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 1

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…

Preface: Overview of the PhD structure

This thesis spans over several areas of psychology: organisational psychology, personality, psychometrics, cross-cultural assessment, and test adaptation with main focus on the last topic. This has guided the structure of the thesis as follows.

The first four chapters (1, 2, 3, and 4) provide an introduction to and a thorough review of the literature which informed the conceptual and theoretical background of this thesis with focus on the Theory of Equivalence and Bias (van de Vijver & Leung, 1997) in adapting work based personality questionnaires.

The fifth chapter (Chapter 5) is an overview of the methodology that is used in subsequent chapters including a glossary of all the terms that will be used recurrently. However, the following chapters will also provide a glossary each, at the beginning of the methods section, which is specific to the terms employed within it to facilitate the reader's understanding.

The next four chapters (6, 7, 8, and 9) report the three main studies but are laid out over four chapters to accommodate the analyses and the hypotheses being tested. These chapters include a comprehensive review of the literature 1) about the specific methods applied in that study and 2) about the statistical techniques used to analyse the results.

Finally, each chapter concludes with a discussion and offers implications and opportunities for future research. Table 1 presents a summary of the content of each of the chapters.

Chapter 1: General Introduction – The importance of cross cultural research. This chapter set the scene for the whole PhD by highlighting the importance of research in cross-cultural assessment. In this chapter we aim to define cross-cultural assessment and the different streams it could take. This is discussed in relation to fairness in assessment and its implication on test development and adaptation. The chapter concludes with the general aim of this thesis.

Chapter 2: Personality across cultures. This chapter discusses personality theory with respect to the challenges, namely the role of culture, in defining and assessing it. The Five Factor Model (FFM) is reviewed in detail as it constitutes the basis of the instrument later used in the analysis. This review focuses on the FFM specifically in the workplace.

Chapter 3: The instrument- Orpheus work-based personality test. This chapter offers a critical description of Orpheus, the instrument used as the basis of this thesis. The psychometric and other technical qualities of the instrument are discussed in light of classical test theory and the literature on response distortion.

Chapter 4: Levels of Equivalence and Bias. This chapter describes and critically discusses the only theory of test adaptation, developed by van de Vijver and Leung (1997). The chapter distinguishes between the different types of equivalence and bias, the relationship between them, their sources and the ways to achieve/minimise them. The end of the chapter offers a suggested reformulation of this theory into a framework of equivalence and bias.

Chapter 5: Overview of methods. A summary of the methods used in the subsequent 4 chapters is provided in this overview. This chapter mainly portrays the relationship between the different studies in the form of a practical framework of test adaptation.

Chapter 6: The translation phase- Using qualitative techniques to support the traditional back-translation method. This chapter describes the qualitative exploratory study applied, on one hand, to increase the accuracy of the translation, on the other hand to identify the common linguistic problems faced during the adaptation process.

Chapter 7: Pre-testing and Cognitive Interviews. This chapter is a quantitative and qualitative approach to assessing the accuracy of the translation. This chapter and the previous one constitute the first part of the suggested adaptation process referred to as the "Quality Control Process". The chapter is aimed in part at providing empirical evidence about the quality of the translation and also at identifying the role of pre-testing in the adaptation process.

Chapter 8: Reliability and Differential Item Functioning Analysis. The first part of this chapter investigates the reliability, difficulty and discrimination of Orpheus across the four cultures. The second part of the chapter focuses on the use of Logistic Regression in assessing differential item functioning across cultures. The results of the Logistic Regression are contrasted with the results of ANOVA for DIF detection.

Chapter 9: Measurement Invariance. This chapter explores the measurement invariance of the multi-lingual versions of Orpheus using Mplus. The analysis incorporates Exploratory and Confirmatory Factor Analysis for identifying different levels of comparability between the cultures.

Table 0.1: Chapter content

Parts of the information in this PhD are copyrighted to Harcourt Assesssment and were therefore removed. For further questions please contact the author.

Chapter 1: General introduction – The importance of cross-cultural research

1.1. Introduction

Interest in cross-cultural psychology has increased significantly over the last two decades (Casillas & Robins, 2005; van de Vijver & Hambleton, 1996; van de Vijver & Leung, 2000; Yeganeh, Su, & Chrysostome, 2004). The scope of some cross-cultural projects and the rise in the number of publications in the field reflect this considerable increase. The following are only but few examples that illustrate this.

Geert Hofstede analyzed data from 50 countries collected by IBM from 116000 employees between 1967 and 1973, to test work-related value patterns across cultures (Hofstede, 1983). His analyses yielded four dimensions that describe the values adopted by different cultures and that influence the way they operate on a business level. This had immense implications in the field of business and organisational psychology. These were mainly in terms of understanding the way organizations function in different countries and also the effect of culture and values on employee performance. In 1995, the Global Leadership and Organizational Behaviour Effectiveness (GLOBE) was launched as a multi-phase project for examining the inter-relationships between societal culture, organizational culture, and organizational leadership across 61 countries (House, Javidan, Hanges, & Dorfman, 2002). The projected expanded Hofstede's five dimensions into eighteen, though Hofstede himself argues that these highly correlate with his original five (Hofstede, 2006). The strength of this research lies in its contribution to leadership and organisational theories that might have previously overlooked cultural variables such as religion, language or political systems (Dorfman, 1996; House, Javidan, & Dorfman, 2001). In parallel but on an educational level, Trends in International Mathematics and Science Study (TIMSS) was launched in 1995 to assess science and mathematics achievement of fourth and eighth grade students from more than 45 countries in more than 30 languages (Hambleton, 2005). Founded by the International

Association for the Evaluation of Educational Achievement (IEA), TIMSS is carried out every four years and designed for monitoring changes in students' mathematics and science achievement over time and ultimately improving learning in those fields in countries all over the world (Martin, Mullis & Chrostowski, 2003). As a final example, the Programme for International Student Assessment (PISA) is an internationally standardised assessment jointly developed by more than 40 participating countries to assess reading, mathematics, and science literacy in 15-year-olds (Grisay, 2003; Le, 2006). PISA is conducted by the Organisation for Economic Cooperation and Development (OECD) and is aimed at assessing capabilities of students near the end of compulsory educations to use their knowledge and skills acquired from schools for meeting societal demands (Le, 2006). The PISA project can significantly improve the understanding and monitoring of the outcomes of educational systems in economically developed and developing countries (OECD, 2006).

A scan of PsychInfo and Educational Resources Information Center (ERIC), two renowned databases in psychology, for publication relating to cross cultural assessment over the past 40 years revealed that 134 articles were published between 1994 and 2003 in comparison to 4 articles between 1964 and 1973 (Casillas & Robins, 2005). The first (1980) and second (1997) editions of the *Handbook of Cross-Cultural Psychology* also mirror the increased action in the field in the past twenty years or so (van de Vijver & Leung, 2000). This trend is not surprising considering the rise in economic interdependence between countries (van de Vijver & Leung, 2000), internationalisation of education (van de Vijver, 1998), prominent migration streams and the rapid demographic changes in Europe (van de Vijver & Phalet, 2004). Unfortunately, the increased interest in the field of cross-cultural psychology is mainly due to the need for exploring new areas rather than a methodical build up of

knowledge on previous work (van de Vijver & Leung, 2000). That is, some test adaptation work is mainly conducted in order to expand the use of some tests into other countries rather than to build up on different methodological approaches in adaptation in order to reach a more conducive one.

1.2. Defining cross cultural assessment

"Cross-cultural psychology is concerned with the systematic study of behaviour and experience as it occurs in different cultures, is influenced by culture, or results in changes in existing cultures" (Triandis, 1980, p. 1). The aim is to assess the generalisability of psychological laws and theories across cultures (Triandis, 1980). Although cross-cultural psychology studies encompass several areas, *cross-cultural assessment* in occupational settings constitutes the main focus of this thesis. Cross-cultural assessment, as a sub area of cross-cultural psychology, involves comparing two or more groups of people, who differ on the basis of their cultural "origin", on one or more variables of interest but using psychometric tools. Psychometrics tests, tools, measures, instruments and questionnaires will be used interchangeably throughout this thesis to refer to any psychological or educational tests that have been standardised and tested for validity and reliability. The variable(s) of interest, which will sometimes be referred to as construct(s), could include intelligence, personality, specific ability or any other variable that can be measured using questionnaires.

1.2.1 Two streams of cross-cultural assessment

Globalisation is continuing to change assessment dynamics making them increasingly complicated. Assessment processes nowadays rarely include individuals from one cultural or ethnic background. More importantly though, the emergence of

Internet testing opened new horizons and made assessing candidates across national boundaries more accessible. Many multinational companies are using tests for selection and recruitment (Bartram, 2001) and several of them are making use of the Internet as a platform for these purposes (Lievens & Harris, 2003). The use of multi-lingual versions of the same questionnaire is becoming essential, though the infrastructure for achieving their successful reproduction in other languages and cultures remains underdeveloped (Daouk, Rust, & McDowall, 2005).

The increasingly diverse and cosmopolitan societies also have implications on the definition of cross-cultural assessment. Testing between cultural boundaries, whether for research or other purposes, is the traditional cross-cultural context for which trans-linguistic tests are being developed. However, testing *within* one country is becoming as complex and cross-cultural as assessment between countries. Within one country, it is common to use a version of a questionnaire in the official language of that country to assess individuals at work. Although this could be the second or third language of the assessee, some argue that this is the business language in that country and all participants wishing to work there should be proficient in it. This undeniably has consequences for the performance of non-native speakers of English (or whatever the language of the country is) as a first language, since the test might be testing them on the construct of interest but indirectly on their proficiency in the target language. Cross-cultural assessment within one culture is another stream in the field that requires a certain level of attention if the assessment process is deemed to be fair to all participants. Assessment in this context is also contributing to the increased need in developing multiple language versions of questionnaires, if comparability between them could be achieved. That is, two multilingual versions of the same tests can be used fairly if they are assessing the same criteria.

1.3. Fairness of assessment

Any assessment method needs to demonstrate fairness and freedom from bias (Baron & Janman, 1996). A selection measure is fair when it predicts future job performance and when it measures the same construct between members of a particular group and those of a standard group (Cleary, 1968 in Baron & Janman, 1996). By particular group, the authors refer to any group such as age, gender, race or any other one that can distinguish between members of one culture. Such extraneous characteristics should not affect results of what the method of assessment is measuring. However, it is important to distinguish between unfairness resulting from bias within the test (intrinsic) and bias extrinsic to the test, more commonly referred to as adverse impact (Rust & Golombok, 1999). Both types of bias lead to unfairness in assessment although the latter is a reflection of a real difference between the groups most often a consequence of social deprivation (Rust & Golombok, 1999). The differences between fairness and bias will be discussed further in chapter 4.

The same applies cross-culturally; an assessment tool is biased if it does not measure the same psychological characteristics across cultural groups (van de Vijver, 2002). For two versions of the same test to be comparable and fair, they need to be assessing the characteristic that they were developed to assess in each culture.

1.3.1 Why Psychometric instruments?

Practitioners around the world are turning to cost effective and efficient measures for assessing employees: psychometrics tests (Bartram, 2005). These are only but a few of the advantages that psychometric tests offer. However, the advantages apply only when the right test has been chosen to assess the right skills. A thorough job analysis is key to choosing the appropriate methods for assessing the core competencies necessary for a job (Anderson & Shackleton, 1993; Robertson &

Smith, 2001; Smith & Robertson, 1993). *Job analysis* entails analysing the behaviour patterns of employees which they are required to do on day-to-day basis in order to develop a *job description.* This in turn can lead to a *person specification*, which involves identifying the essential knowledge, skills and abilities necessary for performing, with competence, the tasks and functions identified in the job description (Woodruffe, 1991). Once the person specification is established, it is possible to correctly choose the methods of assessment necessary for assessing all the attributes needed for the job (Bartram, 2005).

Rust and Golombok (1999) list several advantages of using psychometric tests in organisational settings. They argue that tests are relatively easy to administer to a small or a large group of test takers, making them a cost effective method for assessing a large numbers of participants in a short period of time. Also, unlike some subjective methods of assessment, tests are objective and control for many biases that can arise during the assessment process. For example, interviews are by far the most popular method of assessment across Europe (Shakelton & Newell, 1997). Yet, there is a general agreement in the literature that their validity, specifically that of unstructured interviews, is poor (McDaniel, Whetzel, Schmidt, & Maurer, 1994; Wiesner & Cronshaw, 1988). This is in part due to the subjective nature of this method of assessment, which leads to potential interviewers' biases. However, the procedure and the questions asked in psychometric tests are structured making them less prone to such biases.

Additionally, the internet made it possible to create computer-based versions of tests and cut down on the cost of printing, mailing, and warehousing (Bartram, 2005). This is an attractive quality especially considering that moving to online recruitment was shown to cut the cost of recruitment by 90% (Cober, Brown, Blumental, Doverspike, & Levy, 2000). Moreover, Cober et al. (2000) argue that online recruitment also

reduces the time between recruitment and selection by 25%.

The most important criterion that makes using psychometric tests so attractive is the fact that they are standardised and their reliability and validity are established (Rust & Golombok, 1999). Schmidt and Hunter (1998) showed in a meta-analysis of 85 years of research on 190 selection methods, that general mental ability tests are the highest predictors of overall job performance as well as training performance. More recent studies by Salgado, Anderson, Moscoso, Bertua, de Fruyt, and Rolland (2003) and Bertua, Anderson and Salgado (2005) also mirrored these results across several occupation groups and for both general and specific ability tests. Personality tests, the construct of conscientiousness (discussed in chapter 2) in specific, and integrity tests were also shown to be among the highest predictors of performance on the job and in training (Schmidt & Hunter, 1998). Additionally, personality and integrity tests have relatively little adverse impact, an issue that causes unfairness in the selection process and poses potential legal challenges for organisations (Ones & Viswevaran, 1998).

1.4. Need for trans-linguistic tests

Not surprisingly then, psychometric tests are very widely used (Robertson & Smith, 2001) and applied in 60% of assessment centres (Ryan, McFarland, Baron & Page, 1999). However, nearly all tests used in organisational assessment are developed in Western Europe and the United States (Brown, Green, & Lauder, 2001) and approximately 50% of them are imported for use in other countries (Oakland, 2004). Although most tests are mainly developed and used in highly developed countries, foreign tests are used substantially in the least developed countries and the Middle East (Oakland, 2004). The number of indigenously developed tests is scarce, and high stakes decisions are being made about individuals around the world using Western norms. In such an internationally operating economy, it is essential to have

common grounds according to which individuals from different cultural backgrounds can be compared (Daouk, Rust, & McDowall, 2005). The use of multi-lingual versions of the same questionnaire has become increasingly essential, and test adaptation into other languages and cultures is indeed growing with the popularity of tests (Hambleton, 2005). However, validity and reliability, the most important criteria that distinguish psychometric tests from other methods of assessment, are not easily transferable from the original to the multi lingual versions of tests (Geisinger, 1994). There is a need for more *sophisticated methods* of developing and adapting multi-lingual versions of tests to ensure the equivalence between them and fairness of cross-cultural assessments.

1.4.1 Adaptation and translation

Translation can be described as a new presentation of information in one language but that was originally offered in another language (Reiss & Vermeer, 1984). Recently, translation theories have shifted tremendously in their focus. Whereas a lot of importance and analysis was put on the grammatical syntax of sentences in the early days of development of translation theories, recently the cultural aspect seems to be gaining more attention than the purely linguistic one (Snell-Hornby, 1988). As a result, the term test translation has frequently been replaced by the more accurate term, test adaptation (Geisinger, 1994). Although sometimes used interchangeably to refer to the construction of tests that require cultural and linguistic sensitivity, test adaptation encompasses broader issues than test translation (Casillas & Robbins, 2005). Examples of such issues include aspects that relate to the culture, content and wording that are fundamental for producing comparable versions across culture.

Transforming a test from one language and culture for use in other ones is a sequential process. Translation is one part of this process, which is commonly referred

to as the *adaptation process*. As Hambleton (1992) explains: "producing an equivalent test in a second language or culture often involves not only a translation that preserves the original test meaning, but also additional changes such as those affecting item format and testing procedures" (p3-4). More recently, some researchers started using the term "transadaptation" as a more accurate description of this process (for example, Downing, Bogoslaw, & Juntos, 2002; Zucker, Miska, Alaniz, & Guzman, 2005; Cohen, Gafni, & Hanani, 2007).

Poor test adaptation is the main and most common source of lack of validity of translated tests (Hambleton, 2005). However, the adaptation process is more complicated than it sounds as it addresses a number of complex questions (van de Vijver & Hambleton, 1996). Firstly, does the construct exist in the target culture? Even if it does exist, is it defined and manifested in the same way in both cultures? Additionally, are the questions in the test measuring the same construct in both cultures? Ho (1996), for example, explains that the behaviours associated with being a good son or daughter, known as filial piety, are much broader in China than in most Western countries (as cited in Byrne & Watkins, 2003; van de Vijver & Hambleton, 1996). Therefore the questionnaire used to assess filial piety in China should contain a broader set of questions than the one used in Western countries.

