mean the same thing but concernan is closer to the english one and there is no reason why we should use the more exact one. "el ambio de mi trabajo" (the scope of my work) replaced by "mi typo de trabajo" (my type of work) because it's more achúrate.

76- "las gestiones" (management) changad to "los preparativos necesarios" (the necessary preparations). Also, " a otras" (to others) replaced by "a los demas" (to others) but the second one is much better because it's more formal and proper.

77- "hay momentos" (there are moments) was replaced by "a veces" (sometimes) because it's more achúrate. Also "prara hacer notar my presencia" (to make my presence niticed) was replaced by "a fin de imponer mi presencia" (so that I can impose my presence) because the english one is "to assert my presence" so it's more forceful.

79- "decisiones" (decisions) changed to "implicaciones" (implications) because it's more accurate

81- "machísimos" (extremely) was replaced by "a veces" (sometimes) because it's more achúrate.

83- ""es no bueno" (it's not good) was replaced by "es imprudente" (it's unwise) because it's more accúrate. Also, "lastimar" (hurt) was added because it is in the english one but not in the spanish.

86- remove "en mi trabajo" because it doesnt exist in the english one.

87- add "grave" (serious) because it's in the english one. Also replace "sin haber reflexionado bien antes" (without first having reflected well) by "sin dejar tiempo para la reflexion" (without leaving enough time fore reflection) because this one has the time issue in it and also it['s grammatically more correct.

92- "a veces, la gente me ha dicho que" (people have sometimes told me) was used to replace "se ha llegado a decir de mi" (it has been said of me) because the first one is people have said it in my face rather than between themselves.

98- "por regla general" (as a general rule) was replaced by "en general" (in general) because it's more achúrate. And also "dedicame a los pequenios detalles" (didicate myself to small detail) replaced by "ocupame de lof detalles). And also "ortos" replaced by "los damas" (others) but it's mroe formal that the first one.

100- "segura" (trusworthy) was replaced by "firme" (steady) because it's closer to the english one. Aslo the structure was changad from they say to they know me to be.

103- "se me deja solo" was used to add "when I'm left alone" and the structure was changad a bit to avoid using "quando" twice.

104- "laborales futuras" (future workrelated) was replace by "de mi trabajo" (of my work) because prospects is already about the furute so need to repeat twice.

105- "dispuesto a hacerlo cobrando menos" (i World chrage less for it) was replaced by a more achúrate sentence "por un sueldo menor" (i World do it for less pay).

106- typing error.

107- especialmente (especially) was replaced by "excepcionalmente "(exceptionally) and also added "libre fe estrés"(free from stress) rather than (little stress)

108- "I mind" was replaced by "I resent" and also " pointless rules" was replaced by "rules made for no obvious reason" becasue pointless jeans that they have bno point whatsoever but the other one jeans that they might have a reason but we cannot see it.

109- "el resto" (the others) was replaced by "los damas" (other people) because it's better sadi that way. "hacen me mismo trabajo" (do a job like mine) was replaced by “hacen me tipo de trabajo” (do my type of work) which is better sadi that way and closer to the english one.

112-“conseguir lo que quieres” (acheive what you want) was repalced by “conseguir lo que uno quiere” (achieve what one wants) to make it more formal.

113- the meaning of the sentence was there but the structure was very far from the english one so it was all changad.

115- “no tenria problema” (i have no problem) changde to “no vacilaria” (I Worldn’t hesitate) because it’s more achúrate and also the end of the sentencve was changad for grammitacal purposes, avoiding repetition.

116- “estoy” (i am) replaced by “podria ser” (i may be) because it’s closer to the english one.

117- add “la parte” (the part) and “de mi plan” (of my plan) because they are missing.

118- “laborales” (labor) changad to “trabajo” (my work) because it’s much better to put tit that way.

119- changad “se me da” (it suits me) was replaced by “soy” (i am) becasue it’s more formal and more accurate. Also “opinar sobmre” (expressing opinions) replaced by “juzgar” (judging) and also “producirlas” was added to mean produce

120- changad “in mi opinión” (in my opinión) to “encuentro” (i find). And “ a veces (sometimes) so “suelen ser” (it ofetn is).

121- added “activa”.

122- “Yo diria” (i world say) was replace by “parezo” (I seem to have).

124- “estan alterados” (to be upset) was replaced by “emociones se desatan” (emoting are high).

125- “ siempre prefiero tener la seguridad” (i always prefer having the security) as opposed to being sure and also “what is expectedof me” rather than “what they expect” becasue here we dont know who they are.

126- “”a veces” (sometimes) was replaced by “suelo” (often). Also “tiempo” (time) was added befreo (pressure).

128- change I am a worrier by nature to “I tend to worry about things” becuase the first one jeans that you worry constantly whereas the second is trae but not necessarily all the time and also because there is no Word for “worrier”.

129- add “siempre”

130- “en duda” (questioned or douted) chaged to “considerado” and also “nunca nadie” replaced y “personna jamas” because it’s more formal and stronger in meaning.