Another complexity in test adaptation can result from the fact that some words are non translatable and may need a "passport" in the target language (Daouk, Rust & McDowall, 2005). Some words have different connotations between two cultures even if they are literally equivalent. For example, the expression "everything is coming together" is a simple sentence that can easily be literally translated to Arabic. However, this expression is inherently positive in English but its literal equivalent in Arabic holds negative connotations. This expression in Arabic is understood as "everything *bad* is coming together *at the same time*" whereas the italicised words are

hidden in the underlying meaning. So if this expression is part of a questionnaire assessing people's positive attitude, for example, the conclusion resulting from this item in English should be reversed before it can be compared to the same item in Arabic.

1.4.2 ITC guidelines

For multi-lingual versions of a test to be equivalent, they need to be equivalent linguistically, culturally, psychologically, and also statistically. Equivalence will be the focus of chapter 4 in light of the Theory of Equivalence and Bias (van de Vijver & Leung, 1997). Although the combination of methods used in the adaptation process play an integral role in achieving equivalence between multi-lingual versions of a test, this cannot be always guaranteed (van de Vijver & Leung, 2000). Recently, the International Test Commission (ITC), the European Association of Psychological Assessment, the European Test Publisher's Group, the International association of cross-cultural psychology, the International Association of Applied Psychology, the International Association for the Evaluation of Educational Achievement, the International Language Testing Association, and the International Union of Psychological Science developed (Hambleton, 1994), field-tested (for example Hambleton, Yu, & Slater, 1999) and published (ITC, 2001) test adaptation guidelines to assist test translators and publishers in the cross-cultural adaptation of educational and psychological tests (Geisinger, 1994; Hambleton, 2001).

The guidelines were developed as a response to 1) the incoherence of the technical literature about test translation and adaptation and 2) the evidence of overgeneralisations and inaccuracy of findings due to bad application of transadaptation (Hambleton, 2001). Tests that have been shown to possess good

validity and reliability in one culture were used in other cultures to consider cross-cultural similarities and differences without ensuring that they function as well in the new cultures (van de Vijver & Hambleton, 1996). Moreover, the use of single method designs that rely either on a single translator, back translation, or bilingual judges only (discussed in chapter 5, 6 and 7) was evident in the literature for empirically judging the quality of the translation (Hambleton, 2001). All these factors, in addition to globalisation trends, have contributed to the development of the international guidelines that are available to academics and practitioners around the world. This is mostly important for the increasingly interdependent countries and if, for example, employment assessment practices are to adopt the principle of free movement that labour has between nations of the European Union (Bartram, 2001).

The ITC guidelines are divided into four main areas: context, test development and adaptation, administration, and documentation/score interpretation (Hambleton, 1994, 2001; ITC, 2001). Context, which is divided into two subsections C.1 and C.2, highlights the importance of controlling for any cultural confounding variables that might affect the results, such as familiarity with response scale or insufficient overlap between the construct in the cultures of interest. For example, participants from some countries could be more familiar with a certain response format than others, such as multiple choices, which may advantage them and bias their results. Adjustments to the instruction should therefore be made, such as adding practice exercises, to make the participants more familiar with the stimuli (van de Vijver & Leung, 1997). Test development and adaptation on the other hand (D.1-D.10) focus on the actual process of adaptation from 1) combining several judgement techniques to assess the quality of the adaptation to 2) designing and collecting data in a way that facilitates the use of appropriate statistical analysis to 3) providing evidence of linguistic, cultural, psychological and statistical equivalence between tests. Although the guidelines do

not specify how this should be done, several papers provide specific examples and procedure for achieving this (for example, Hambleton, 2001; van de Vijver & Hambleton, 1996; van de Vijver & Leung, 1997). As for the last two areas administration (A.1-A.6) and documentation/ score interpretation (I.1-I.4), these are very similar to the guidelines for test use in general in terms of minimising sources of environmental bias during test administration and ensuring confidentiality. However, they also stress on the importance of making comparison and interpreting results according to evidence of equivalence across groups (Hambleton, 2001; van de Vijver & Leung, 1997).

1.5. General aim of this research

As will be discussed in full details in the following chapter, the Big Five model of personality does not have explanatory power for explaining differences between people. Additionally, this model is highly dependent on language because it originated from a scan of the dictionary for words that distinguish people from each other. Language is therefore central to this research since the meaning of words and the way they are put together affect the description of personality.

This thesis builds up on previous research for developing the methodological approach adopted, in order to offer a framework for adapting work-based personality tests and developing comparable versions in other languages and cultures. Through this research, we explore the use of this multi-method process of test adaptation, based on the ITC test adaptation guidelines, to assess the statistical equivalence across four different languages: Arabic, Chinese (Mandarin), English, and Spanish. This is operationalised using Orpheus, a work based personality questionnaire based on the Big Five model. In the literature review, we will first focus on the concepts of personality and culture, the technicalities of the psychometric test used in this

research, and the Theory of Equivalence and Bias (van de Vijver & Leung, 1997). At the end of chapter 4, we suggest a reformulation of van de Vijver and Leung's Theory of Equivalence and Bias into a theoretical framework for test adaptation and then discuss specific hypotheses in subsequent chapters where they are being tested.

Chapter 2: Personality across cultures and jobs

"We should take care not to make the intellect our god; it has, of course, powerful muscles, but no personality."
Albert Einstein

2.1. Introduction

In this chapter, we will explore the definitions of culture and personality and the relationship between them and behaviour. This is crucial to understanding the use of objective tests for assessing personality. Personality and culture are both manifested behaviourally. Objective tests rely on such explicit behaviours to understand personality, but personality is confounded with culture, which makes such assessment of personality harder to achieve.

We will then introduce the Big Five Model, still the most prominent personality model, which forms the basis of many psychometric tests (Digman, 1990). The Big Five Model also forms the structure of Orpheus, the test that this research is based on, which will be discussed in detail in chapter 3. We will mainly focus on the impact of language on the development of this model and the effects of using language based questionnaires to assess personality.

After reviewing the literature on the origin of the Big Five Model, we will revisit some of the studies that aimed to replicate the Big Five factor structure in non-western cultures (such as Piedmont & Chae, 1997; Yang, 2000). Whilst many of these studies have managed to replicate a five factor structure, most of them failed to reproduce it with the same factor loading as the original American one. Additionally, indigenous studies conducted in China revealed that a six factor structure represents a better fit for the Chinese sample than the original five-factor one (Cheung et al., 2001).

Finally, we conclude the chapter by focusing on validation studies of the Big Five Model in the workplace. Although evidence might be conflicting in terms of the

power of personality tests in predicting certain job-related criteria (Barrick & Mount, 1991; Salgado, 1997; Furnham, Forde & Ferrari, 1999; Ones, Viswesvaran, & Diltchert, 2005; Zhao & Seibert, 2006), it is evident that personality assessment is dependent on the type of job, job analysis and other criteria provided on the job such as autonomy (Tett, Jackson, & Rothstein, 1991; Barrick & Mount, 1993).

2.2. Personality

The term "personality" is commonly used in everyday language with an understanding that it refers to psychological aspects that differentiate individuals. These are sets of enduring predispositions and tendencies of individuals to think, feel, and act in certain ways, which translate into predictable behaviour across situations (Feist & Feist, 1998; McCrae & Costa, 2003, Ones, Viswesvaran, & Dilchert, 2005). While there is broad agreement in the literature about the stability of personality across time and situations, there is inconclusive evidence about the exact influences of nature and nurture in shaping it; although evidence of both has been reflected in a number of twin studies (McCrae & Costa, 2003; Rust & Golombok, 1999). Many psychological theories converge on the idea that both nature and nurture play a role in shaping behaviour but diverge in the amount of emphasis their theories put on either. For example, Jung recognises the influence of nurture, such as parental influence, on personality but he considers that biologically shaped predispositions are the determining factors of one's personality (Feist & Feist, 1998). Similarly, Eysenck argues that strong biological basis exists for some personality characteristics (neuroticism, psychoticism, and extraversion) but not for others (agreeableness and conscientiousness) (Feist & Feist, 1998). The influence of nature and nurture have also been demonstrated in the workplace, where studies of monozygotic twins reared apart showed that part of the variation in job satisfaction was due to genetics (30%)

but the other part (70%) was due to environmental and other factors (Arvey, Bouchard, Segal, & Abraham, 1989).

We argue that there are three essential parts to understanding personality and its assessment and these are: biological factors, culture and behaviour. Behaviour is the end product that is used to make assumptions about an individual's personality. However, personality is influenced by both culture and genetically inherent characteristics, which are both manifested behaviourally. We will therefore explore the definition of culture before focusing on assessment of personality using psychometrics.

2.3. Culture

Rosinski (2003) explains that a "group's culture is a set of unique characteristics that distinguish its members from another group" (p 20). A "group" could be a nation, an organisation, a society, a gender, an age or any other group distinguishable from others (Rosinski, 2003). Therefore it is not only nations that have their own "unique set of characteristics" or "culture" but also other groups such as women, elderly, political groups and so on.

Every group's culture could have an effect on the individual's behaviours and each individual has several cultures that he or she belongs to. That is, an individual's identity could be considered as "the personal and dynamic synthesis of multiple cultures" (Rosinski, 2003, p 21). A person can be German, a student, a female, an atheist and many more simultaneously. Nevertheless, the stronger the culture of the group, the more likely it is to predominate in shaping behaviour and forming identity (Peterson, 2007). For example, a strong organisational culture can lead employees to exhibit behaviours (goal orientation, people orientation etc) that are congruent with the culture of their organisation rather than their national culture.

The challenge in understanding cultural differences is that cultural factors might sometimes be overgeneralised, which could lead to stereotyping (Peterson, 2007). In other words, a good understanding of a national culture might lead to wrongly assuming that all its nationals will typically exhibit certain behaviours associated with that culture. This is not always true because of the effect of other cultures as well as personality on behaviour. For example, it is possible for the cultural gap between a young Japanese girl and her grandparents to be bigger than between her and a young French girl. In this case, the age and gender cultures are stronger than the national culture, which might have created less stereotypical nation related behaviours (i.e. Japanese or French) but more teenage girly behaviours. Therefore, key to understanding cultural differences, is to understand that people from different cultural backgrounds can be similar in many ways such as how organised or punctual they are, even when these characteristics are not typical of their national culture (Peterson, 2007).

While culture is multifaceted, in this thesis we will mainly focus on one type of culture and use this term primarily to refer a group level phenomenon that derives from the national origins of the individual (i.e. Chinese, Italian, Hungarian), and is typically shaped by history, religion, shared values, politics etc, and affects individuals from that group collectively.

This is not a definition of culture but as a fragmentation of the complex notion of culture to explore one specific aspect of it. This should inform our understanding about national cultures but should not limit our view of culture to a static one.

2.4. The relationship between personality and culture

Hofstede (1981), one of the pioneers in cross-cultural research, distinguishes between three levels of mental programmes that influence our behaviour, namely:

individual, collective and universal. The universal level contains biologically-shaped behaviour shared among all human beings such as crying and laughing. The collective level on the other hand, contains all that is shared between certain groups of people, such as language, gestures, comfortable physical distance maintained between individuals and so forth. Finally, the individual level is unique to every human being and is characterised by personality. Each person's behaviours are the product of individual, collective and universal mental programmes (Table 2.1).

Hofstede's mental programmes	**Influenced by**	**Observable outcome**
Individual	Personality	Behaviour
Collective	Culture	Behaviour
Universal	Biological	Behaviour

Table 2.1 Hofstede's Mental Programme

The various theories discussed in this chapter differ in the approach and terminology they employ but the unifying issue is that personality and culture have been recognised as playing a role in shaping behaviour. In practice, psychometric tools that purport to assess personality do so by measuring behaviours that are considered as the manifestation of specific personality characteristics. However, it is practically impossible to measure personality while ignoring all cultural influences that may affect respondents' replies and consequently their assessment reports. Thus, culture can act as a confound variable in personality assessment as shown in figure 2.1 below and should be controlled for to ensure fair and accurate assessment.

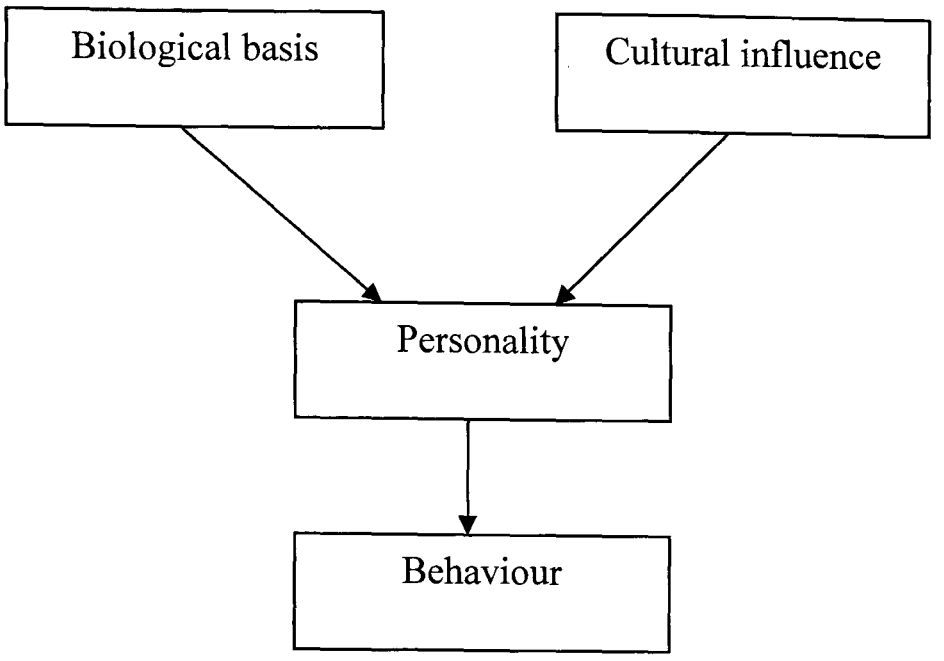


Figure 2.1: relationship between culture, personality, genetics and behaviour

2.5. Methods of Personality Assessment

Although assessment of personality is quite challenging, it is still of great concern to many psychologists and practitioners alike. Interest in measuring personality began in the 19th century mainly for detecting psychopathology and could be categorised as projective and objective assessment (Hogan, 2008). Projective tests rely on abstract or figurative pictures that individuals interpret whichever way they perceive them. The assumption is that individuals will project their own experiences onto the picture. This type of tests is usually used to understand the individual's mental representations in addition to getting the individual to articulate unconscious conflicts (Urist, 1977). This is based on the psychoanalytic assumption that individuals sometimes project some of the conflicts, emotions and feelings that they cannot tolerate onto external objects around them (Cattell & Kline, 1977; Feist & Feist, 1998). The Rorschach and the Thematic Apperception Test (TAT) are examples of projective tests that were used for assessing personality clinically and also in the workplace (Hogan, 2008). Projective tests are no longer used for assessing employees in the workplace due to their questionable validity and inappropriateness within that

context (Rust & Golombok, 1999).

Objective assessment gained momentum after World War II with the increased interest of the American Air Force in predicting officers' effectiveness and screening them for psychiatric problems (Digman, 1990; Hogan, 2005). However, practitioners and theorists alike lost interest in personality assessment at the end of the 1960s when research findings drew attention to response distortion and its challenge to validity. Response distortion is characterised by respondents faking their answers and presenting themselves in a way that does not characterize them accurately (Barrick & Mount, 1996). This issue will be discussed further in chapter 3 and 4 as it constitutes a significant challenge to personality assessment.

Personality measurement regained attention in the 1990s mostly in the field of organisational psychology as a result of methodological advances that provided evidence of their validity in predicting job-related criteria (Baron & Janman, 1996; Hurtz & Donovan, 2000; Hogan, 2005). For example, evidence from Barrick and Mount's (1991) meta-analysis revealed that some personality traits from the Big Five model are good predictors of future job performance. A more important reason for their revival is that personality tests, unlike ability tests, have been shown to have little adverse impact (Hogan, 2005; Ones & Viswevaran, 1998). That is, personality tests are less likely to discriminate between groups of people based on criteria that are unrelated to what the test is measuring.

Objective personality tests are usually used as a method of assessing personality traits by asking a list of questions, clusters of which measure a specific trait (Ones, Viswesvaran, & Dilchert, 2005). When individuals respond to single items on personality inventories, they are providing self-descriptions about themselves. These descriptions can lead to an understanding about their typical behaviour and consequently their personality. Some psychometric tests adopt a *type approach* and

some others adopt a *trait approach* to personality. Type questionnaires categorise individuals as one particular personality type (Rust & Golombok, 1999). For example, a type questionnaire will classify an individual as either an extrovert or an introvert. Trait questionnaires on the other hand describe the individual's preferences on a continuum. In contrast to Type questionnaire, a Trait questionnaire can lead to describing a person as striking a balance between extraversion and introversion. Some argue that types are merely extreme scores on continuously distributed trait dimensions (McCrae & Costa, 2003). Based on a trait approach, the Big Five model, also known as the five factor model (FFM), forms the basis of many psychometric personality tests and is argued to be the best universal representation of personality structure for reasons that will be explained in the next section (Digman, 1990; McCrae & Cost, 1997). According to the FFM, personality can be described through five main traits that represent the extreme end and these are: Extroversion/Introversions; Agreeableness/Tough mindedness; Neuroticism/Emotional Stability; Openness to experience/ Close mindedness; and Conscientiousness (opposite not named) (Digman, 1990; McCrae & Costa, 1997; McCrae & Costa, 2003).

2.6. Why the Big Five?

Unlike evidence based research such as Eysenck's, the Big Five Model is not a theory of personality but a data driven description of some components that differentiate between individuals (Hogan, 2008). Although a five-factor structure first emerged in the 1930s (Rust & Golombok, 1999), McCrae and Costa (1985 in Digman, 1990; McCrae & Costa, 1997; McCrae & Costa, 2003) played a major role in making it a popular model after developing their Big Five questionnaire NEO-PI-R® and testing it across many cultures and languages (Digman, 1990). They adapted their

questionnaire to Germanic (i.e. German), Indo-European (i.e. Portuguese), Hamito-Semitic (i.e. Hebrew), Sino-Tibetan (i.e. Chinese), Bantu (i.e. Kenya), Malayo-Polynesian (i.e. Malay), Uralic (i.e. Finnish), Altaic (i.e. Korean) language families[1] and several other languages and managed to replicate a five-factor structure as found in their original American sample (McCrae & Costa, 1997; McCrae & Costa, 2003). However, the replication of the factor structure did not always follow the exact composition of the original American one, an issue that will be critically discussed later in this chapter. Additionally, some indigenous studies that aimed to replicate the Big Five Model found that a six factor model was a better representation of personality in China (Cheung et al, 2001). These two points will constitute much of the discussion that will follow in chapters 6, 7, 8 and 9.

2.6.1 The Origin of the Five Factor Model

Allport and Odbert first instigated the emergence of trait psychology in 1936 when they decided to study language in order to better understand personality (Digman, 1990). They conducted a lexical search in the English dictionary to gather all the trait names that are used in the English language to differentiate between people (Digman, 1990; Dawda, 1997; Rust & Golombok, 1999; Costa & McCrae, 2003). However, Allport & Odbert were not the first to pick up on the importance of language as a tool to differentiate between people. Galton in 1884 was actually the first to suggest that "individual differences between people would have become encoded throughout history in single linguistic terms that would occur in all the world's languages" (as cited in Rust & Golombok, 1999, p 155). He foresaw that people would use language to refer to any differences or similarities they encounter

[1] "Language families are groups of languages with a common historical origin that have cognate terms and share certain features of grammar and syntax." (McCrae and Costa, 1997, p510)

between each other. During their search, Allport and Odbert found around 17,000 trait names of which nearly 5,000 were personality trait names.

Cattell (1946) followed up on their study and managed to group the 5000 trait names into 35 clusters (Digman, 1990; Rust and Golombok, 1999; Costa & McCrae, 2003). He then subjected this data to peer rating and factor analysis to find 16 factors, which later formed the basis of his renowned personality questionnaire 16PF®. Interestingly, no-one in the literature has ever reported the same complex structure as Cattell's (Digman, 1990). Yet, the 16PF is still a very widely used questionnaire.

Fiske (1949) adapted Cattell's 35 clusters into 22 traits and subjected them to self, team mate and supervisor ratings. The factor analysis results showed a clear five-factor solution across the three ratings. This study is known to have marked the emergence of the five factor model, though Rust and Golombok (1999) argue that Thurstone (1936) was the first to attain such a factor structure while Barrick and Mount (1991) attribute it to McDougal (1932).

In 1961, Tupes and Christal, the acclaimed originators of the Big Five model alongside Fiske and Thurstone (McCrae & John, 1992), conducted studies in the American Air Force using Cattell's scales. They reported a consistent five-factor solution across a series of six studies (Digman, 1990; McCrae & John, 1992; Rust & Golombok, 1999; McCrae & Costa, 2003). Norman in 1963, Borgatta in 1964 and Smith in 1967 were also able to replicate a five-factor structure following up from the work of Tupes and Christal and Cattell (as cited in Digman, 1990).

2.6.2 Collectivism and Individualism

The main work on the five-factor model and the replication of the NEO-PI-R structure discussed above was conducted on North American samples (Yang, 2000). North America is considered to be an individualistic type of culture and the patterns

found in such samples may not necessarily be representative of personality across other cultures particularly not collectivistic ones. Triandis and Gelfand (1998) argue that individualist and collectivist societies differ on four basic attributes:

1. how the individual defines their self;
2. whether personal goals have priority over in-groups goals and vice versa;
3. whether relationships the individual forms are exchange relationships that are by choice or communal ones that are predetermined; and finally
4. whether the individual views attitudes are more valuable than norms or vice versa.

Hofstede (1980) explains that individualist cultures emphasize the "I", autonomy, emotional independence, individual initiative, privacy, pleasure seeking etc (also see Brewer & Chen, 2007). However collectivist societies value the "we", collective identity, emotional dependence, group solidarity, sharing, duties and obligations etc. These cultural factors have huge indirect implications on behaviour and consequently on personality, which would be reflected in the personality structure on a group level in either types of culture. Further, Triandis and Gelfand (1998) argue that individualistic cultures are not all the same, and that neither are collectivistic ones. They distinguish between horizontal collectivist (HC), horizontal individualist (HI), vertical collectivist (VC), and vertical individualist (VI). The difference between them is based on social relationships and can be illustrated in the following examples: whilst both are individualistic, HI people want to be unique and distinct from the group whereas in VI people want to acquire status and be distinguished through competing with others. Similarly within collectivistic cultures, HC people are characterised by individuals who see themselves as similar to others and follow shared goals without necessarily submitting to authority. However, VC people are willing to sacrifice their own goals for the in-group and submit to the will of authority figures. This distinction suggests that the five-factor structure might differ or even not hold in

other collectivistic cultures.

2.6.3 Replication of the five factors across cultures

The replication of Allport and Odbert's 1936 study in other non North American cultures has yielded inconsistent results (McCrae & Costa, 1997). In Germany, for example, factor analysing the trait names found in the dictionary and other references replicated the five-factor structure (Ostendorf, 1990 in McCrae & Costa, 1997). However, a similar indigenous study in Hungary found that only four out of five factors were found to replicate (De Raad & Szirmak, 1994 in McCrae & Costa, 1997). Similarly in China, Yang and Bond (1990) used Chinese descriptors and others derived from the Tupes and Christal (1961) study to replicate the five-factor structure, which they did. Again, only four out of the five factors mapped on to the Big Five.

Other studies attempted to replicate the five-factor structure either through adapting Western Big Five measures or by developing indigenous personality questionnaires. Piedmont and Chae (1997) for example adapted the NEO-PI-R® into Korean and tested its validity compared to the Korean version of MBTI® (Myers Briggs Type Indicator). In their first study, they compared the scores of Korean and American test takers on the English version of NEO-PI-R® and found that scores on four out of five factors were significantly different. However, a follow up study where Koreans were given both the English and the Korean versions showed a similar response pattern on both versions. Although this might have been the result of a methodological problem, Piedmont and Chae interpreted the findings as a cultural difference, since participants who took both versions scored similarly. Similar results where also reported by McCrae, Costa and Yik in 1996 with the Chinese culture in

Hong Kong (Piedmont & Chae, 1997; Yang, 2000). Although a five-factor structure was replicated in many cultures, the factor loadings were not consistent which has implication for the comparability of personality across cultures. Factor analysis usually works by showing which questions (or items) are loading together. That is, questions designed to assess a particular Big Five characteristic, should load together. These studies showed the emergence of five factors but with discrepancies in factor loading. For example, questions that assess Agreeableness might have conglomerate or loaded randomly under the five factors, so there is no factor that is assessing Agreeableness per se. It therefore becomes arguable that the Big Five are not necessarily the same across all cultures. Obtaining five factors in different cultures is not full evidence for the universality of the BFM, as whilst the structure might replicate the content can differ.

Cheung et al. (2001) developed an indigenous personality test for the Chinese culture arguing that adapting tests from Western cultures is bound to omit culture specific characteristics. Moreover, Cheung et al. (2001) argue that it is unlikely to find a factor solution that better fits other cultures based on questionnaire that assumes that five factors constitute the best fit. In a series of studies, Cheung et al. gave Chinese students the indigenous Chinese Personality Assessment Inventory (CPAI) (Cheung et al, 1996) and the Chinese version of NEO-PI-R®. A confirmatory factor analysis revealed that a model with 6 factors fits the data best. The first five factors mapped onto the Big Five whereas the sixth factor was labelled "interpersonal relatedness" and assumed to be indigenous to the Chinese culture as it covers the "interdependence concern in the Chinese personality" (Cheung et al., 2001, p425). This characteristic mainly relates to the part of personality that is influenced by living in a collectivistic society that emphasises shared values. Additionally, when the researchers subjected the NEO-PI-R® data to exploratory factor analysis they found that extraversion, and

agreeableness cross-loaded, whereas neuroticism loaded on several factors. This suggests that the content of extraversion, agreeableness and neuroticism is different between China and Western cultures where the Big Five originally emerged, such as in the US and Europe.

To address the issue that a student sample may be unrepresentative of the Chinese culture, Cheung et al. (2001) replicated this same study with a non-student sample and found similar results in the exploratory factor analysis whereby agreeableness and extraversion cross-loaded, and neuroticism had mixed loading on several factors. CPAI showed that interpersonal relatedness loaded independently again with the non-student sample, leading again to a six factor solution with better fit than a five factor one in the Chinese sample.

To investigate whether the six-factor model is superior to the five-factor one, Cheung et al. (2003) replicated the same study on a non-Chinese sample. The six-factor model still fit the data in the confirmatory analysis with the interpersonal relatedness loading independently as the 6^{th} factor. However, unlike their findings with the Chinese sample in the previous studies, the exploratory analysis showed that the Big Five factors retained their original structure with the non-Chinese multiethnic group. Cheung et al. (2003) deduced that extraversion and agreeableness need further investigation, especially in collectivistic cultures since the pattern of cross loading was evident in more than one of their Chinese samples as well as in other studies in Asian countries.

McCrae and Costa (1997) argue that the five-factor structure is in fact universal as it replicated in several cultures; however, comparability between cultures cannot necessarily be established because of differences in factor loading. That is, regardless of the loading and cross loading of individual items on factors, personality across cultures can be summarised with five main characteristics.

Whilst the findings from Cheung and her colleagues' study suggest that a six-factor model represents a better fit to the Chinese culture, they are inconclusive in terms of the universality of this model. However, it is evident that Western developed tests and theories are not fully representative of personality across cultures. Also, the six-factor model seems to undeniably fit the Chinese culture better than the five-factor model, and further research in this area is needed to investigate whether the same would apply to other cultures.

2.6.4 Big Five at work

Personality assessment has increasingly become central to the activities of work psychologists as it has been for personality psychologists for years. It is argued that since organisations are made of people, it is important to understand them in order to better manage them and to optimize organisational outcomes and meet organisational needs (Bartram, 2004; Hogan, 2004). However, for many years, personality assessment has been consistently criticised for being a lower predictor of job performance in comparison to cognitive ability tests (Hurtz & Donovan; 2000). Some argue that this is mainly due to the lack of agreement on one ubiquitously accepted personality taxonomy (Barrick & Mount, 1991). Conversely, Tett, Jackson, & Rothstein (1991) argue that some of the problems in studying the validity of personality tests in predicting job performance are methodological ones. They explain that there are several meta-analytic issues that are often overlooked, such as analysing confirmatory and exploratory studies together or analysing significant findings from matrices with non significant correlations, which can distort the findings if no corrections are applied.

Nonetheless, several large-scale meta analyses have reported the predictive

nature of some personality constructs for several aspects of organisational behaviour and outcomes, such as overall job performance, training performance, team performance, entrepreneurial skills, managerial and leadership skills (Barrick & Mount, 1991 and 1993; Tett, Jackson, & Rothstein, 1991; Salgado, 1997; Ones, Viswesvaran, & Diltchert, 2005; Zhao & Seibert, 2006), job satisfaction (Judge, Heller, & Mount, 2002), and motivation (Furnham, Forde & Ferrari, 1999; Ones, Viswesvaran, & Diltchert, 2005). However, some of these constructs were shown to be consistently more predictive than others although this depends firstly on the performance criteria and secondly on the occupation group of interest. While many studies focused on other personality models and their relation to workplace behaviour, we will concentrate on the studies on the five-factor model.

As discussed earlier, Barrick and Mount's (1991) meta-analysis was instrumental in drawing the attention back to personality assessment in the workplace. While investigating the predictive validity of the Big Five factors across five job categories (professionals, police, managers, sales, skilled/semi-skilled) and three job criteria (job performance, training performance, and personnel data), Barrick and Mount (1991) found that conscientiousness consistently predicted performance across all jobs and criteria. Salgado's (1997) meta-analysis mirrored part of these findings in the European Community (EC) whereby (high) conscientiousness and (low) emotional stability had the highest validity criteria among the Big Five for predicting job performance in the EC across all jobs though not for managers group. Similarly, another meta-analysis by Hurtz and Donovan (2000) also found conscientiousness to be the highest predictor of job performance among the Big Five.

Although conscientiousness has been shown to predict job performance, this might not be true across all types of jobs and work environments. For example, Salgado's (1997) meta-analysis also investigated the relationship between the Big

Five and training performance in the European Community (EC) from studies between 1973 and 1994. Unlike his findings with overall job performance, conscientiousness was not as good of a predictor of training performance as were Openness to experience and Extraversion. Robertson, Baron, Gibbons, MacIver, and Nyfield (2000) argue that this might be due to the complexity of the construct of conscientiousness. They explain that, in NEO-PI-R for example, conscientiousness is measured as a single construct but covers nearly six different facets. Some of the qualities associated with these facets are clearly associated with conscientiousness (i.e. careful and orderly), some others are not (i.e. achievement).

The confounding findings about conscientiousness across jobs highlight the idea that there is no "perfect personality" that would be suited to all jobs. Personality and job performance cannot be considered in vacuum since these two are strongly influenced by other environmental factors in the workplace and also the work criterion of interest. This is another draw back to assessing the validity of personality assessment because personality can only predict work performance as a function of some other elements in the work environments. To illustrate, Barrick and Mount (1993) investigated the relationship between the Big Five and job performance specifically in low and high autonomy jobs. They distinguished between work environments with low autonomy, which restrict the amount of behaviour associated with personality that individuals can exhibit at work, and high autonomy environment, which maximise this kind of behaviours. Their findings revealed that managers with high scores on extraversion and conscientiousness performed better in high autonomy jobs whereas those with low scores on agreeableness performed better in low-autonomy jobs. This implies that there are other factors that can affect job performance, other than personality characteristics, that should be carefully considered when assessing validity of any method of assessment.

Similarly, studies investigating teams through the five-factor model have yielded invaluable information about team performance, communication between team members, likelihood of lasting together, and workload sharing. However, the findings they yielded about personality assessment are very much associated with other factors such as cognitive ability. For example, Barrick, Stewart, and Mount (1998) investigated the influence of the individual level characteristics on overall team performance. Their findings showed that teams with high conscientiousness (mean score) and high cognitive ability perform better than low conscientiousness and low cognitive ability teams. They also found that Emotional stability and Agreeableness are paramount for higher team performance, and that interestingly enough, teams with no disagreeable or introverted individuals performed better than other teams. As for communication, teams that did not include members with low conscientious reported more communication and workload sharing and less conflict than other teams. However, teams with highly a disagreeable member reported more conflict, less communication and less workload sharing. Team functioning, creation and make-up are widely discussed in the literature though most of the available information is descriptive and based on case studies (Furnham, Steele, & Pendleton, 1993). These findings provide empirical evidence about personality factors influencing team performance and functioning. This accentuates the value of personality assessment in the workplace given that there is a need to assess people individually, for instance in recruitment, but also for other purposes such as team building and development.

2.7. Summary

Addressing the validity of personality assessment is a complex area of research due to all the confounding factors associated with job performance and also the methodological issues which may affect any results. Nevertheless, personality

assessment has been shown to be of importance in personnel selection, especially when the personality traits assessed are based on a thorough job analysis (Tett, Jackson, & Rothstein, 1991). Researchers and academics are heading towards an agreement about the Big Five model as a comprehensive taxonomy of personality universally (Barrick & Mount, 1991), though alternative models for predicting job performance are also being explored such as competency based ones (e.g. Batram, Kurz, & Baron, 2003). However, further indigenous research has revealed that a six-factor structure might be considered a better fit in some cultures as it was shown in China (Cheung et al, 2001). Other studies based on the adaptation of Western-developed tests have consistently revealed a five-factor structure, though the factor composition in each culture was not exactly the same. In this research, we adopt a confirmatory approach to investigate the stability of the five factor model in the Arab world, China and Spain through simultaneous adaptation of a Western developed Big Five measure, Orpheus. We argue that a replication of the five factor structure across these cultures could be considered as evidence of the universality of the Big Five model, though not necessarily as evidence that the Big Five Model is the best fit model for assessing individuals across cultures. In the next chapter, we will outline the psychometric properties of Orpheus, the factors that constitute it, and the audit scales that were designed to control for response bias.

Chapter 3:The Instrument-Orpheus work-based personality test

"A personality test, like music, is an expression of the inner self. Orpheus represents the conscious expression and Eurydice represents the inspiration, but also the unconscious".
John Rust

3.1. Chapter overview

This chapter provides an introduction to the psychometric properties of Orpheus personality questionnaire and other technical qualities. We will first start by describing the theory behind Orpheus and the interpretation of its scales and then describe other technical characteristics associated with it such as the audit scales and within subject standardisation. We will conclude the chapter with a summary of the technical qualities of Orpheus in comparison to a selection of other widely used tests in occupations settings in the UK such as NEO, OPQ32 and HPI.