133- fly by night (if found)

135- “mi jornada laboral” (my working day) was replaced by “todas las horas que estoy en el trabajo” (all the hours that I am at work).

136- all the esntence was changad becasue it didnt sound very formal and accurate.

137- add “jamas” (never ever) to make it stronger in meaning.

138- typing mistake

141- “y no se sente bien” (not feeling well) was removed because it’s not i the english one. And the structure of the sentence was chaged to make it closer tot he englihs.

143- Added “produccos” (I produce) becaues it’s in the english, and also “los” repaced by “mis” qand “mucho mejores” (much better) replaced by “mejores resultados” (my best).

144- typing listake.

145- “sadly” changad to “unbfortunately”. Also removed “realmente” (really).
149- typing mistake

151- changad the whole sentence because it sounds clumsy and too long and can be said in a shorter way.

154- “delego” (delegate) was replaced by “de dejar” (to leave) because it’s closer to the english one.

157- “equipo” (team) replaced by “empleados” (staff pr employees).

159- “opiniones” (oponions) repaced by “sentimientos” (feelings) because it’s mroe accurate.

161- typing mistake.

162- reduced the sentence to a smaller one that is closer to the english.

165- “salir” (turn out) replaced by “ir” (go) and also added (probablemente” (probable).

166- typing mistake. And added “special” because it’s i the english one.

168- “mucho” was removed because it’s not in the english version. Also “demasiado” (too much) was added later becaue it is in the english one.

169- “se” (I know) was replaced by “estoy seguro” (i’m sure) and also “saben” (they know) was replaced by “me consideran” (consider me) because it’s closer to “think of me” which is in english.

170- “dedico” (dedicate) changad to “tomare” (take) because more accurate. Also removed “de lo necesario” (than necessary) because it’s not necessary and lengthens the sentences without a good reason. Added “para hacer un trabajo bien” (to do a job well) becaue it’s in the english. “no es realment necesario” (it’s not really inecessary” replaced by “relativamente insingnificante” (relatively insignificant” becaseu it’s much closer to the english one and the first one means that it’s not necessarily whereas the other one means to a c eertain extent not necessary.

171- como todos (like everyone) was replaced by “como la mayoria de la gente” and also removed “creo que” (I think)

172- “no esta bien valorar” (it’s not good to assess) was replaced by “es erróneo juzgar” (it’s wrong to judge). Changad the rest of the sentence to “ según su impacto sobre los bienes de una empresa” (according to its effect on a company;s profit) more accurate, less clumsy and more formal.

173- “es mi puento fuerte” (is my strong point) was replaced by “ mi ventaja mas importante” (my most inoprtant advantage) becasue it’s the closest to the englishone as it doesnt exist in Spanish.

174- the sentence was all changad because it’s not accurateand it’s clumsy. “empeno” (to make an effort) replaced “hincapié” whcih is the same but less formal. Also added “quedarme abierto” (leave myself open). And the last sentaence was replaced by “a ser influenciado or las ideas de los demas” (to be influenced by pther people’s ideas) because the first one was longer, clumsy and not straight forward. And also they used “foreign ideas” instead of “other people’s ideas”.

175- “particular” was added because it is in the english version. also “applicabilidad” replaced by “implicaciones practicas” becasue the fist one is about implementation whereas the english one is about pros and cons and benefits. So the replacement is much more accurate.

176- added “jamas” tos ay never ever also removed “las” from befote personas becasue it’s not there in the english one.

177- the sentence meant “dont tolerate routine well” but in english “a little too intolerant” means that you know that you should be more tolerant but you are not so it was changad toa n closer one.

178- add “really”. Also “alterar” (alter) was replaced by “torcer” (twist) because it’s more exact but they both mean the same.

179- replaced by a sentence that is closer in structure to the english one alsthough meaning is not very different.

180- as a general rule replaced by in general. Also “puerta a puerta” is literally the same as the snglish one but means “delivery person” so it was reoplaced by “vendedor ambulante” which means “mobile saleperson” which is the cultural equivalent in Spanish.

181- removed “en algún momento” (at any time) because it doesnt exist in the english one.also the structure of the sentence was chaged to put it in the passive.

182- replace “el motive” (motive) with “el razonamiento” (reasonning).

183-there is a more exact transaltion so it was used.

184- “tranquilo” replaced by “flojo” because it also has a negative sense to it.

185-“cuando fracaso” (when I fail) was replaced by “el fracaso” (failure). Also the sentence was restructured to become passive becasue it’s better grammatically.

186- check if we can use imbecil

188- remove “muy” because it’s not a lot”.

189- add “tengo un poco la tendencia” (i have a little tendecy” because it’s not in the spanish version and also that’s the closest to “a little too prone”.

190- “proyectos” (projects) changad to “tareas” (taks).