3.2. Description of Orpheus

Orpheus© (Rust, 1996) is a work-based personality questionnaire developed for selection, promotion, appraisal and staff development purposes (Rust, 2001). The test measures the individual's preference at work though 190 items and is scored on a four point Likert scale (1=Strongly Disagree, 2=Disagree, 3=Agree, 4=Strongly Agree). These items give information about five major scales, seven minor scales, and 4 audit scales. The first five scales (or factors) Fellowship, Authority, Conformity, Emotion and Detail measure the Big Five traits of personality but in relation to work related attitudes, beliefs, behaviours and interests (Rust & Golombok, 1999). Traits are relatively enduring measurable variables that account for the differences in styles of thinking, feeling and acting between people and are descriptive of the person's typical state (McCrae & Costa, 1997; Cronbach, 1990; Kline, 1993). A trait is usually described in terms of extreme ends, for example "good" on one end and "bad" on the other. "Good" and "bad" are just examples of two extremes but this example does not

imply that one extreme of a personality trait is good and the other is bad. Moreover, the assumption is that these two ends of a trait are on a continuum and people could lie anywhere between them (Rust& Golombok, 1999).

The big five traits are generally referred to as: Extraversion vs. Introversion, Tough-mindedness vs. Agreeableness, Openness to experience vs. conservatism, Neuroticism vs. Emotional stability, and Conscientiousness (opposite not named) (Barrick & Mount, 1993). Although Orpheus measures the big five, it employs different trait names that are more appropriate for use in the workplace. For example, the term Neuroticism has connotation that might be offensive in the work context and was therefore replaced by Emotion.

3.3. Orpheus five major scales

Orpheus reconceptualises the big five model as a domain theory of personality (Rust, 1996). The five domains are Social, Organisational, Intellectual, Emotional, and Perceptual and are essential contributors to the individual's psychological life (Rust & Golombok, 1999). The following table 3.1 describes the relationship between the big five, Orpheus scales, and the five domains.

The Big Five Model (Costa and McCrae, 1992)	**Orpheus Scales**	**Domains**
Extraversion vs. Introversion	Fellowship	Social
Tough-mindedness vs. Agreeableness	Authority	Organisational
Open mindedness vs. conservatism	Conformity	Intellectual
Neuroticism vs. Emotional stability	Emotion	Emotional
Conscientiousness (other side not named).	Detail	Perceptual

Table 3.1: Mapping Orpheus scales and domains to the Big Five Model

The five Orpheus factors are measured on a stanine scale that ranges from 1 to

9, and are described in terms of extreme scores on either end. However, individuals could lie anywhere on the scale and could be described as having a strong, moderate or slight preference to one side or the other of the scale. The Orpheus technical manual (Rust, 1996) and Rust and Golombok (1999) have been used as the main references for the following section.

Fellowship scale assesses the Extraversion/ Introversion trait of the big five model. This scale represents the social domain, which describes the way people interact and form relationships with others. People differ in how much they like to be around others, and how much interaction they like to have on day-to-day basis. This characteristic would also be reflected in their preferences at work. For example, low scorers on fellowship tend to prefer a certain degree of independence in the workplace whereas high scores are likely to perform at their best when working with others or having lots of interaction with clients and /or colleagues at work.

	Independence at work							Working with others
Fellowship	1	2	3	4	5	6	7	8	9

Authority, on the other hand, assesses the tough-mindedness/agreeableness trait. The organisational domain people live in influences their position on the Authority scale. Hierarchies and social status are both a consequence and a determinant of the individual's level of Authority. Some people prefer to exercise more power and make more tough decisions while others might prefer to make decision collaboratively. However, some might prefer a moderate balance between the two. These preferences could reflect where individuals would work at their best within the organisational hierarchy. Low scores on this scale tend to prefer a more collaborative approach in making tough decision while high scorers on this scale tend to enjoy making tough decision on their own without having to consult with others.

	Collaborative approach in decision-making							Making tough decision independently
	2	3	4	5	6	7	8	9

The third trait, **Conformity**, assesses the Openness to experience/ conservatism trait, first suggested by Rokeach (1960) then Rogers (1961) (as cited in Rust & Golombok, 1999). Conformity corresponds to the Intellectual domain where reason and knowledge are the basis of our judgment. Our beliefs and attitudes towards the methods that should be used at work are reflected in our judgement about how to approach work. Low scorers on conformity tend to be more innovative, prefer to explore new solutions for problems and enjoy finding new ways of doing things at work. This scale does not measure innovation per se. It measures whether people *enjoy* developing new ideas and new ways of doing things at work but not *how good* they are at doing so. High scorers, on the other hand, generally prefer working according to an established set of rules or guidelines and to follow tried and tested methods in their work.

	New ways at work							Following tried and tested methods
Conformity	1	2	3	4	5	6	7	8	9

The Big Five's Neuroticism trait is assessed through the **Emotion** scale. Evidently, this trait represents the Emotional domain that drives individuals on everyday basis. The control people have over their emotions tends to influence their motivation at work. The more in control people are of their emotions, the more likely they are to enjoy working in hectic atmospheres because it does not "get to them" easily. However, the less people are in control of their emotions, the more likely they are to be affected by the ups and downs of performing several tasks on the go. Low scorers on this scale tend to perform better under stress and to enjoy juggling several tasks on the go. Conversely, high scorers tend to be more sensitive to other people's

feelings at work and might prefer working in less hectic atmospheres where they can concentrate on few tasks at a time.

	High stress tolerance							Low stress tolerance
Emotion	1	2	3	4	5	6	7	8	9

The last trait, **Detail,** assesses Conscientiousness. It represents the Perceptual domain in the sense that individuals' actions are the result of their views about what is significant and important in the world. People are different in their organisational skills and the attention they like to give to details and perfection. Some are more concerned with the overall concept, its implications and applications rather than its details. And again, some like to strike a balance between the two. Low scorers on detail are likely to be interested in the broader view of problems and are less tolerant for repetitive tasks. However, high scorers stand out in the extent to which they perfect tasks that require particular care.

	Bigger picture							Tolerance for routine
Detail	1	2	3	4	5	6	7	8	9

Table 3.2 provides examples of positive and negative items for each scale whereas appendix 1 presents all the items that measure the five major scales in Orpheus.

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

Table 3.2: Example of positive and negative items for each Orpheus scale

The seven minor scales are Proficiency, Work-orientation, Patience, Fair-mindedness, Loyalty, Disclosure and Initiative and are based on Prudentius' theory of integrity (Rust & Golombok, 1999). Integrity tests are designed to assess individuals' tendency to engage in behaviours that are commonly viewed as counterproductive, such as theft, by measuring their attitudes towards those issues (Ones & Viswesvaran, 1998). However, the items measuring integrity in Orpheus have a disguised purpose since they do not overtly ask about attitudes towards counterproductive behaviour. Questionnaires with this type of questions are usually called personality-based integrity tests and are presented in appendix 2 (Ones & Viswesvaran, 1998).

Orpheus can therefore be described as a personality questionnaire that measures the big five traits and seven integrity traits in the work context. As discussed earlier, the Big Five Model is based on language, which constitutes the basis of this Thesis. Integrity on the other hand does not share the same linguistic origins as the BFM and will therefore be excluded from the analysis of the Thesis. Orpheus will be treated as a measure of the big five at work only. However, the audit scales Dissimulation, Ambivalence, Despondency and Inattention that monitor any attempt of manipulating responding will be used because they provide valuable information about response patterns. Two of the audit scales reply of mathematical methods for assessing response patterns, which makes them valuable for use in cross-cultural settings. The four scales will be discussed in full detail in the next section

3.4. Audit Scales

Whilst personality tests are regaining interest as pre-employment screening method, it is argued that their biggest flaw is that respondents can inflate their scores if they want to (Rosse, Stecher, Miller, & Levin, 1998). For example, Furnham (1997) showed that, when instructed to fake their responses on the NEO-PI-R® S form,

participants were able to do so but only on agreeableness, neuroticism, and conscientiousness scales. Ballanger, Caldwell-Andrews and Baer (2001) also reported similar results with a clinical sample whereby participants were able to fake their scores on extraversion, neuroticism, agreeableness, and conscientiousness but not on openness to experience. In an attempt to prevent test takers from faking their responses, Heggerstad, Morrison, Reeve and McCloy (2006) compared using Likert scales and multidimensional forced-choice response formats to control for response distortion. They found that participants were still able to fake their response with either response formats. These studies reflect how difficult it is to prevent test takers from distorting their answers when they wish to do so. Therefore *monitoring* or *measuring* the amount of response distortion seems to be the most suitable approach that test developers can take.

Conversely, Ones, Viswevaran, and Reiss (1996) showed in a meta-analysis that controlling for social desirability in Big Five measures does not affect their predictive validity of job performance, task performance, counterproductive behaviour and (to a lesser extent) training performance. However, Rosse, Stecher, Miller, and Levin (1998) argue that Ones et al. did not consider the type of job as a moderator of these relationships. That is, the prediction will not be the same for different types of jobs. Additionally, their study showed that the rank order of applicants to be hired changed significantly after adjusting for response distortion on the conscientiousness scale, which is known to be the strongest predictor of performance at work (e.g. Barrick & Mount, 1991). This indicates that response distortion has huge implication on the selection of future employees. Some employees are getting onto the job by giving a more positive impression of themselves, and as a result jeopardising more suitable employees' chance of getting through the selection process. Paradoxically, some jobs require candidates to have a high level of impression management such as

PR and sales jobs. Impression management should always be considered in the context of the job in order to accurately establishing its prediction of job performance.

3.4.1 Reasons for response distortion

There are several reasons for distorting responses, whether intentional or not, that have been reported in the literature such as 1) motivation, as in the case of job applicants 2) lack of self-insight 3) intentional sabotaging of tests or 4) cultural orientation. For example, Rosse, Stecher, Miller, and Levin (1998) showed that response distortion was higher among job applicants than among job incumbents. This indicates that job applicants tend to present themselves in a more positive light in order to increase their chances of getting the job. Therefore their motivation is different than that of actual job incumbents. Lack of self insight and intentional sabotages have been widely discussed as possible explanations of response distortion (Furnham, 1997). As for the cultural orientation, Extreme Response Style (ERS) and Acquiescence Response Style (ARS) are two forms of response distortion that have been showed in the literature to differ across cultures (Cheung & Rensvold, 2000). For example, groups high on ERS tend to come from cultures that value sincerity and conviction, and therefore choose extreme responding to reflect that. Similarly, ARS is characterised by one group consistently scoring higher or lower than other group(s) (Rust & Golombok, 1999), which could be the result of culturally induced behaviours (Cheung & Rensvolt, 2000). This issue will be discussed in more detail in chapter 4.

3.4.2 Orpheus and response audits

Orpheus uses four audit scales to measure different types of response

distortion, and these are: Ambivalence, Dissimulation, Despondency, and Inattention. Scores on these scales range form 0 to 3, with 0 indicating no distortion of scores and 3 indicating extreme and deliberate distortion of scores. Reports with scores of 3 on any of the audit scales will be produced with a warning that the results are of questionable value.

3.4.2.1. Dissimulation and Despondency

Dissimulation measures social desirability responding, that is faking good whereas Despondency measures faking bad. Both of these are measured though the disclosure scale, one of the seven integrity scales mentioned above. These two scales for are made up of 16 items, for example:

Copyrighted information

Although respondents are expected to present themselves in a positive light especially in recruitment contexts, their scores are likely to be treated as dishonesty the higher the score on dissimulation.

3.4.2.2. Inattention and Contradiction

Inattention and contradiction checks are carried out before standardising the scores. Inattention assesses the tendency to complete the questionnaire haphazardly.

Copyrighted information

3.4.2.3. Acquiescence and within-subject standardisation

Acquiescence is another form of response bias manifested through the tendency to agree to most items or disagree to all (Rust & Golombok, 1999). Although acquiescence will not totally disappear, its influence can be reduced by introducing

items written both positively and negatively. However, Orpheus also makes use of within-subject standardisation also known as ipsative rescaling as a technique for controlling for acquiescence (Cheung & Rosenvold, 2000). This is computed for each individual separately by deducting the score on every item from the mean of all responses then dividing it by this candidate's standard deviation as follow

$$z = \frac{x - \bar{x}}{sd}$$

whereby $\bar{x}$ is the participant's average on all items, x is the participant's score on a specific item and *sd* is the participant's standard deviation based on his or her scores. Here is an example that best illustrates how this technique works (table 3.3).

	Q1	Q2	Q3	Q4	Q5
Person A	1	2	3	4	5
Person B	5	4	3	2	1
Person C	1	2	1	1	2
Person D	5	5	5	4	4

Table 3.3: Example data on a 1 to 5 Likert scale from a personality questionnaire.

If this is our data from a personality questionnaire, persons C and D are acquiescing in opposite directions whereas persons A and B are using the "whole" scale. So if we standardize the scores based on the Mean and SD of the whole sample, which would be 3 and 1 respectively for a normally distributed sample on a 1 to 5 Likert scale, we will notice that person C has all his scores below the mean and person D all above the mean. This is of course because we are comparing them to the overall mean 3. However, if we take the mean and SD of person C, then mean=1.4 and SD=1.1 therefore, the z score of this person will be between -0.07 and +0.5 so now this scale is not all positive or all negative. The same applies for Person D, his mean will be 4.6 and SD=1.1 so the his or her scores will range between 0.36 and -0.54

Therefore, the within subject standardisation adjusts the means of all the participants so that all of them would have answers above and below the mean. However, the real mean for some is 3 whereas for others its 4.5 Likert scale. Yet, all participants are now using the whole or at least more of the scale. Further discussion about response bias cross-culturally will follow in chapter 4 as well as further discussion about within subject standardisation.

3.5. Scale inter-correlation

Orpheus scales are considered to be independent with a low inter-correlation ranging between –0.06 to0.29 (Rust, 1996). This is necessary because the big five model assumes measuring five independent scales (Rust & Golombok, 1999). Rust and Golombok (1999) argue that most big five questionnaires assume that this assumption is satisfied by producing a five factor structure. However, the inter-correlation between those five factors is rarely reported (Rust & Golombok, 1999). The advantage that low inter-correlated scales provide is the breadth of the personality profiles they can produce. For example, a questionnaire comprising of two highly positively correlated scales A and B, will produce only two possible personality profiles: low A low B and high A high B. It will therefore not be possible to produce a profile for someone high on A and low on B. However, if A and B are not correlated, then there are four possible personality profiles: low A low B, low A high B, high A high B, high A low B. This characteristic allows for a greater distinction between individuals being assessed.

3.6. Norms

Orpheus was standardised on a sample of 427 respondents, 275 females and 138 males, from 20 companies and a variety of job roles including: teachers, security

staff, accountants, managers, drivers, engineers, scientists, sales, marketing, police officers, insurance claims negotiators, secretaries, clerks, HR personnel, insurance underwriters and so on. The age of the sample ranges from 16 to 62, with an average of 30.67 and a standard deviation of 11.01. The educational level ranged from no qualifications to Thesis and percentage of ethnic backgrounds in the sample resembled that of the UK working population.

3.7. Reliability

Regardless of the tool being used or what is being measured, measurement is subject to error that causes unwanted variation in scores (Cronbach, 1990). Error can be random, which is very undesirable, or systematic, which is not dangerous because it is "organised" or "constant" (Kline, 1993). For example, if a scale adds one kilo, it will do so to all measures systematically and therefore will affect all measurement equally. However, random or unsystematic error is significant because it can affect the accuracy of measurement (Kline, 1993). For example, if the scale randomly adds or deducts kilos from every measurement, the comparison between the different measures becomes inaccurate.

According to the theory of True Score, any observed score (X) that the measurement reveals is made up of two components: the true score of the person (T) and error (E) as shown in the following formula (Rust & Golombok, 1999):

$X=T+E$

Error could result from many sources such as the environment, participants not feeling good, guessing, badly written instructions and so on (Kline, 1993). Reliability of a test refers to how accurate it is in measuring what it is purported to measure (Rust & Golombok, 1999). That is, reliability refers to the error associated with the test. There are several types of reliability that can be computed, but the main concern of

reliability studies is test retest reliability, that is, whether people's scores would agree if they were tested twice (Cronbach, 1990). Internal consistency is another widely reported form of reliability but it falls short in that is neglects the sources of the variance (Cronbach, 1990). Internal consistency examines the relationship between items, which should be strong if the items are measuring the same variable (Nunally, 1978). Although some argue that high internal consistency is a pre-requisite of validity, this could actually result from asking relatively the same questions in the test (Kline, 1993). Parallel form reliability is another form that is usually employed when two versions of the same test exist (Rust & Golombok, 1999). The same participants take the two versions and a correlation of their scores estimates the reliability coefficient (Cronbach, 1990). Although having parallel form can be useful for decreasing the exposure of participants to the test especially if it is widely used, it is not very commonly applied. Split-half reliability is a more widely used form of reliability and consists of randomly splitting the test into two halves to create a "pseudo-parallel form" then correlating the results (Rust & Golombok, 1999).