Appendix 12: Arabic V3

Copyrighted information

Appendix 13: Chinese V3

Copyrighted information

Appendix 14: Spanish V3

Copyrighted information

Appendix 15: Reasons for Arabic amendments based on V3

6- “على اتخاذ قرارات” (make decision) was replaced by “اطلاق احكام” (make judgments) because it’s more accurate

8- “ما أفكر فيه” (what I think about) was replaced by “ما يجول في خاطري” (what’s on my mind)

16- أفضل من البوح بها “better than saying the truth” was put back instead of the أكثر عقلانية من البوح بها “more wise than saying the truth” because it sounds much smoother in Arabic than the one that it closer to English.

17- ”أبالغ أحياناً “ (I over-react sometimes) was replaced by “أحياناً اميل الى المبالغة “ (sometimes I tend to over-react) because the English version had “tend to”.

21- occasionally and sometimes have the same translation in arabic.

23- Traditionalist is noun but it doesn’t have an exact word to describe it in Arabic, the closest equivalent is “someone who holds on to traditions”

26- “الطباع الرديئة” (bad attitudes) was replaced by “ استعمال النباهة اساسي" (using slyness/cleverness is essential) because the first one was a wrong translation probably because it sounds like rudeness. Also important was changed to essential.

27- “real power and influence” was translated to “any sort of power or influence” because that’s a more proper way of saying it rather than translating it word by word.

28- “احدى أولويات الشركة الأكثر أهمية” (one of the company’s most important priorities) is the closest and most natural way of saying “the company’s top priority”

39- “بطريقة لا تسهّل الأمور” (a way that doesn’t make things easy) was replaced by “بطريقة غير متعاونة” (unhelpful way)

44- “ الأمور الايجابية تنتج دوماً عن قول الحقيقة “ (saying the truth always leads to positive things) there is no close expression to (it always pays to tell the truth).

58- “نقيّم” (assess) was replaced by “نطلق الاحكام” (judge) because it’s more accurate.

109- “ ابذل جهداً كبيراً“ (I do a big effort) was used to explain “ I go out of my way” because that’s the closest expression.

111-“يفرض عليّ “ (the is forced on me) was removed because working according to strict guidelines already assumes this and it’s not in the English version.

137- double negative, dishonest makes it that way.

166- “واضحة ودقيقة “ (clear and specific) was replaced by “.وجيزة “ (concise) because it’ more exact translation.

189- “أميل الى الظن بأن أخطائي كبيرة “ (I tend to think that the size of my mistakes is massive) was replaced by “أميل الى المبالغة في حجم اخطائي” (I tend to exaggerate the size of my mistake) because exaggerate was mssing

Appendix 16: Reasons for Chinese amendments based on V3

19:”properly” was missing from the sentence

28: I don’t know what we changed here. There are two changes in this item. One is the measure word which I put was wrong and is now replaced by the correct one. The other word was added to make the sentence more smooth.

31: “my” was added next to “mind” because it exists in the English one. However, the sentence was understood anyway because you don’t speak somebody else’s mind.

33: “things” was replaced by very similar word but that would be better for the flow of the sentence. Also, “when” was added to rearrange the structure of the sentence. And “difficulty sleeping at night” was replaced by an idiom that means “flipping in bed and not being able to sleep because of stress”. “things” in Chinese can be expressed by one word which is the one I used, also can be expressed by using a combination of two words – which is the word I used plus another word, but the meaning does not change.

35: “praising” was replaced by the same word “praising” but the first one is usually used for praising the performance whereas the second one is directed to people.

39: “helpless” changed to “unhelpful” because it was a wrong translation.

41: “aggressive” was replaced by “pushy”. Also this made rearranging the sentence necessary. The word we are using here to replace the old one is because this word has the meaning which is closer to the original English version.

42: got back to the original one why? The original one highlights the point that “everyone should live a comfortable life, and this is main point of living a life, so working hard to get promotion is not right, because it makes a person’s life hard”, whereas the one I change emphasizes “if we don't need to worry about promotion, then everyone can live in a more comfortable way.”

45: “manipulate” was replaced by “manipulate” but the first one was for machines whereas the second one is for humans. “them” at the end of second half of the sentence was moved to the front of the second part, in this way, the sentence sounds more smooth.

46: “to calm down and get a job done” was replaced by “get a job done” because “to calm down” does not exist in the original English one

49: “against” was replaced by “against” because the first one means 2 countries against each other whereas the new one relates to being emotionally against something. . Some other words were added in to help rearrange the sentence, but these words do not change the meaning of the sentence.

56: don't know what we changed here. It looks like the whole sentence was replaced by another sentence; but in fact, two sentences have very similar meaning, just the second one express the original meaning a bit better and we are looking for the best translation, so we changed this one.

59: typing mistake and also "better than" was replaced by "rather than". The first change is typing mistake, the second change is "rather than".

71: added "almost"

78: "always" replaced by "particularly".

79: "prefer" could be in Chinese "willing" or "like" so it was changed to "like" because it explains prefer better

82: the first sentence meant "if I can do what I want" so it was replaced by "if I could completely without limitation do what I want".

101: remove "people I know from work" was replaced by "people from work" because both are understood in the same way but the original English one doesn't have "I know".