3.7.1 Reliability in Orpheus

The only form of reliability that is reported in Orpheus is split-half. Split half reliability for the five major scales range from 0.73 to 0.81 and a standard error of measurement from 0.87 to 1.04. Although these reflect a good split-half reliability, test-retest would have been another good form of reliability to assess to provide a better estimate of error associated with Orpheus.

3.8. Validity

Test scores are said to be valid for a particular use when there is evidence to

support that use (Kline, 1993; Rust & Golombok, 1999). Validity is mainly concerned with the soundness of the inferences that are made from tests (Cronbach, 1990). However, there are different approaches to measuring this and these represent the different forms of validity (Anastasi, 1988).

Face validity the most basic form of validity that measures whether the test *appears* to be measuring what it is claiming to measure (Kline, 1993). This can be established by giving the test to potential test takers and asking them to rate what they think the test is measuring. It is important for a test to be face valid because test takers' attitude and faith in the test can have an effect on their responding.

Concurrent validity is usually measured against a certain criterion. It is established by correlating the test with an existing one that measures the same constructs (Rust & Golombok, 1999). However, there are several problems associated with this method. Some tests are very widely used but so not necessarily measure the construct of interest as well as other less famous tests. Yet, it is unlikely that test developers will choose the less popular test to validate their own (Kline, 1993). Additionally, Kline (1993) argues that if a good measure exists and can be used as a benchmark, what is the need for developing a new test? Rust and Golombok (1999) argue that this type of validity should never be used on its own. However, they explain that it could be very useful especially in identifying when the old and new test do not correlate.

Another criterion related type of validity is predictive validity, which measures how well a test can predict a certain criterion (Rust & Golombok, 1999). The main challenge to this type of validity is accurately identifying and measuring the criterion of interest (Kline, 1993). For example, occupational psychologists could be interested in how well a test predicts future job performance. One way of measuring it is by giving the test to a group of individuals a part of the recruitment process then

assessing their performance few months after they have been selected. Performance appraisals could be used to measure this criterion; however, it is important to ensure that they are validly assessing work performance.

Content validity on the other hand, applies to a small sample of tests that have a very clearly defined domain (Kline, 1993). A test is said to be content valid if its content fully represents the construct that is being measured (Haynes, Richard & Kubany, 1995). For example, say we are interested in developing a questionnaire to measure pathological gambling as defined by the DSM-IV-R (APA, 2000). This test would be content valid if the items tap on the ten diagnostic criteria listed in the DSM-IV-R. This type of validity is usually established by giving the questionnaire to experts in the field and asking them to rate its content (Kline, 1993).

Finally, construct validity is the last form of validity that encompasses all what have been previously discussed (Kline, 1993). Construct validity is concerned in whether a test is measuring what it is claiming to measure and all other types of validity try to tackle this issue from different angles. Some common methods of construct validity are replicating the factor structure and cross validating with another test or with other criteria that assess the same construct (criterion related validity). As for the relationship between validity and reliability, reliability could be seen as a pre-requisite for validity but does not necessarily guarantee it. However, if a test is valid it is definitely reliable (Nunnally, 1978).

3.8.1 Validity in Orpheus

Content validity of Orpheus was demonstrated by presenting the items that have the extreme positive and negative loading on each scale and these are as follow:

Copyrighted information

Although the items seem to represent well the content of the scales, these were not cross-checked with experts in the field as it is usually suggested (Kline, 1993). Construct validity of Orpheus was also established based on criterion validation. A total of 10 rating scales, one positive and one negative inspired from the items with extreme loading on each of the five major scales were chosen for supervisor rating. All participants from the standardisation sample were rated by their supervisors on a five point likert scale ranging from below average, to average, a little above average, much above average to exceptional. Table 3.4 below shows the results of this validation study.

Supervisor rating scale	Correlation with Orpheus scale
	Fellowship
Team skills	0.19**
Ability to work independently	-0.14*
	Authority
Ability to make friends with colleagues	-0.13§
Ability to make tough decisions	0.24**
	Conformity
Ability to generate new ideas	-0.25**
Obedience to company policy	0.23**
	Emotion
Level of self confidence	-0.25**
Tendency to worry	0.08
	Detail
Attention to detail	0.23**
Breadth of vision	-0.22**

Table 3.4: Correlation Between supervisor ratings and Orpheus major scales

* $p<0.05$ (two tailed)
** $p<0.01$
§ $p<0.05$ (one tailed)

In summary, Table 3.5 below summarises the technical qualities of Orpheus and compares them to a selection of widely used tests in occupational contexts in the UK: NEO, OPQ32, and HPI

Name of test	Orpheus	NEO	OPQ	HPI
Scales	Five domains (Organisational; Social; Emotional; Intellectual; Perceptual) represented in five scales	Big five (Extroversion, Agreeableness, Neuroticism, Openness to experience, conscientiousness)	Three domains (Relationship with People; Thinking Styles; Feeling and Emotion) represented in 32 scales	Seven scales that map on to the big five model
Unique features	▪Work-based ▪4 audit scales ▪Low inter scale correlation	▪ Most extensively studies test ▪ Adapted into more than 30 languages and cultures ▪ Made the Big Five famous ▪ Applied in clinical, counselling, occupational, health, behavioural medicine, and educational settings	▪ Work-based ▪ Available in more than 30 countries ▪ 86 regional norms ▪ Available in ipsative and normative versions	▪ Specifically tested to look at predictive validity in relation to several occupational groups
Validity	▪ Criterion related ▪ Content	▪ Convergent ▪ Discriminant ▪ Construct	▪ Content validity in relation to Big Five ▪ Concurrent with 16PF, OPP, HPI, MMPI, MBTI, NEO, IPIP ▪ Criterion related (Intellect, leadership, entrepreneurial, interpersonal, creative).	▪ Concurrent with MMPI, MBTI, 16PF, MVPI, and more in the US. ▪ More than 400 predictive validity studies between the US and UK
Reliability	▪ Split-half	▪ Internal consistency ▪ Test retest	▪ Internal consistency ▪ Test-retest	
Norms available for practitioners	▪ 427 across the UK ▪ International norms investigated in this Thesis	▪1301 across the UK ▪ 14 international norms including the English UK version	▪ 2028 across the UK ▪ 86 regional norms with median 979.5	

Table 3.5: Orpheus and other work-based personality tests

3.9. Conclusion

At the time Orpheus was developed, it was an innovative test as it reconceptualised the Big Five in the workplace. The language employed in questionnaires used in clinical settings might be inappropriate for use in the workplace, such as "neurotic". The contextualisation of Orpheus in the workplace overcame this problem and made it unique in that respect. Additionally, the audit scales also add value to Orpheus as they measure several possible ways of response distortion. In practice, this is particularly insightful in development settings whereby negligence or resistance on behalf of the test taker can be detected and explored by the practitioner. One of the potential problems with adapting Orpheus, however, is its reliance on English idioms, manifested in items such as "I sometimes get hot under the collar when people act in an unhelpful manner". This can be a problem when the English version of Orpheus is used with a multilingual workforce or when it is adapted into another language. Furthermore, Orpheus lacks cross-validation studies especially when compared with other prominently used tests in the UK.

Chapter 4: Levels of Equivalence and Bias

"A book is a version of the world. If you do not like it, ignore it; or offer your own version in return."
Salman Rushdie

4.1. Chapter Overview

This chapter offers a critical discussion of the concepts of bias and equivalence with particular reference to cross-cultural test adaptation from a psychometric perspective. Whilst van de Vijver and Leung's (1997) theory of bias and equivalence has informed much of our thinking, we provide this chapter as 1) an overview of the theory and its development throughout the last decade and 2) a critical evaluation of the terminology employed in the literature of test adaptation and cross-cultural assessment. The overview incorporates a description of the different forms of bias (construct, method, and item bias) that threaten validity of cross-cultural comparisons and their effect on the different types of equivalence between tests (construct inequivalence, construct equivalence, measurement unit equivalence and scalar equivalence). For each form of bias described we outline how it manifests itself and the practical steps to address and diminish or negate its effects. While doing so, we will distinguish between some closely related concepts such as: equivalence and inequivalence, bias and equivalence, bias in psychometrics and bias in cross cultural research, uniform and nonuniform bias, types and sources of bias, validity in psychometrics and validity in cross cultural research, etic and emic approaches, and divergent and decentred approaches.

We will also address in this chapter the confusion in the terminology frequently referred to in the statistical and theoretical literature on cross-cultural test adaptation in general. To illustrate, terms such as "equivalence" and "invariance" are sometimes used interchangeably and refer to the same concept. However, the former is more theoretical

whereas the latter is a statistical term. As another example, the terms "conceptual" and "cultural" equivalence sound like different concepts but actually refer to the same one. Finally, while unifying terminology and defining it; we will suggest a restructuring for van de Vijver and Leung's theory of equivalence and bias into a theoretical framework.

4.2. 2. Introduction

The ITC guidelines (1994, 2000, and 2001) and van de Vijver and Leung's (1997) Theory of Equivalence and Bias laid the foundations for unifying practice in cross-cultural assessment. However, it is argued that the ITC guidelines suffer from being difficult to implement in practice (Hambleton & Li, 2004) even though several articles are being published to provide examples of their practical implementation (such as van de Vijver & Hambleton, 1996; van de Vivjer & Tanzer, 1997; Hambleton, 2001).

van de Vijver and Leung's (1997) theory, on the other hand, provides a comprehensive summary of the issues related to test adaptation and groups them under a Theory of Equivalence and Bias. Although based on challenges encountered in test adaptation, this theory is somehow explanatory without being fully descriptive first. That is, some concepts are very clearly explained in terms of *why* they create bias but not as clearly in terms of *what* are all the sources that lead them to do so. For example, the theory provides a comprehensive explication of why mistranslation creates item bias, but there are many other sources (such as psychological bias) that can lead to item bias and these are scattered in the literature, as will be discussion below. Method bias, on the other hand, is explanatorily and descriptively well defined and offers an answer to both questions: "what are the sources of method bias?" and "why do these sources lead to

method bias?"

We argue that the van de Vijver and Leung (1997) theory is a comprehensive one but could be restructured to become more descriptive as well as being explanatory. This can be achieved amalgamating the main sources of bias that have been discussed in the literature and adding them to the theory. This will consequently render this theory more easily applicable in practice. In attempting to do so, we will first start by defining the terms equivalence and bias in different context, then move on to the Theory of Equivalence and Bias, and conclude with a restructured framework of equivalence and bias.

4.3. Defining equivalence and bias

Equivalence and bias are two closely but inversely related concepts fundamental to cross-cultural comparisons (van de Vijver & Leung, 1997). For scores to be comparable and equivalent between two groups of interest, they need to be unbiased. Paradoxically, equivalence is always investigated first in practice followed by an assessment of bias; perhaps due to the fact that tests that are equivalent should be free from bias.

The terms bias and equivalence have slightly different meaning when used in classical test theory, where comparisons are mainly within one culture, than in the field of cross-cultural testing, where comparisons are between cultures.

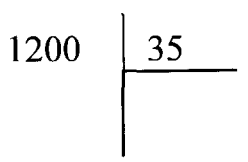
4.3.1 Bias in classical test theory

Freedom from bias is one of the four psychometric principles in classical test theory considered as the hallmarks of any good measurement tool. Bias is very closely associated with fairness whereby a test is considered to be biased if it is "unfair to a group of individuals who can be defined in some way" (Rust & Golombok, 1999, p83). In other words, a biased test is one that unfairly discriminates against a specific group of test takers.

Bias can be subdivided into three categories: item bias, intrinsic bias, and extrinsic bias. An item is biased when one group of individuals is more likely to answer it correctly (or endorse it in the case of personality tests) than an equally knowledgeable group (Zumbo, 2006). Typically, these groups refer to age, gender, ethnic minority or any other group within one culture. For example, in Lebanon in general both public and private schools follow either American or French schooling systems. So let us consider a Lebanese national maths exam with the following item:

Calculate: $35\overline{)1200}$

This is a long division symbol as referred to in the American schooling system and requires students to divide 1200 by 35. This item could be considered biased to students who studied in the French systems since the same long division in that schooling system is written as follow:



Lebanese students from French schools who are as "knowledgeable" as their counterparts from American schools (i.e. of the same mathematical ability) will be less likely to get this item right. However, this is not due to their ability to calculate this long division, but to their familiarity with that specific long division symbol.

Intrinsic bias is similar to item bias, but the bias occurs on the overall score of the test rather than on one item in specific. Therefore two equally knowledgeable groups of respondents are likely to get significantly different mean scores on the test but on the basis of characteristics that relate to the test rather than the construct being measured. That is, it is not their mathematical ability that determines how well they do on the test, but some characteristics that related to the test itself. For example, a test that comprises of several items such as the long division example provided earlier would result is lower scores for the group of students who are not familiar with the symbols used but not because of their mathematical ability.

Extrinsic test bias on the other hand, also known as adverse impact, is characterised by differences in responding between two groups but due to characteristics unrelated to the test (Rust & Golombok, 1999). The differences between the groups in this case are real, but unfair mainly due reasons such as social deprivation. As an example, Rust and Golombok (1999) explain that some immigrant groups end up living in deprived areas within cities where the schools are of poor quality, which affects their overall academic achievement. Their scores could therefore be lower than those of other relevant groups (such as nationals) but due to lack of opportunities. Although this indicates that group differences are real, the source of this discrepancy is unfairness in the world and can lead to a cycle of deprivation if not recognised and dealt with. Bias is

therefore an anomaly, external or internal to the test, which affects the validity of the inferences it produces or the comparison between people on the same test.

4.3.2 Bias in cross-cultural research

In cross-cultural research, bias can be detected on two levels. The first level is what was discussed in the earlier section, where groups of interest are within one culture (such as age, gender, ethnic group etc.), whereas the second level is between countries whereby the groups of interest are between cultures. When scrutinising one version of a test, freedom from bias needs to be established before making comparisons between individuals. When several versions of the tests are to be scrutinised, they first need to be free from bias (item, intrinsic and extrinsic) independently in order to allow within culture comparisons. However, they also need to be collectively unbiased against any culture group, in order to allow between group comparisons. On that level, there are three forms of bias that can be distinguished: construct, method, and item bias (van de Vijver & Leung, 1997; van de Vijver 1998; van de Vijver & Poortinga, 2005). In the field of cross-cultural testing, bias does not only affect the validity of scores in one test, but to the validity of score comparison between the cultural groups of interest.

4.3.3 Equivalence in classical test theory: validity and reliability

Equivalence, as defined by Cambridge advanced learner's dictionary (2000), refers to "having the same amount, value, purpose, qualities, etc." The concept of equivalence is closely associated to certain forms of reliability and validity (Arffman,

2007).

Reliability refers to the extent to which any form of assessment is free from error (Rust, 1996) mainly in terms of accuracy and repeatability of scores. Several forms of reliability, such as inter-rater reliability, test retest, parallel forms, and split half (Kline, 1993; Rust & Golombok, 1999), rely on the concept of equivalence. For example, test retest reliability refers to the comparability of scores of a group of participants who take a test on two different occasions (Kline, 1993). Therefore test retest reliability considers the *equivalence* between the scores across time. However, parallel forms reliability relies most on equivalence, since it assumes that for two versions of a test are parallel, they need to be comparable in difficulty and discrimination.

Validity, on the other hand, refers to "the extent to which a test is measuring what it is purported to measure" (Rust & Golombok, 1999, p 64). This definition explains the most important form of validity: construct validity. A test is construct valid if it measures the underlying construct it is designed to measure. In practice, validity can be established by running several studies such as concurrent, convergent, divergent, and predictive validity (Kline, 1993). Concurrent validity for example is assessed by correlating the newly developed test with another test that measures the same construct. Say we are interested in assessing the concurrent validity of a newly developed test that purports to measure the Big Five Model; we could correlate its results with those of NEO-PI-R®, which is famously known for assessing the same constructs. Arguably, high correlation between the two indicates that they are measuring the same construct. In other words, concurrent validity measures how equivalent two tests are in measuring the same construct or constructs.

4.3.4 Equivalence in cross-cultural research

In cross-cultural research, equivalence is still very much associated with reliability and validity but refers to whether there is any difference in measurement level of within and between group comparisons (van de Vijver, 1998). That is, two multilingual versions of a test are equivalent if 1) test takers within one culture can be meaningfully compared and 2) test takers between two cultures can be compared. Conversely, the former can be established without the latter, but not vice versa as shown in table 4.1 below. So it might be possible to have two tests that function well in each country, but the results of test takers in one culture cannot be compared to the results of test takers from the second culture. However, if this level of comparison is possible, then we can certainly assume that the test functions well in each culture separately. Equivalence between multilingual versions is analogous to the concept parallel forms reliability in classical test theory. Two forms of the same test (usually in the same language) are considered parallel if they are equivalent, that is, if they are equally reliable (Rust & Golombok, 1999). The same applies cross-culturally between the different language versions of the same test. For two linguistic versions of the same test to be parallel, they need to function similarly in the two cultures. However, there are several levels of equivalence that can to be established in order to determine how comparable tests are and how reasonable it is to compare participants within or between cultures. These are: construct, measurement unit, and scalar equivalence, which will be discussed in detail in the subsequent sections.