107: "describe" was replaced by the same word "describe" but the first one is used to describe something whereas the second one is used to describe oneself. "especially" was replaced by "especially", but the second one sounds better and have a closer meaning to the original version.

109: "I will find a way" was replaced by "use all possible ways" to explain "go out of my way" this replacement can mean that you would do anything, whether it's good or bad whereas the first one was always finding bad ways.

112: change "aggressive" was replaced by "aggressive" but the first one means "ready to fight" which is very negative whereas the second one is "self-assertive" or "forceful".

114: "style" was replaced by "style" but the first one is for the style of writing but the second one is for the style of doing something.

119: "comment on something" was replaced by "judge" because it more accurately describes the original one.

121: "is" replaced by "can be" also, "dream life" was replaced by "fantasy life". Also, "full of" was added to help "active" because in Chinese you need to use some words to explain other in a better way. Also "active" was replaced by "active" because the first one is more energetic but the second one is "functioning, in operation".

124: "efficiently" replaced by "exceptionally" because it describes the English version accurately.

125: “always” was added because it was missing in the translation. Also, “safe” was replaced by “secure”

127: added one word so that “I think” will become “expect”. They all fall under expect but one is think and the other is expect.

129: “always” was added because it was missing from the sentence. Also, “one of the rewards” was replaced by “sufficient reward”.

136: : change “aggressive” was replaced by “aggressive” but the first one means “ready to fight” which is very negative whereas the second one is “self-assertive” or “forceful”.

151: “I will find a way” was replaced by “use all possible ways” to explain “go out of my way” this replacement can mean that you would do anything, whether it’s good or bad whereas the first one was always finding bad ways.

159: “dealing” was replaced by “hide” because you cannot say in Chinese “keep my feelings to myself” even though you can use “keep a secret” but this verb cannot be used with feelings.

164: “people” was deleted because there is no need to say focus of people’s attention, “focus of attention is sufficient” especially that it doesn’t exist in the English version.

166: changed the order of the sentence, not for grammatical purposes, but to put the emphasis on the more important bit of the sentence.

174: “open” was replaced by the same but the first means literally open and the second one is “willing to”.

176: “supervisor” changed to “people who supervise my work” also “supervise” was changed to a better more common one.

184: added one word so that “I think” will become “expect”. They all fall under expect but one is think and the other is expect. Also “not active” was replaced by “laid back” because it’s more accurate and describes it better.

Appendix 17: Reasons for Spanish amendments based on V3

4- "que nunca me atrevería a decir" was replaced by "de las que nunca hablaría" is a more accurate translation and a better way of saying it. It sounds more Spanish that way.

10- "se puede medir por su capacidad de incrementar los beneficios" was replaced by "se juzga por su habilidad para incrementar beneficios." Because the sturcture of the sentence is better in terms of grammar and vocab (accuracy).

14- "es mejor no intentar cambiarla" was replaced by "no tiene sentido hacerles cambiar de opinión." Because the first one is unnatural because it's too literal. Also, less clumsy that the first option.

16- " es preferible no decir la verdad" (not preferible tos ay the truth) was replaced by "no es sensato decir la verdad" (not sensible to say the truth) because it's more accurate tranlation

20- " comprometidas" (engaged or difficult) was replaced by "incisivas" (sharp) because it's closer the English "penetrating questions" although it's not wrong, it's more accurate

30- If someone accidentally gave me too much change I would always tell them.
30- Cuando me devuelven demasiado cambio de más en un establecimiento siempre se lo digo.
30- Cuando, par error, me dan cambio de más en un establecimiento siempre se lo digo.

40- Sentence was rearranged to stick to the structure of the orinigal English one and also because it sounds better. Also "me dejaran a mis anchas." is an expression that is closer in meaning to the English expression "left on my own devices".

41- The sentence was rearrange and 2 words changes to make closer to the meaning in the English version

42- Todos (all) was removed because in spanish you can add a suffix to explain who it refers to so it's unecessary. Tranquilos (calm) was replaced by "de manera desahogada" (in a better way) becasue it's closer to the english "comfortably". conseguir un ascenso. More achúrate in general.

58- "A la hora de juzgar" (when judging) was added becasue it was inaccurate in the older version. The whole sentence was restructured to sound more formal but also more like Spaniards would talk.

66- una was removed because it's not in the english one so it's more accúrate now. "preparo de antemano" (prepare in advance) was re0placed by "pienso" (think) because it's closer to "consider" in the english version..

76- "dejar que los demás lo gestione" was replaced by "dejarles las gestiones a otras" becasue it's better grammar. Also, "siempre" (always) was replaced by "con tal" (so long as) because the first one is wrong and is not natural spanish.

85- trabajos manuals (manual labor) was replaced by "trabajando conlos manos" (working with my hands) because it's much more accurate.

93- "preocupo temo" to make it more formal and closer to the english version.

94- "no fiarme tanto" was replaced by "desconfiar" (distrust) because it's better said that way

96- "Con toda sinceridad" (in all honesty) was added because it was missing. Also, "a veces" (several imes) was replaced by "suelo" (tend to) because it's closer to the english version.

100- segura/o the o was added to include femenine and masculine.

106- "gana" (gained) was replaced by "consigue" (attained) althought the first one is the exact literal translation, the second one sounds better. Also "engañando" (deceiving) was replaced by "engaño" (deceit) because it's more accurate gramatically and meaning.

106- "poco predispuesta al estrés" (with little predesposition to stree) was replaced by "con especialmente poco estrés" (with epecially very little stress) because it's closer to the english version, soa more accurate transaltion.

124- "disminuir tensiones" (reduce tension) was replaced by "tranquilizar a la gente" (calm people down) because it's a more accurate translation though not wrong. "compañeros" (friends) was replaced by "la gente" (people) because it's more accurate.

127- "juzgo demasiado" (i judge people a lot) was replaced by "a menudo soy demasiado duro juzgando" (i am oo harsh in my judgment) because the first one is a worng transaltion.

131- "que soy demasiado autoritario" (tha I am very authoritarian) was replaced by "que me impongo demasiado" (that i impose myself a lot) becasue it's closer to the english one.

133- "disponer" (arrange) was replaced by "seguir" (follow) because the first one is wrong. "atajos fáciles" (easy shortcuts) was replaced by "descentrarse" beause the first one is worng.

134- "Que entregar un trabajo en un plazo previsto" was replaced by "entrega ajustadas" because it's better said that way enstead of making it too long and also better structure.

143- "en equipo" (in a team) was replaced by "en colaboración con otros" (in collaboration wih others because this is the correct translation.

149- he sentence was restructured to sound more Spanish ven thought the original structure is like the English one.

155- "la visión necesaria (necessary vision) was added because it's in the english version.

161- In important matters people eventually come round to my way of thinking. "siempre convenzo a los demás de que mi forma de pensar es la mejor" (i always convince people that my way of thinking is the best) was replaced by "la gente suele acabar pensado como yo" (peopl end up thinking like me) becaseu the first one is wrong and this one is more accurate and formal.

163- En el deporte tengo muy mal perder (in sports, i don't like loosing) was replaced by "Me resulta muy desagradable perder en un juego (i dislike loosing at a game) because it's much more accurate.

165- The sentence was restructured becasue it was grammaticaly incorrect and doesnt flor well in spanish.

166- "facilidad" (easiness) was replaced by "habilidad" (ability) becasue it's more accurate. Also added "concias" because it is in the english versión.

173- Sentence was restructured and changed because it wasn't very accurate and also now it's more formal.

174- I often make a point of leaving myself open to being swayed by other people's ideas.

174- "opiniones" (opinions) was replaced by "las ideas" (ideas because it' mroe accurate. Also the sentence was restructured to sound less clumsy and more formal.

Appendix 18: Arabic back translation V4

Copyrighted information

Appendix 19: Chinese back translation V4

Copyrighted information

Appendix 20: Spanish V4

Appendix 21: Randomly selected items for coding

Arabic Dyads/triads 1:

Item 4: أقولها “I say it” was replaced by أتكلم عنها مع أحد “talk about it with somebody else” to emphasize that you wouldn’t ever discuss it with someone rather than just a matter of not being able to talk about it because you cannot express it well.

Code 1: Not exact literal translation

Item 12:” التعاطي” (dealing) was replaced by “التعامل” (dealing) because the first one can be used to say dealing with people but is more used as dealing drugs. Whereas the second one is used with people only.

Code2: Synonym used in different context in the TL

Arabic dyads/triads 2:

166- “واضحة ودقيقة “ (clear and specific) was replaced by “وجيزة.“ (concise) because it’ more exact translation.

Code 3: Not exact literal translation

189- “أميل الى الظن بأن أخطائي كبيرة “ (I tend to think that the size of my mistakes is massive) was replaced by “أميل الى المبالغة في حجم اخطائي” (I tend to exaggerate the size of my mistake) because exaggerate was missing

Code 4: not accurate translation of expression

Chinese dyads/triads1:

19: the 2 statements are similar but were replaced with a more comfortable one that sounds better in mainland china and also because this one is word by word translation but in mandarin you don’t really say it this way.

Code 5: literal translation of expression replaced by a culture specific one
Code 6: better structure of sentence

51: 被设定了 was added to make the sentence passive like the original English one. 最终期限 and 完成期限 both mean deadline but the second one is the type of deadline that is related to a task or job so even though the word work was already mentioned in the sentence, you need to use the second one because it is more work related and should be used in this context. 很 means “very or extremely” so it was changed to 最 which means “most” so the combination of this word that was replaced and the 2 after it will lead to “essential” whereas if it was kept as it was it would be “important” but nor necessarily “essential” so the strength of the word will decrease.

Code 7: not following grammatical structure of OL
Code 8: synonym used in different context in TL
Code 9: not exact literal translation of word

Chinese dyads/triads 2:

41: "aggressive" was replaced by "pushy". Also this made rearranging the sentence necessary. The word we are using here to replace the old one is because this word has the meaning which is closer to the original English version.