4.3.5 Equivalence and Invariance

The terms "equivalence" and "invariance", and their opposites "inequivalence" or "nonequivalence" and "noninvariance", are sometimes used interchangeably in the literature on cross-cultural score comparability to refer to the same idea. However, it is important to distinguish conceptually between measurement invariance (Meredith, 1993) and the types of equivalence (scalar, measurement unit and construct). The former is a statistical term that refers to statistical evidence of equality in factors (configural invariance), factor loading (weak invariance), intercept (strong invariance), and residual variance (strict invariance) (Wu, Li, & Zumbo, 2007). 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 types of equivalence will be discussed in more detail later in this chapter whereas measurement invariance will be extensively explained in Chapter 8.

Invariance is defined as "not varying" (The American Heritage, 2007), and is composed of the prefix "in" and the word "variance". The prefix "in" is a Latin negative prefix which makes "invariance" a negative word (Ferguson, 1997). Thus, the opposite of invariance, noninvariance is a double negative word because it is composed of two consecutive negative prefixes "non" and "in". Although double negative *sentences* are common across languages, double negative *words* are particular to the English language (and perhaps few others) and do not necessarily exist in all other languages. This makes these words less straightforward and harder to understand especially in the field of cross-cultural comparisons, which attracts researchers from all over the world. Since equivalence and invariance refer to slightly different concepts, it is not possible to use of

term to refer to both. For this reason, we argue that the term *invariance* should be used as a concept that could be either achieved or not. That is, instead of referring to *measurement non-invariance*, it is possible to replace it with *measurement invariance not achieved.*

4.3.6 Relationship between equivalence and bias

Equivalence refers to the comparability of scores between the different versions of a test, whereas bias represents the issues that threaten the comparability between the adapted versions (van de Vijver, 1998). There are several types of bias that can affect the different forms of equivalence. The existence of bias in cross-cultural test adaptation is undesirable as it might lead to inequivalence between the multilingual versions on at least one of the following levels of equivalence hierarchically: construct, measurement unit, or scalar. Tests need to satisfy the different types of equivalence in that order since certain types of equivalence are prerequisites for later types. For example, it is possible to have tests that are equivalent on a construct level, but not on the measurement unit and scalar levels. However, it will not be possible to have measurement unit equivalence without satisfying construct equivalence. Table 4.1 illustrates the levels of equivalence and comparisons, which will be explained in section 4.4.

Construct equivalence	Measurement unit equivalence	Scalar equivalence	Possible comparison
Yes	No	No	Within cultures only
Yes	Yes	No	Within culture and indirectly between cultures
Yes	Yes	Yes	Within cultures and between cultures directly

Table 4.1: *Levels of equivalence and comparisons.*

4.4. Types of equivalence

The ultimate goal of any test adaptation process is to reach equivalence between multilingual versions of tests. There is a degree to which two versions could be equivalent, accordingly, different types of comparisons can be allowed. van de Vijver (1998) differentiated between four types of equivalence necessary for full test equivalence and these are: construct inequivalence, construct equivalence, measurement unit equivalence, and scalar equivalence. However, we will distinguish between three types of equivalence only by collapsing construct equivalence and inequivalence together under "construct equivalence" as explained in the following section.

4.4.1 Type 1: Construct Equivalence

4.4.1.1. Definition of construct inequivalence

Construct inequivalence refers to situations where psychometric tools are measuring different constructs in different cultures (van de Vijver & Poortinga, 1998; van de Vijver & Hambleton, 1996). This is due to the constructs being dissimilar in the given cultures (van de Vijver, 1998; van de Vijver & Hambleton, 1996; van de Vijver &

Poortinga, 2005). Psychological constructs are defined through behaviours, values, attitudes or norms that are sometimes different across different cultural groups (Rust & Golombok, 1999). van de Vijver and Poortinga (2005) portray the concept of inequivalence by comparing the definitions of intelligence in Kenya, Zambia, Japan, Europe and America discussed in the following studies by Munday-Caste (1974), Sperpell (1993) Azuma Kashiwagi (1987). In Europe and the US, intelligence is mainly associated with academic achievement. Whereas in Kenya, Zambia and Japan, the definition of intelligence encompasses the academic achievement to cover culturally defined behavioural aspects such as respecting others and using appropriate language in conversations (for further discussion about intelligence across cultures, see van de Vijver & Tanzer, 1997; Sternberg, Nokes, Geissler, Prince, Okatcha, Bundy & Grigorenko, 2001). Therefore the construct of interest in this case is not equivalent and cannot be measured using the same tool across languages and cultures.

4.4.1.2. Definition of construct equivalence

On the other hand, construct equivalence, also referred to as functional or structural equivalence refers to situations where the tools are measuring the same construct across the different languages. This type of equivalence is established by replicating the patterns of correlations between the different comparison groups (van de Vijver, 1998). When different language versions of the same test result in similar factorial structures in the different cultures, it is assumed that they are measuring the same constructs across cultures (van de Vijver & Poortinga, 2005; van de Vijver, 1998). However, factorial or construct equivalence does not guarantee full equivalence between

the tests (Byrne & Watkins, 2003).

4.4.1.3. Relationship between construct equivalence and inequivalence

The distinction between construct equivalence and construct inequivalence seems confusing, as there is a great deal of overlap in their definitions. This confusion is partly due to the similarity in the terms *construct equivalence* and *construct inequivalence*. When van de Vijver and Poortinga (2005) discussed van de Vijver and Leung's (1997) Theory of Equivalence and Bias in their chapter, they replaced the term *construct equivalence* with *structural or functional equivalence*. Perhaps they did so to overcome this confusion and highlight the distinction between the two concepts.

Whilst replacing *construct equivalence* with *structural equivalence* might seem to solve the problem, we argue that there is further complexity in the definition of these types of equivalence. To clarify, *construct inequivalence* refers to situations were the tests are not measuring different things across two or more cultures but mainly because "constructs are associated with different behaviours or characteristics across cultural groups" (van de Vijver & Tanzer, 2004, p10). Construct equivalence reflects situations where tests are measuring the same issues between cultures. These two concepts seem like one, *construct equivalence*, which have or have not been achieved regardless of the reason for not reaching equivalence. Some problems with test construction, which will be discussed in section 4.5.1 under *construct bias*, may in fact lead to *construct inequivalence*. To summarise, if the test is free from construct bias, then it is possible to achieve *construct equivalence* between two or more cultures. However, if it is affected by construct bias, then it will fail to do so and therefore become *construct inequivalent*.

Similarly, *measurement unit equivalence* and *scalar equivalence*, which will be examined in the next section, could also be reached or not depending on freedom from certain types of bias. Yet, van de Vijver and Leung (1997) did not add measurement unit inequivalence and scalar inequivalence as additional types of equivalence. Rather, they portrayed these two types of equivalence on a continuum whereby equivalence is on one end and inequivalence on the other.

We therefore argue that *construct equivalence* and *construct inequivalence* should be merged into one type, *construct equivalence*. Should construct equivalence be a continuum, two version of the same test could, on one end, achieve construct equivalence or, on the other end, be construct inequivalent. As a further elaboration, failing to achieve construct equivalence can arise from construct bias, which in turn has several potential sources: either the constructs are fundamentally different in the different cultures or there is a malfunction in the test. To conclude, this concept of equivalence is a dimension of similarity that multilingual versions of tests could either reach fully or not at all and we will, hereafter, refer to it to as *construct equivalence*.

4.4.2 Type 2: Measurement unit equivalence

4.4.2.1. Defining Measurement equivalence

Measurement unit equivalence refers to situations where participants from different cultures perceive and interpret observed measures (items) similarly (Byrne & Watkins, 2003; Muller, 1995). The assumption is that the origin of one of the scale is shifted, in other words, participants in one group score consistently higher or lower than participants in the other group. To illustrate, van de Vijver and Tanzer (2004) explain that

two scales could be measuring the same construct, say temperature, but using different scales, say Kelvin and Celsius. Although temperature (as the construct of interest) is being measured with either tool, they cannot be directly compared. The reason is that 30°C is not equal to 30°K, but rather to 303°K. However, since Celsius and Kelvin have a constant difference of 273° between them, it is possible to convert one to make it comparable to the other. When two instruments have measurement unit equivalence, it is possible to compare them directly only if the offset or constant difference between them is identified, which in practice is hardly ever the case (van de Vijver, 1998). Therefore, when measurement unit equivalence is achieved it is possible to compare differences measured in each group (van de Vijver & Leung, 1997; van de Vijver & Tanzer, 2004; van de Vijver & Poortinga, 2005). For example, if a Croatian and a Korean version of a depression questionnaire have measurement unit equivalence, differences in depression between genders in Korea can be compared to difference in depression between genders in Croatia. Conversely, no direct comparison can be done between a participant in one culture and another participant from the second culture. As a final point, when measurement unit equivalence have not been reached between groups, the tests might still be measuring the same construct in each culture independently but no comparison whatsoever can be made between the cultures.

4.4.3 Type 3: Scalar equivalence

4.4.3.1. Defining scalar equivalence

Scalar equivalence could be seen as the ultimate goal to be reached for assuming full score comparability between different language versions (van de Vijver, 1998; van de

Vijver & Leung, 1997; van de Vijver & Tanzer, 2004). Scalar equivalence is full equivalence between two measures, indicating that they are functioning in the same manner across any cultures of interest. If scalar equivalence is achieved, tests are assumed to be bias free.

4.4.4 Measuring construct, measurement unit and scalar equivalence

Principal component analysis is sometimes applied to assess construct validity (Gierl, 2000); however, exploratory factor analysis *with target rotation*[2] is also a common method for construct equivalence (van de Vijver & Leung, 1997; Osterlind, Miao, Sheng, & Chia, 2004). More recently analysis of covariance structures, such as confirmatory factor analysis [CFA] within structural equation modelling, is becoming more common for assessing construct equivalence (Byrne & Watkins, 2003; van deVijver, 1998; Krishnakumar, Buehler, & Barber, 2004, Meckler & Mullen, 1997; Muller, 1995). There are several levels of measurement invariance (Meredith, 1993) that can be investigated in CFA each of which has implications of assuming different levels of equivalence.

The levels of measurement invariance (the statistical counterpart of the concepts of equivalence discussed above) are:

1. configural invariance, which assumes equality in factors,
2. weak invariance, which assumes the configural in addition to equality in loading
3. strong invariance, which assumes weak in addition to equality in intercept and finally
4. strict, which assumes strong invariance in addition to equality in residual variance (Wu, Li, & Zumbo, 2007).

---

[2] Target rotations are essential for cross cultural research but not available in popular programmes such as SPSS. However, van de Vijver and Leung (1997) provide a procedure for applying it using SPSS.

Although this will be explored in more detail in chapter 9, it is worth mentioning that similarity in factor structure and factor loading between groups of interest indicates equivalence of construct between them. This means that they are all measuring the same construct within each culture. However, evidence of similarity in factorial structure within each group does not guarantee equivalence across the groups (Byrne & Watkins, 2003). In other words, the tests could be measuring the same construct in each culture (construct equivalence) but not necessarily be directly comparable (measurement unit or scalar equivalence not achieved).

4.4.5 Relationship between the three types of equivalence

Equivalence is hierarchical by nature whereby lower equivalence levels need to be achieved first before assuming equivalence on higher levels. Construct equivalence is the lowest level of equivalence and is a prerequisite for the next level of equivalence, measurement unit. When measurement unit equivalence is achieved, it can be taken for granted that construct equivalence has also been reached. Similarly, when scalar equivalence is established, both measurement unit and construct equivalence are considered to have been fulfilled as well. The less bias between the tests, the more likely they are to achieve higher levels of equivalence.

4.5. Types of Bias

Hambleton (2005) suggested that poor test translation is the main challenge to the validity of adapted versions of psychometric tools. In fact, mistranslation is one source of bias that affects one specific type of bias, that is, item bias. However, there are several

types of bias identified in the Theory of Equivalence and Bias that challenge the validity of tests and consequently the equivalence between them, these are: a) construct, b) method, and c) item bias (Hambleton & van de Vijver, 1996; van de Vijver, 1998; van de Vijver & Leung, 1997; van de Vijver & Poortinga, 2005). The three types of bias could arise from different sources. It is crucial at this point to differentiate between *type of bias* (also referred to as *form of bias*) and *source of bias* since the two concepts are closely related yet different. As mentioned earlier, different types of bias may affect different types of equivalence. However, each of these biases can arise form one or more sources. For example, linguistic mistranslation could be a potential source of bias that results in item bias. The existence of item bias challenges equivalence between multi lingual versions of tests. So in this case:

- Source of bias: linguistic mistranslation
- Type of bias: item bias

In the following sections, we define each type of bias and illustrate potential sources with practical examples, followed by ways of detecting and controlling it.

4.5.1 Construct bias

Construct bias is the first and most general form of bias, which affects the construct equivalence between tests. In short, construct bias occurs when constructs being measured are not equivalent between given cultures (Byrne & Watkins, 2003). Construct bias can arise from two main sources, differential construct manifestation and construct under-representation, which could be dealt with using two approaches, convergence and decentred approach. In the following sections, we will discuss the conceptual differences

between these two sources as well as the approaches for managing them using practical examples to illustrate.

4.5.1.1. Sources of construct bias

Differential construct manifestation (DCM)

This first source of construct bias could result from the fact that the construct, although it exists in both cultures, it is defined and exhibited differently in each culture (van de Vijver, 1998; Hambleton & van de Vijver, 1996; van de Vijver & Tanzer, 1997; van de Vijver & Leung, 1997; Byrne & Watkins, 2003). This type have not been given a label in the Theory of Equivalence and Bias but we will refer to it as *differential construct manifestation bias* since the same construct manifests itself differently in each of the cultures of interest. This source of bias is mostly associated with the cultural differences where the constructs do not fully overlap.

To illustrate, the concept of *differential construct manifestation*, we will consider the example of depression. Differences in the symptoms of depression between the Eastern and Western cultures are a good example of this concept. Although for a number of years depression was thought to exist only in advanced industrialised societies (DSM-IV-R-TR, 2000), extensive research in this field revealed that this psychological disorder does exist in pre-industrialised societies as well (Sulaiman, Bhugra, & De Silva, 2001). However, it emerged that its existence in the latter societies was masked by the use of western diagnostic tools and procedures. Depression as described in the DSM-IV-TR is usually experienced through a feeling of guilt and sadness. Nevertheless, in some cultures, it is experienced and expressed through physiological symptoms such as

headaches and nerves in Latino and Middle Eastern cultures, tiredness and imbalance in Asian cultures, and heart problems in the Middle East (DSM-IV-R-TR, 2000).

A study by Sulaiman, Bhugra and DeSilva (2001) suggests that Dubai nationals somatise the symptoms of depression much more than their western counterparts. The authors explain that physical symptoms are more culturally acceptable in Arab culture than psychological ones, which affects the definition and manifestation of this psychological disorder in that part of the world. It is therefore common in Dubai to consider complaints such as: "I can't breathe" or "I have a headache, a stomach-ache or backache", in the diagnosis of depression. The behavioural symptoms of depression and the way they are interpreted by individuals are qualitatively different between Eastern and Western cultures (for further reading about differences in manifestation of depression, see Okasha, el Akabawi, Snyder, Wilson, Youssef & el Dawla, 1994; Cheng, 2001; and Kleisman 2004). The content of the clinical depression scale in the Arab world should therefore differ from the one in used in the West (Sulaiman, Bhugra, & De Silva, 2001).

In summary, although the construct exists in the different cultures, they do not overlap fully. In order to have construct equivalence, the questionnaires need to measure the same construct in both cultures. In this case, translating or adapting a test into another culture is not enough to secure construct equivalence. Therefore, depression cannot be measured using the same tools across the different cultures due to construct bias.

Construct under-representation (CUR)

Another source of construct bias is: *construct under-representation* (Messick, 1989, 1995; Emberston 1993 in Van de Vijver & Leung, 1997). This is characterised by

insufficient sampling of the behaviours that explain the construct. In other words, the construct is insufficiently represented in the content of the original questionnaire. This is parallel to the concept of *content validity* in classical test theory, which assumes that the test measuring a certain construct should be fully representative of this construct (Kline, 1993). *Construct under representation* is an anomaly related to the original instrument because it does not cover all the essential dimensions and facets that define the construct (Messick, 1995). Generally, for constructs to be under-represented, the original test is usually either too short to make valid deductions from it or the items are too badly written to tap on the construct it is claiming to measure (Downing, 2002).

As an example, let us consider emotional intelligence (EI) as the construct of interest. Although there is no real consensus about the specific constituents of emotional intelligence (perhaps due to the novelty of this construct and the lack of literature about it), several researchers distinguish between trait EI and ability EI (Petrides, Pita, & Kokkinaki, 2007). Several theories have been developed to explain either type of EI, most of which overlap. BarOn EQI (BarOn, 2002) and TEIQ (Petrides & Furnham, 2003) are two emotional intelligence questionnaires based on different emotional intelligence theories though some of their scales overlap. In BarOn EQI and TEIQ, the subscales "interpersonal" and "emotionality" respectively measure the ability of individuals to communicate their feelings and understand other people's feelings. "Interpersonal skills" is an essential part of emotional intelligence without which the construct of emotional intelligence will be underrepresented. Consequently, a test of EI that does not measure "interpersonal skills" is likely to be affected by construct bias when it is adapted because of the construct under-representation in the original test.