Code 10: not accurate translation of word

107: "describe" was replaced by the same word "describe" but the first one is used to describe something whereas the second one is used to describe oneself. "especially" was replaced by "especially", but the second one sounds better and have a closer meaning to the original version.

Code 11: synonym used in different context in TL
Code 12: not accurate translation of word

Spanish dyads/triads1:

14- "es mejor no intentar cambiarla" was replaced by "no tiene sentido hacerles cambiar de opinión." Because the first one is unnatural because it's too literal. Also, less clumsy that the first option.

Code 13: Unnatural literal translation

165- The sentence was restructured becasue it was grammaticaly incorrect and doesnt flor well in spanish.

Code 14: grammatically incorrect sentence

Spanish dyads/triads 2:

125- " siempre prefiero tener la seguridad" (i always prefer having the security) as opposed to being sure and also "what is expected of me" rather than "what they expect" becasue here we dont know who they are.

Code 15: not exact literal translation
Code 16: grammatically incorrect sentence
143- Added "produccos" (I produce) becaues it's in the english, and also "los" repaced by "mis" qand "mucho mejores" (much better) replaced by "mejores resultados" (my best).

Code 17:missing word
Code 18: not accurate translation

Appendix 22: Initial codes based on randomly selected items

Codes from sample amendments	Initial Codes template
1: Not exact literal translation **3:** Not exact literal translation **4:** not accurate translation of expression **9:** not exact literal translation of word **10:** not accurate translation of word **12**: not accurate translation of word **15:** not exact literal translation **18:** not accurate translation	Code 1: Literal translation more appropriate (LTMA)
2: synonym used in different context in TL **8:** Synonym used in different context in the TL **11:** synonym used in different context in TL	Code 2: Context Dependent Synonym (CDS)
6: better structure of sentence	Code 3: Better Structure (BS)
7: not following grammatical structure of OL	Code 4: Sentence Grammatically Nonequivalent (SGN)
5: literal translation of expression replaced by a culture specific one **13:** Unnatural literal translation	Code 5: Literal Translation not most appropriate (LTNMA)
14: grammatically incorrect sentence	Code 6: Grammatical Mistake (GM)
16: grammatically incorrect sentence **17**:missing word	7: Missing Word(s) (MW)

Appendix 23: Briefing for inter-rater reliability

The aim of this exercise is to compress the information below into smaller and more specific pieces of text. Please read through the codes in the table carefully to understand their meaning and the differences between them. Then read the comments next to each item below, and assign the codes or codes than represent all the information in it.

Item 132: محبطا "depressed" was replaced by أفقد عزيمتي "loose my strength and will" because the first one is too strong especially that the original one in English is "discouraged". Depressed is Arabic is more clinical and its meaning is way too strong for expressing this point.

Item 136: لا يستحقوا عناء توظيفهم "are not worth hiring them" was replaced with يسيؤون اليه أكثر مما يفيدونه "do bad to the work environment rather good". The English one is "are more trouble than they are worth" but in this context, the sentence cannot be translated literally because it wouldn't transmit the right meaning. The expression that was used to replace it expresses the intended meaning better.

Item 189: "أميل الى الظن بأن أخطائي كبيرة " (I tend to think that the size of my mistakes is massive) was replaced by "أميل الى المبالغة في حجم اخطائي" (I tend to exaggerate the size of my mistake) because exaggerate was missing

Item 28: "احدى أولويات الشركة الأكثر أهمية" (one of the company's most important priorities) is the closest and most natural way of saying "the company's top priority"

Item 45: 幼稚 replace by 天真. Both of them can mean "naïve" but the first one Is more "childish" whereas the second one is more "naïve". Also the order of the sentence was changed because it sounds better.

Item 62: 它们自己的时间出现 means "they will appear in their own time" was replaced 特定的时间里自然出现 means "they will appear naturally by themselves at some point in the future". Although the first one seems to be more appropriate when translated to English, however the second is what is common and more normal to use in Chinese and the first seems unnatural to say.

Item 49: "against" was replaced by "against" because the first one means 2 countries against each other whereas the new one relates to being emotionally against something. . Some other words were added in to help rearrange the sentence, but these words do not change the meaning of the sentence.

Item 109: "I will find a way" was replaced by "use all possible ways" to explain "go out of my way" this replacement can mean that you would do anything, whether it's good or bad whereas the first one was always finding bad ways.

Item 66: “una” was removed because it’s not in the english one so it’s more accurate now. “preparo de antemano” (prepare in advance) was replaced by “pienso” (think) because it’s closer to “consider” in the English version..

Item 127: “juzgo demasiado” (i judge people a lot) was replaced by “a menudo soy demasiado duro juzgando” (i am oo harsh in my judgment) because the first one is a worng transaltion.

Item 6: remove “cuando trabajo en” (when I work in) because it’s not needed and it doesn’t exist in the English version. Also added “particular” and changed “opinion” (making and opinion) to “tomar juicios” (making judgments). The adjectives had to be changed with that for grammatical reasons.

Item 17: “negativa” (negative) changed to “exagerada” (exagerate) because there is no exact Word that explains “over-react” and also the first one was negative but over-react is not necessarily that negative. Also “gente” replaced “personas” because it’s more formal and general which lead to a change in the stucture of the sentence.

Appendix 24: Theme 1: Accuracy of Translation

ode	Example	Broad Codes	Specification	Theme	9.8. Affects
: Literal translation ore appropriate	"aburrida"(boring) and "tediosa" (tedious)	1:Literal translation more appropriate	Any word that has a literal translation that can be used in the TL	Accuracy of translation	Linguistic equivalence
: Sentence rammatically non-equivalent	"I do not like parties" and " I hate parties" have the same meaning but one is negative and the other poitive	2: Sentence, phrase/clause, or word grammatically inequivalent	Any sentence word, phrase/clause, or sentences that does not follow the same grammatical structure as the OL (passive vs. active voice; translation into different lesical category ex. Adjective into noun)		
10: Word(s) Grammatically Nonequivalent	"I take risks" and " I am a risk taker" have similar meanings but one is a verb and the other an adjective				
7: Omitted Word(s)	Such as "very"	3: Wrongly omitted or added word	Any word that does/ does not exist in the OL, that was wrongly omitted/ added to the TL		
9: Wrongly Added Word	Such as "almost"				
14: Composed Words in TL	"Dikka" (precision) and "Dikka fi el wakt" (precision in time) meaning punctuality				
8: Wrong Meaning	Change has several meanings (money or modification) and could be translated wrongly	4: lingo-syntactic mistake	Any word, phrase/clause, or sentence that contains a grammatical or semantic mistake		
6: Grammatical Mistake	"I has eaten"				

Appendix 25: Theme 2: Language idiosyncrasies

Code	Example	Broad Codes	Specification	Theme	Affects
4: Context Dependent Synonym	"التعاطي" (altaati) means dealing but usually drugs, or commerce. "التعامل" (altaamoul) means dealing but with people.	1: Context Dependent Synonyms	Any word that has several synonyms in the TL that are used in different contexts	Language idiosyncrasies	Cultural equivalence
3: Better Wording or Structure		2: Sentence Formulation	Any wording or sentence structure that affects the quality of the writing (not the meaning) in the target language.		
16: Unnatural or informal wording					
15: Words Nonexistant in TL	"fantasy" does not exist in Arabic. It could be replaced with لايخلا which also means fiction and imagination	3: Words Nonexistant in TL	Any word that doesn't have an equivalent in TL		
19: Elaboration	如果 means "if" is sometimes added to sentence to refer to hypothetical scenarios even if it was not in the English verion	4:idiosyncratic omissions or additions	Any sentence that needs words to be removed or added to the original language in order to make it more specific to the target language		
20: Shrinking	"Top priority" has to be reduced to "priority" because a priority is at the top already.				
13: Idioms	"fly-by-night schemes" or "go out of my way"	5: Idioms	Any expression in the OL whose meaning cannot be understood from literal translation		

Appendix 26: Theme 3:

Code	Example	Broad Codes	Specification	Theme	Affects
11: Leading Literal Translation	Manipulate in Chinese could be either done in a "unjust way" or " in a positively smart way".	10: Leading Literal Translation	Any word that could be positive or negative in Original language but leading in the Target language	Connotative meaning	Psychological equivalence
12: Different Magnitude	"nunca" and "jamás" both mean "never"	11: Different Magnitude	Any word that is the exact literal translation in the TL but is not the most appropriate equivalent because its magnitude is stronger or weaker in the TL than in the OL		
5: Literal Translation not most appropriate	"discouraged" (mouhbat) in Arabic was replaced by "my determination decreases" (takoullou azimaty)	12: Literal Translation not most appropriate	Any word whose literal equivalent does not convey the same meaning or feeling in TL		

Appendix 27: Email for participants

Hi
I am a PhD student at City University, London working on cross-cultural adaptation of personality tests from English to Spanish, Arabic, and Chinese.
XXX language is of central interest to my research, as the grammar and sentence structure is very different to English.
Will you be able to help with this important research?
I have developed a XXX version of Orpheus- a work based personality Questionnaire and need 200 native XXX speakers to complete it.
In return, I will provide you with a report describing your personality preferences in the workplace.
All the names will be kept confidential, and no individuals can be identified from the final data set.
Here is the questionnaire with the user name and password:
http://www.staff.city.ac.uk/psychstudies/Orpheus/Survey/Ar.h
tm
username – avgs
password – xlnt1 (the second character is L for Lima and the last is the number one)

Many thanks in advance for all your help!!