4.5.1.2. Comparing the two sources of construct bias

When contrasting these two sources of construct bias, it is evident that differential construct manifestation is an anomaly extrinsic to the test whereas construct under-representation is an intrinsic one. The communality between them, however, is that they both represent a threat to the validity of construct being measured by the test and its equivalence between cultures.

4.5.1.3. Dealing with construct bias

Construct bias is a source of anomalies that affects the most fundamental outcome of any test adaptation process: equivalence between the constructs measured. Two approaches for dealing with construct bias in comparative cross-cultural research are convergence and decentred approaches (van de Vijver & Leung, 1997). The former approach is "emic" (culture specific) by nature, whereas the latter is "etic" (culture general) (Den Hartog, House, Hanges, Ruiz-Quintanilla, & Dorfman, 1999). The terms emic and etic were originally coined by Pike (1967; in Den Hartog et. al, 1999), to distinguish between phonemics and phonetics, which are respectively sounds that are language specific versus sounds that are used in all languages. Emic and etic were later adopted by social scientists such as anthropologists and behavioural scientists (i.e. psychologists) to refer to concepts that are specific to a culture or a human being and those that are shared by all humans or cultures. In cross-cultural research, a project is said to adopt an emic approach when it focuses on attributes and behaviours that are specific

to one culture. Whereas an etic approach involves attributes and behaviours that are more culture general, and cross-cultural comparison is possible.

Dealing with DCM: Convergence approach

In cross cultural test adaptation practices, the convergence approach entails developing emic versions of the questionnaires to investigate the same constructs in the different cultures. Before attempting to adapt tests, researchers can start by reviewing the literature for the underlying theory behind the concepts measured in the test and its psychometric properties, and/or by collecting information about characteristics and behaviours associated with a construct in different cultures using survey method (van de Vijver & Hambleton, 1996). If this investigation confirms that the constructs suffer from differential construct manifestation, then the adapted version is subjected to fundamental changes in the conceptualisation of the underlying variables of interest. In other words, the adapted version is rendered culture specific and scores cannot be directly compared to the original version. Accordingly, the adapted version could include all the culture specific characteristics associated with the construct, even those not covered or those manifested differently in the original culture. As a result, two versions, free from construct bias can be developed. Yet score comparability cannot be established. To illustrate this better, comparing groups directly using measures developed by convergence implies that the tests are measuring the same construct in each culture. However these cannot be compared directly or indirectly, as it would be like comparing the verbal skills of one group to the numerical skills of another. This is a case of construct equivalence but measurement unit inequivalence. That is, the questionnaire as measuring the same constructs in both cultures, but using different scales that cannot be directly

compared.

Dealing with CUR: Decentred approach

On the other hand, the decentred approach involves the parallel development of the measurement tools across the different languages based on data collected from all the cultures of interest. The GLOBE cross-cultural leadership project (Global Leadership and Organizational Behaviour Effectiveness), for example, was launched in 1993 to assess the relationship between the societal culture, organisational culture and organisational leadership in 61 different countries (House, Javidan, Hanges, & Dorfman, 2002). The researchers were interested in uncovering the attributes that each culture values as essential for outstanding leadership. They hypothesized that characteristics of "charismatic/transformational" leaders will be endorsed universally (61 cultures) as fundamental characteristics for outstanding leadership. However, they were aware that the "charismatic/transformational leadership" might be coined with differing behaviours in the 61 cultures of interest (Den Hartog, et al., 1999) but they were still interested in measuring the same leadership characteristics using the same questionnaire in order to facilitate cross-cultural comparison. As a result they adopted the decentred approach to develop a measurement tool that assesses leadership characteristics. They used survey method with 15,022 middle managers from the 61 different cultures to collect information about the characteristics that enhance and impede leadership (Hollensen, 2001). After subjecting the data to first and second order factor analysis, the leadership attributes were clustered under six dimensions, three of which overlapped with Hofstede's cultural dimensions[3]. Therefore, the end result was a questionnaire that draws on all the cultures of interest and can be used across all of them.

[3] To read more about the GLOBE project see http://www.thunderbird.edu/wwwfiles/ms/globe/index.asp

Comparing the convergent and decentred approaches

Some concepts, such as depression discussed earlier, can be very culture specific and a convergent approach is more suitably applied in those cases. Culture specific questionnaire can be developed and the same construct can be measured in each culture while respecting the cultural difference and providing a fair and accurate assessment in each culture. On the other hand, a decentred approach is ideal for situations where a common questionnaire can be developed based on information from all cultures involved. Both approaches can secure construct equivalence; however, the former cannot lead to measurement unit equivalence whereas the latter could. That is, adopting the convergent approach leads to measuring the same construct in all cultures without being able to do a comparison between them. However, adopting the decentred approach leads to measuring the same construct in all cultures and also possibly to do a certain level of comparison between them. In summary, either of these methods is a useful techniques for protecting construct equivalence.

4.5.2 Method bias

Method bias takes its name from the methods section because it relates to topics usually covered in that section of any journal article or thesis (van de Vijver & Leung, 1997). Method bias could emerge from three main sources: the instrument itself, the data collection process, and the characteristics of the sample. However, there are several methods that could be employed to minimise the different sources of method bias. These will be discussed in detail in the following section.

4.5.2.1. Sources of method bias

Method bias is another type of bias that could present an obstacle for tests in reaching equivalence (van de Vijver & Poortinga, 1997; van de Vijver, 1998; van de Vijver & Hambleton, 1996; van de Vijver & Poortinga, 2005). Method bias is not related to the conceptual development of the questionnaire, but to the data collection process, tools, and participants. Therefore Method Bias encompasses three types of bias: Instrument, Administration, and Sample bias.

Instrument bias

Instrument bias refers to characteristics that can affect candidates' scores but that relates to the measurement tool rather than the characteristic being measured. An example of such characteristics could be familiarity with the response format or style of responding (van de Vijver & Poortinga, 2005). Social desirability responding and purpose of talking a test and participants' motivation will also be discussed as additional sources of instrument bias.

Source 1: Familiarity with response format

Psychometrics tests differ in the response format they employ (such as Likert scale, multiple-choice, or open-ended questions) which could constitute or result in a form of instrument bias. The reason for this is that certain cultures can be more familiar with one type of response format than other cultures. In some countries, for example, questionnaires are only filled out for governmental or legal purposes (Fife-Schaw, 2006) People in these countries are different from their European counterparts, who are more used to taking questionnaires for scientific research, and might respond to personality

tests in a similar way than they would respond to governmental surveys (Fife-Schaw, 2006). As another example, certain schooling systems rely more on multiple-choice exams since an early age which makes respondents from those schools more familiar with this response format than their students whose schooling system generally employ open-ended questions. This type of bias falls under the umbrella of instrument bias since the scores of respondents are affected by criteria associated with the instrument rather than the construct being measured.

Source 2: Response style

Response style refers to participants' style of responding, which could be affected by culture. Two main forms have been prominently discussed in the literature and these are Extreme Response Style (ERS) that affects measurement unit equivalence and Acquiescence Response Style (ARS) that affects scalar equivalence (Cheung & Rensvold, 2000). Although this form of bias relates to the participants' cultural background, response styles fall under instrument bias because the response format of the questionnaire is what leads participants to exhibit their response styles.

Extreme Response Style (ERS)

Evidence in the literature suggests that there are groups of participants who are likely to exhibit Extreme Response Style (ERS), which is a tendency to endorse extreme answer options rather than middle ones (van de Vijver, 1998; Cheung & Rensvold, 2000; Hui & Triandis, 1989 in van de Vijver & Poortinga, 2005). On a personality questionnaire with a Likert scale from 1 to 5, for example, participants who belong to groups high on ERS are likely to systematically endorse responses 1 and 5 and avoid

middle ones. Low ERS participants are likely to cluster around the middle (2, 3 and 4) whereas no ERS participants' scores are likely to be spread on all the scale. The means of a high ERS group, a no ERS group, and a low ERS group are therefore relatively the same, around the middle of the scale. The example in table 4.2 illustrates this concept clearly by comparing a set of hypothetical data from three ERS level groups. Even though the means of the three groups are identical, the items themselves are not endorsed similarly by participants in the three groups, and therefore cannot be directly compared. Cheung and Rensvold (2000) argue that groups or cultures high on ERS tend to value sincerity and conviction, and therefore choose extreme responding to reflect that. While cultures low on ERS appreciate modest and non-judgmental individuals and thus tend to endorse less extreme answers. The presence of ERS, whether high or low, suggests that responses have different meanings to different groups (Cheung & Rensvold, 2000), making the results incomparable between them. In the presence of ERS, measurement unit is inequivalent between the groups.

Acquiescence Response Style (ARS)

ARS on the other hand is characterised by one group consistently scoring higher or lower than other group(s) (Rust & Golombok, 1999; Cheung & Rensvolt, 2000). Considering the example of the personality test above, a group with high ARS is likely to agree or disagree to an item more than other groups, which makes it uniformly different than them; higher or lower. The uniformality of this response style makes comparison between groups possible, but indirect. That is, if a group scores consistently higher than another, then the difference between them is systematic. Table 4.2 illustrates the

relationship between high ARS and no ARS groups. The means of the two groups are different, but the style of responding is relatively comparable. If, for example, group A (high ARS) are consistently agreeing more than group B (no ARS), the difference between them is systematic. Therefore the difference between genders in group A can be compared to the difference between genders in group B. However, a person's score from group A cannot be directly compared to another person's score in group B. This is because the participant from group A who answered 1 is responding similarly to the participants who responded 2 in group B.

ARS poses a threat to scalar equivalence, the direct comparison of scores, but does not create measurement unit inequivalence (Cheung & Rensvold, 2000). That is, scores are comparable, but not directly a described above. Cheung and Rensvold (2000) explain that acquiescence could be the result of believing that a higher score is a better score or other culturally induced behaviours.

Example	Group	Definition	Raw score	Mean
Example A				
	High ERS	Extreme scores	1,5,1,5,1,5,1,5,1,5	3
	Low ERS	Middle scores	2,3,4,2,3,4,2,3,4,3	3
	No ERS	Spread scores	1,2,3,4,5,1,2,3,4,5	3
Example B				
	No ARS	Spread on the scale	1,2,3,4,5,1,2,3,4,5	3
	High ARS (say positive)	More agreement than No ARS group (+1)	2,3,4,5,5,2,3,4,5,5	3.8

Table 4.2: Hypothetical example of A) high, low and no ERS groups of 10 participants each and B) high and no ARS groups of 10 participants each

Source 3: Social desirability responding (SDR)

Other types of cultural differences in responding have also been discussed in the literature. For example, Lalwani, Shavitt and Johnson (2006) argue that collectivistic cultures, such as in Singapore, are likely to score high on the social desirable responding (SDR) scale; the likelihood of presenting oneself in a socially desirable way. However, they also differentiate between two types of SDR, self-deceptive enhancement (SDE) and impression management (IM).

Impression Management (IM)

IM assumes that test takers are responding in a way that makes them resemble the majority of the population or the norm. The rationalisation behind this pattern of SDR is that collectivistic cultures are more interdependent and follow, to a certain extent, a set of shared values and goals (Kabasakal & Bodur, 2002; Lalwani, Shavitt & Johnson, 2006). Therefore they tend to endorse items that involve behaviours favoured by the majority.

Self Deceptive Enhancement (SDE)

SDE assumes that respondents are presenting themselves in a better light by exaggerating their own abilities and skills or deceptively enhancing their image. Individualistic cultures on the other hand are less dependent and tend to follow values and goals that are independent of other group members. Uniqueness is an attractive quality in such cultures that drives respondents to adopt items that present them in more positive and unique way than the rest of the group.

Communalities between IM and SDE

Both SDE and IM affect the way participants perceive items on a test, and therefore affect their score on SDR (manipulating self-image), but in different directions. Social desirability responding, in its two forms, is considered as another manifestation of instrument bias, since the scores are affected due to the method of measurement rather than the construct being measured. Consequently, both types of SDR styles may affect participants' scores on a test, perhaps more on items that are culturally dependent.

Source 4: Purpose and Motivation

Perhaps another type of instrument bias relates to the purpose and question format of specific questionnaires. Schmit and Ryan (1993) investigated the differences in factor structure on the NEO-FFI between job applicants and college students. They found that the data from the student sample fit the NEO-FFI model better than data from job applicants. They also labelled one of the factors from the job applicants' data as "ideal-employee" factor because it combined all the desirable work-related personality characteristics. This suggests that the purpose of taking the questionnaire also has an impact on participants' scores. van de Vijver and Hambleton (1996) referred to this as the motivation behind test taking. The method of assessment used in this context had implications on the scores of test takers, due to the motivation behind taking the test rather than the characteristics being measured.

Source 5: Fakability of items

On the other hand, the test itself and perhaps the way items are written could result in another source of instrument bias. A study by Ballenger, Coldwell-Andrews, and

Baer (2001) showed that, when instructed to present themselves in a more positive light, participants from a clinical sample scored significantly different than participants from another clinical sample that received only standard instructions on the NEO-PI-R®. In another study, Furnham (1997) instructed three groups of participants consecutively to fake god, fake bad, and respond honestly to NEO-PI-R® and found that consciousness, agreeableness and neuroticism are fakable dimensions.

Implications of Instrument Bias

In conclusion, instrument bias is an anomaly that affects test scores because of issues related to the instrument such as: familiarity with response style, extreme response style, acquiescence response style, self deceptive enhancement, impression management, motivation, and finally the fakability of the test. These need to be minimised because they erroneously affect inferences drawn from cross-cultural research.

Dealing with instrument bias

Familiarity with response format is relatively difficult to measure, though asking key stakeholders from the target cultures could help answer this question. The Arab world, China, Spain and the UK are all countries where questionnaires have been used though not necessarily to the same extent. Organisations across these four cultures are increasingly using personality questionnaires in their recruitment and development practices. Additionally, there is no time limit associated with the test so it is therefore assumed that familiarity with response format is not a challenge in the adaptation of Orpheus across those languages and cultures. From a motivation point of view, all

participants who took part of the cross-cultural adaptation of Orpheus did so voluntarily with only the feedback report as motivation. The assumption was therefore that participants' motivation level would not radically contaminate the data.

Dealing with Response Style

ERS and ARS can be dealt with during the test construction phase or after the data collection. While developing the test, writing positive and negative items forces respondents to break their pattern of agreement or disagreement to items (Rust & Golombok, 1999). Participants who acquiesce positively, for example, will agree to the item: "I enjoy being the heart of a party" but will have to disagree if they are faced with the opposite of this item: "I hate being the heart of a party". In Orpheus, several methods were put in place by the test developers to control for ARS. Firstly, the nearly half the 190 items in Orpheus are written in a positive direction whereas the other half in a negative direction. This however does not guarantee that participants will answer all items attentively and will not contradict themselves, so four Audit scales discussed in Chapter 3 were also developed to measure any contradictions on these items (Rust & Golombok, 1999).

On the other hand, Cheung and Rosenvold (2000) suggest dropping items that show ERS from the questionnaire as the most conservative method of dealing with ERS. They argue that the limitation of this approach is that it might have implications on the construct validity of the test depending on the content and the number of items that show signs of ERS. As for dealing with ERS, Orpheus relies on "within subject standardisation"- also known as ipsative rescaling- to control for ERS. This method entails deducting the mean of each candidate's responses from his or her response on

every item and then dividing it by this candidate's standard deviation:

$$z = \frac{x - \bar{x}}{sd}$$

whereby $\bar{x}$ is the participant's average on all items, x is the participant's score on a specific item and *sd* is the participant's standard deviation based on his or her scores. Cheung and Rosenvold (2000) argue that this method renders comparison between subjects inadequate because each candidate's SD is different from others' SD. However, they argue that this method is effective when a test comprises of a large number of items with low inter-item correlation.

Dealing with Social Desirability

As with most personality questionnaire, Orpheus has a built-in honesty check that measures self-deceptive enhancement. However, further analysis will be conducted on the cross-cultural data to examine the two forms of SDR. The honesty check is also used to measure the degree of fakability that the participants might have attempted while taking the test.

4.5.2.2. Administration bias

Administration bias is associated with environmental or communicational differences in administering the tests, and which lead to affecting the scores of the comparison groups differently (van de Vijver, 1998). The theory of True Score, the basis

of classical test theory, entails that candidates' observed score is a combination of their true score and some random error as shown in the formula below (Cronbach, 1990; Kline, 1993; Rust & Golombok, 1999; Fife-Schaw, 2006):

Observed score=True score + Error

The error could be positive or negative, enhancing or underestimating the true score, and could be related to the test itself or to external nuisances (Fife-Schaw, 2006). Error unrelated to the test could result from administration bias, such as poor instructions, inconsistency in administering the test, poor physical environment (lighting, noise etc), or participants' feeling at the time of test taking (illness, fatigue, stress etc) (Kline, 1993; Rust & Golombok, 1999).

Source 1: Test Instructions

Tests usually come with a set of standardised instructions to be read during test administration, and also recommendations on best practice in administration settings. Any instructions made by the test administrator that do not follow these recommendations may bias test takers' responses either negatively or positively and result in error (Rust & Golombok, 1999). Fair comparison between candidates relies on an equal treatment of all participants. Receiving the same instructions about taking the test in a relatively similar environment is one way of ensuring this equality in treatment. The ITC guidelines (2005) on computer-based testing, clearly state that the valid and reliable interpretation of scores assumes that the test has been administered in a standardised way.