Lina

Appendix 28: Arabic items for cognitive interview

Copyrighted information

Appendix 29: Chinese items for cognitive interview

Copyrighted information

Appendix 30: Spanish items for cognitive interview

Copyrighted information

Appendix 31: Confidence rating scale

Similarity Scale

Not similar at all

No very similar

Similar

Very similar

Exactly the same

Answer options

Strongly Agree

Agree

Disagree

Strongly Disagree

Appendix 32: Cognitive interview protocol

Preliminaries and framing (see other document)

a. Present each question on a card

b. Can you explain the meaning of this questions/statement in English?

2- Present answer options

3- Start Probing (nodding and active listening all along)

a. How did you come up with this answer?

b. Can you tell me a specific example of what was going through your head that made you come to this decision?

c. what are the key words in this statement to you?

d. What do they make you think of? How does this relate to your answer?

e. Can you detect any words in this statement that could be problematic (i.e. be misunderstood) in this language?

4- Present the item in English

5- Present answer options

6- Start probing

i. If the answer is different:

ii. Can you repeat this item in your own words?

b.

i. Is there anything is this sentence that you do not understand fully? (if yes, give dictionary and show the closest definition)

ii. How did you come up with this answer? (maybe some more of the aforementioned probes)

iii. What do you think might be the difference between the item in Chinese and English?

iv. Do you have suggestions to make the Chinese item more similar the English one?

v. What does "the word that was changed" mean to you?

vi. How confident are you that the item is now equivalent to the English one? (present another confidence diagram)

c. If the answer is the same:

i. Is there anything is this sentence that you do not understand fully? (if yes, give dictionary and show the closest definition)

ii. Can you repeat this item in your own words?

iii. How did you come up with this answer?

iv. Do you think there is a difference between the item in Chinese and English?

v. Is there anything you would change in the Chinese item to make it more similar the English one? If yes:

vi. What to you is "the word that was changed"?

Appendix 33: Consent form

Consent Form for taking part in a project at City University

Working title: Towards a culture-free model of the Big Five - a cross-cultural investigation of the Orpheus in five different languages

PhD candidate: Lina Daouk
Supervisors: Dr Almuth McDowall
Professor John Rust

I agree to take part in the above City University research project. I have had the project explained to me, and I have read the Explanatory Statement, which I may keep for my records. I understand that agreeing to take part means that I am willing to:

- be interviewed by the researcher Lina Daouk
- allow the interview to be audiotaped

Data Protection

This information will be held and processed for the following purpose(s):

- to assess the quality of a translation from English to TL
-

I understand that any information I provide is confidential, and that no information that could lead to the identification of any individual will be disclosed in any reports on the project, or to any other party except for the supervisors of the researcher: Dr Almuth McDowall, Professor John Rust and Lina Daouk. No identifiable personal data will be published. The identifiable data will not be shared with any other organisation.
I understand that this information will be used only for the purpose set out in this statement and my consent is conditional on the University complying with its duties and obligations under the Data Protection Act 1998.

Withdrawal from study

I understand that my participation is voluntary, that I can choose not to participate in part or all of the project, and that I can withdrawal at any stage of the project without being penalised or disadvantaged in any way.

Name:
....(please print)

Signature: ....
Date: ....

Appendix 34: Debriefing for cognitive interview

Introduction and briefing

Working title: Towards a culture-free model of the Big Five - a cross-cultural investigation of the Orpheus in five different languages

PhD candidate: Lina Daouk
Supervisors: Dr Almuth McDowall
Professor John Rust

Thank you for taking part of this interview. This session will start by describing the project, the purpose of this interview, and finally the process of the interview.

As part of my PhD project, we are currently working on developing five multi-lingual versions of a work-based personality questionnaire-Orpheus- into English. Arabic, French, Mandarin, and Spanish. In order for these tests to be comparable/equivalent in all languages, they need to be equivalent linguistically, culturally and metrically. We have already developed the questionnaire in all those languages and the process has involved several speakers of these five languages in order to increase the accuracy of the translation. However, in order to be even more sure that the questions are well translated, we gave the questionnaire to Mandarin speakers, in English and in Mandarin. After careful statistical analysis, we found that there are some items that are answered differently when presented in different languages. So, in this interview, we would like to explore how each one of these questions is understood in Mandarin so that we can build a better understanding of the culture, language, and comparability between the English and Mandarin versions of each question.

As for the interview, it will last for a maximum of an hour and a half. During the interview, you will be present with questions on a card, and asked to rate whether you Strongly Agree, Agree, Disagree, or Strongly Disagree to each of them. You will then be asked a few questions and you are encouraged to think aloud while answering the questions, and to give as much information as you can, whether you think it's relevant or irrelevant. Your active participation is really important for us and really appreciated.

You will also be rewarded 10 pounds per hour for your greatly appreciated input in this interview. Please take some time to read the consent form carefully and sign it. Once you've read it, signed it, and are happy with it, we can start our interview.

Best Regards,

Lina Daouk
PhD candidate
City University
Tel: 07863330863
Email: L.Daouk@city.ac.uk

Appendix 35: Arabic V5

Appendix 36: Chinese V5

Copyrighted information

Appendix 37: Spanish V5

Copyrighted information