Source 2: Test administration across cultures: Computer based testing

Although administration bias can have an effect on participants' score, it generally has small consequences on equivalence between multi lingual versions of tests. An exception to this is the case where the instructions being given in the two cultures are significantly different. However, the most severe forms of administration bias result from computer-based testing.

Some countries have limited access to Internet and computers, which can affect participants' performance on computer-based tests. A culture group that has access to slower Internet connection in its county is likely to spend more time taking the test than another group using a faster Internet connection and consequently their performance might be affected by fatigue or stress.

On the other hand, computer based administration can also be a threat because it might lead to sample bias. That is, due to the method of administration some participants become marginalised from the sample. We will discuss this in more detail in the following section under sample bias (digital divide).

Source 3: Dealing with administration bias

Administration bias can be dealt with using standardised verbal instructions and standardised testing environments (Rust & Golombok, 1999). The British Psychological Society recognised the importance of this issue and as a result launched two qualifications, Test Administration and Level A, which focus on training test administrators in standardising testing procedures. Nevertheless, standardising testing environments could create a problem for researchers with little funding to spend on organising testing centres and even more so for cross-cultural researchers whose data needs to be collected from several countries. The International Test Commission (ITC)

has recently developed best practice guidelines for Test Use (2000) and Computer-Based Testing (2005) to help practitioners overcome and control biases caused by test administration.

Orpheus data was collected using predominantly a computer-based version with standard instructions at the beginning of the questionnaire. In China, the data was mainly collected using paper and pencil version of the test but was conducted under standardised conditions. Crucially, differences between paper and pencil and computer based testing have been shown to be negligible (Bartram & Brown, 2004). Additionally, the electronic version of Orpheus was carefully designed to contain the instructions and the 190 items all on the same page as to avoid prolonging the time of test taking for some participants due to slower internet connection in their country.

Sample bias

Sample bias was originally the third type of bias that falls under method bias (van de Vijver & Tanzer, 1997; van de Vijver, 1998; Byrne & Watkins, 2003) but it was later removed from the theory (van de Vijver & Poortinga, 2005), though there is no clear reason for that. Sample bias is a potentially dangerous type of bias because it does not necessarily affect the equivalence between multi-lingual versions, but the validity of the inferences that test users make of them. Sample bias is directly associated with specific characteristic of the sample, such as age, gender and education level and their consistency across the comparison groups.

Source 1: Samples of convenience and snowballing technique

When collecting data to make inferences about the general population, it is

important to collect it from a sample representative of the one it is being generalised to. There are two systematic methods of sampling that maximise the likelihood of a sample resembling the general population: random sampling and matching (van de Vijver, 1998; Pelham, 1999). Random sampling consists of randomly selecting participants from the overall population whereas matching consists of matching participants in two groups on some specific criteria, such as age or educational level.

In real life, researchers tend to rely on samples of convenience and snowballing techniques as the easiest sources of data. Samples of convenience are direct contacts of the researcher, such as friends, family, co-workers, and so on and snowballing technique relies on using the sample of convenience to recruit future participants. This method of haphazard sampling has strong implications on generalisability, and also on cross-cultural comparison (McCrae & Costa, 2003). Although it is a practical method of data collection, it is likely to contain certain percentages of age, gender and education in one group that are unequal to the other cultural groups of interest. As an example, let us imagine that a UK-based researcher is interested in comparing British and Chinese individuals on the Big Five personality characteristics. The researcher uses a sample of convenience in the UK, which consists of friends, family members, and co-workers who are likely (at least the majority) to be around the same age group as the researcher (say 50) and relatively the same socio-economic background. However, the easiest way for this researcher to collect data in China is through a friend who is a lecturer in a university in a remote area in China. Regardless of the quality of the adaptation of the tests and the construct equivalence, relying on samples of convenience and snowballing technique can show differences in the tests but due to the sample rather than inequivalence of the construct

being measured and vice versa. In this case, comparing an older group in one culture (UK) to a significantly younger group in another culture (China) can result in mean differences on items due to age differences. In fact, Costa and McCrae (1992 in McCrae and Costa, 2003) suggested in a cross sectional study that two personality characteristics agreeableness and conscientiousness increase with age. Having older participants in the Chinese group can lead the researcher to wrongly deduce that Chinese people are more conscientiousness and agreeable than their UK counterparts. However, these differences are only caused by reasons extraneous to the test itself that relate to the characteristics of the sample, which in this case in age.

Source 2: Digital Divide

Collecting data electronically could also be another example of sample bias. In this case, the method of administration (computer-based) favours a sub-group of the general population that has access to computers. This phenomenon is referred to as digital divide and illustrates the way administration bias can lead to sample bias. As an example, according the International Telecommunication statistics (2005), the average of PC per 100 inhabitants in Europe is 30.21% compared to 2.24% in Africa, so data collected electronically from both continents does not represent the same fraction of people there. As another example, the average of PC per inhabitant in the UK is nearly double that of Europe, 62.88%, whereas in Lebanon, it is 11.45%. Accessibility to Internet and computers (administration bias) therefore limits the participation of certain percentages of the population in research (sample bias). It is likely that people who have access to computer and Internet are systematically different than those who do not have access to

them in terms of socio-economic status, education etc.

Source 3: Self-Selection bias

Self-selection bias (or non response bias) is another source of sample bias, which arises from people choosing to take part of a research or not. Pelham (1999) argues that "people who choose to answer surveys are systematically different from people who choose not to do so" (p88). As a hypothetical example, let us consider that a group of researchers at a university are interested in measuring students' attitudes to exams. They hand out surveys to all 5000 students at the university and receive 1000 back. The analysis shows that students believe exam questions are too difficult and do not reflect the material learnt in classrooms. Although a sample of 1000 is large, it only constitutes 20% of the overall sample and might therefore not be a representative one. For one, students who respond to these surveys are likely to be those who failed their exams and found an opportunity to relief their frustration. The other 4000 students who did not take part of the survey create sample bias by choosing not to participate. It is therefore important to consider the response rate of any data collection process as it may produce misleading information (Pelham, 1999).

Implications of sample bias

In cross-cultural comparisons, sample bias is dangerous because of two outcomes it could result in. Firstly, the incomparability of samples can lead to wrongly assuming equivalence between multi-lingual versions of tests. To illustrate, let us consider Costa and McCrae's (1992) findings that agreeableness increases with age. Let us also hypothetically consider that Chinese are more agreeable than their British counterparts. A

younger Chinese sample compared with an older British one can lead to the assumption of equality of construct between the two cultures whereby this is not the case. That is, if our samples are significantly different in age we might mistakenly assume that there are no differences between the two cultures on agreeableness.

Conversely, sample bias could lead to assuming inequivalence but due to differences in the sample rather than cultural ones. That is, assuming that one culture is more agreeable than the other but due to the characteristics of the individuals in the sample rather than actual cultural differences.

4.5.2.3. Dealing with method bias

Finally, as discussed earlier, sample bias can be controlled for by using random sampling or matching techniques. However, even random sampling suffers from other forms of bias, such as self-selection bias. Pelham (1999) proposed substituting paper and pencil surveys by phone interviews as a way of increasing response rate. Although this might constitute a viable solution in market research, it is unpractical to conduct personality inventories over the phone. The most practical solution currently used is collecting data about gender, age and any other relevant characteristic from test takers to measure and account for any differences between samples. By doing so, researchers can account for the variance explained by factors that do not relate to what is being measured but affect it in any case.

4.5.2.4. Item bias

Item bias, better referred to as differential item functioning (DIF), is the last form

of bias that could obstruct the achievement of comparability of scores. An item is considered biased if members from different cultural groups but with similar score groups score differently on the item (van de Vijver, 1998). In the context of ability testing, an item is biased when members of one group (i.e. females) with a similar overall score on the ability measured to members of another group (i.e. males) are more likely to get the item wrong or right. Similarly, in the context of personality testing, an item is biased when members of one group with a similar score on a construct than members of another group, are likely to endorse the item differently. This definition, however, could be misleading, because not all discrepancies in performance between groups are the result of item bias. When DIF is detected, it is crucial to differentiate whether this discrepancy in performance is the result of *item bias* or *item impact* (Slocum, Gelin, & Zumbo, 2003; Zumbo, 2006).

4.5.2.5. Distinction between item bias and item impact

If the differences in performance between the groups of interest are due to anomalies at item level, the incongruence in performance is therefore caused by variables extraneous to what the test is measuring. That is, individuals from different groups are performing differently because of the malfunction in the test rather than differences on the construct being measured. In this case, DIF is referred to as item bias. However, this discrepancy could also be the result of an existing and real difference between the groups on the variable of interest, such as the difference between men and women on empathy, on one of the scales of BarOn Eqi (BarOn, 2002). In this case, DIF is referred to as item impact. That is, the item is rightly differentiating between two or more groups of

respondents due to existing difference measured by the item. Little emphasis has been put on the development of techniques that help to distinguish between item bias and item impact (Zumbo, 2006). Nevertheless, the literature on statistical techniques for detecting differential item functioning between groups is vast and is discussed in full detail in Chapter 8.

4.5.2.6. Sources of item bias

Although the Theory of Equivalence and Bias (van de Vijver & Tanzer, 1997; van de Vijver 1998, van de Vijver & Poortinga, 2005) does not specify any sources that can be grouped under item bias, it provides some possible causes that could create bias in an item. These examples apply mainly to the context of cross-cultural comparisons whereby the groups of interest are cultures. Mistranslation of items and the irrelevance of their content in the target cultures are two sources of item bias that have been discussed in the literature (van de Vijver & Hambleton, 1996). Yet, these have not been clearly and systematically classified in the theory as the sources of method bias. Whilst we will rely on the definition and examples van de Vijver and Leung's theory and other examples from the literature, we will define and explain sources of item bias and label them as follow:

a) linguistic,
b) psychological and
c) conceptual (cultural) bias.

Moreover, Chapter 6 presents the findings of a qualitative study that reveals possible sources of item bias, which we will incorporate in this section as examples, but explain in full as a study in chapter 6. We will first clarify the distinction between *linguistic,*

psychological and *conceptual equivalence* and the other types of equivalence previously discussed before defining and discussing the sources of item bias.

Linguistic, psychological and conceptual equivalence

Linguistic, psychological and conceptual equivalence are three types of equivalence that can be distinguished from the other types of equivalence previously discussed. Whereas construct, measurement and scalar equivalence operate at the scale level, *linguistic, psychological* and *conceptual equivalences* function on item level.

Linguistic equivalence focuses on similarity of wording between items and is the primary goal in the "rendering of items into the target language to capture the meaning of the original item" (Butcher, Cheung, & Lim, 2003; p3).

van de Vijver and Jeanrie (2004), Butcher (2004), and Butcher , Cheung, and Lim (2003) also recognise the importance of psychological equivalence in reaching full equivalence between items. van de Vijver and Jeanrie (2004) classify items as psychologically equivalent when they serve the same psychological purpose in all languages. That is, the psychological effect the item reflects in one culture, should be similar to the one reflected in the second culture.

Finally, conceptual equivalence (also referred to as cultural equivalence) focuses on the cultural suitability of wording in each culture (Marsella et al., 2000 in Hambleton, 2001; Cheung, 2004 b). This is particularly important in reaching equivalence between items because items that are culturally out of context become meaningless even when they are well translated.

These types of equivalence will become clearer when we define each of them, illustrate with examples, and outline the sources of item bias that threaten each of them.

Type 1: Linguistic bias

Linguistic bias results from mistranslation, inappropriate use of wording, or even mistranslation of idioms and the use of colloquialism (van de Vijver & Jeanrie, 2004; Marsella et al 2000 in Cheung 2004 b). In an example by Hambleton (1996), the questions "where does the bird with webbed feet live?" proved to function differently with a Swedish sample than with most other European samples. The term "webbed feet" was translated to "swimming feet" in Swedish, which rendered the item easier in the Swedish version than the English one. Therefore members of one group (Swedish), with the same overall ability on the test as their counterparts in the other group (English), became more likely to answer this question correctly. However, this likelihood is due to an anomaly in the item rather than a discrepancy in ability between the groups.

Mistranslations as such challenge the *linguistic equivalence* between multilingual versions of tests, by threatening the attainment of the same literal and connotative meaning on an item (van de Vijver & Jeanrie, 2004). Although linguistic equivalence is important, it is not sufficient for comparability of items.

Source 2: Psychological bias

Psychological bias relates to situations where the psychological impact of the item is not the same in the two or more given cultures (see Cheung, 2004 a). The psychological function the item serves, especially in personality testing, affect the way it is viewed by test takers. The differential psychological effect can arise even when the translation is accurate and the linguistic equivalence is achieved. For example, emotions could have different intensity across cultures, which make respondents in different

cultures exhibit stronger or weaker psychological reactions towards an item (Marsella et al. 2000 in Cheung, 2004 b). This source of item bias challenges the *psychological equivalence* between items, that is, the equivalence of the psychological effect the item has in the different languages versions (van de Vijver & Jeanrie, 2004). Items should therefore be adapted in a way that would ensure that the multilingual versions of the same item serve the same psychological function in both languages.

Source 3: Conceptual bias

Conceptual bias, also referred to as cultural bias, is another source of item bias that relates to the relevance of the item content to the target culture (for examples see van de Vijver & Tanzer, 1997; van de Vijver & Hambleton, 1996; Byrne & Watkins, 2003; Hambleton, Merenda & Spielberger; 2005). That is, the concepts covered by a certain item need to be meaningful in the target culture. This source of bias is also independent from the quality of the translation. Cheung (2004 a) provides an example from the MMPI-2 that best illustrates this source of bias. The item "I used the play hopscotch and jump rope", even if well translated, will present a problem in cultures where hopscotch and jump rope is not common as a children's game. Therefore content should also be adapted to be culturally appropriate in order to avoid creating bias.

This source of bias challenges the *conceptual equivalence* between multilingual items (Jeanrie & Bertrand, 1999; Cheung, 2004 a). Conceptual equivalence accounts for the suitability of the situations and information contained in each item and also to the equivalence in calibration such as using metric vs. imperial systems (Jeanrie & Bertrand, 1999).

4.6. Conclusion

In summary, it is evident that bias in all its forms constitutes a threat to the equivalence between multi lingual versions of tests. The relationship between types of equivalence is hierarchical whereby one should be achieved before the other so different types of bias affect different types of validity (van de Vijver & Leung, 1997). The types of bias covered in van de Vijver and Leung's Theory of Equivalence and Bias are what we would like to call *higher order* ones, which have major implications on scores comparability. However, there are *lower order* types of bias such as linguistic, conceptual and psychological bias that do not constitute a major part of this theory though they are acknowledged by the authors. Higher order types can take several forms, which are referred to as lower order, though they are all types of bias. Lower order bias can results from different sources, and these sources consequently lead to higher order bias. For example, linguistic, psychological and conceptual biases all fall under item bias. But each of them is still a type of bias, not sources of bias. Here is an example to illustrate this classification:

- Higher order bias: Item Bias
- Lower order bias: Linguistic Bias
- Source of lower order bias: Grammatical inconsistency

In this case, grammatical inconsistencies between the tests (such as passive and active voices) will lead to linguistic bias. Linguistic bias is a type of bias that affects the tests on the item level and therefore leads to a higher level of bias, item bias. As another example, instrument bias can be considered as a lower order type of bias and can arise from several sources, such as familiarity with response format or social desirability responding. If instrument bias exists, it will lead to a higher order type of bias, in this case method bias.

This categorisation can facilitate the understanding and application of the theory in practice. Further discussion about lower order bias in item bias will follow in chapter 6 based on the findings of study 1.

Linguistic, conceptual and psychological biases are the basic types of bias to be avoided in order to ensure that item bias is not existent. Once this is achieved, the tests have already achieved a certain level of equivalence that we argue should be referred to as linguistic, conceptual, and psychological equivalence. The same classification used for bias can also be applied to the concept of equivalence. Thus, these three types of equivalence are prerequisite to the higher order types of equivalence discussed in van de Vijver and Leung's Theory of Equivalence and Bias: construct, measurement unit, and scalar equivalence. As discussed earlier, the relationship between types of equivalence is hierarchical therefore linguistic, psychological and conceptual equivalence precede construct, measurement unit, and scalar equivalence as illustrated in figure 4.1 below.

As for the relationship between the types of bias, we argue that it is sequential. That is, bias can occur at different times during the adaptation process. For example, item bias (and the lower order types that relate to it) can only occur during the translation part of the process. On the other hand, method bias can occur between recruiting the sample and administrating the test. Finally, construct bias is likely to precede the test adaptation altogether.

As result of the literature review, Figure 4.1 below represents the suggested Theoretical Framework of Equivalence and Bias that builds primarily on van de Vijver and Leung's (1997) Theory of Equivalence and Bias. This framework incorporates higher and lower order types of equivalence and bias as well as methods to deal with each of

them. This provides an easy to follow visual aid that can guide any test adaptation process. The big rectangles with bold writing represent the higher order equivalence and bias whereas the ellipses represent the lower order types and the long rectangles represent the sources of bias. The unidirectional arrows represent the sequential relationship between bias and equivalence and their order of possible occurrence. The bidirectional arrows represents the possible methods that can be used to deal with the particular types of bias is points to.

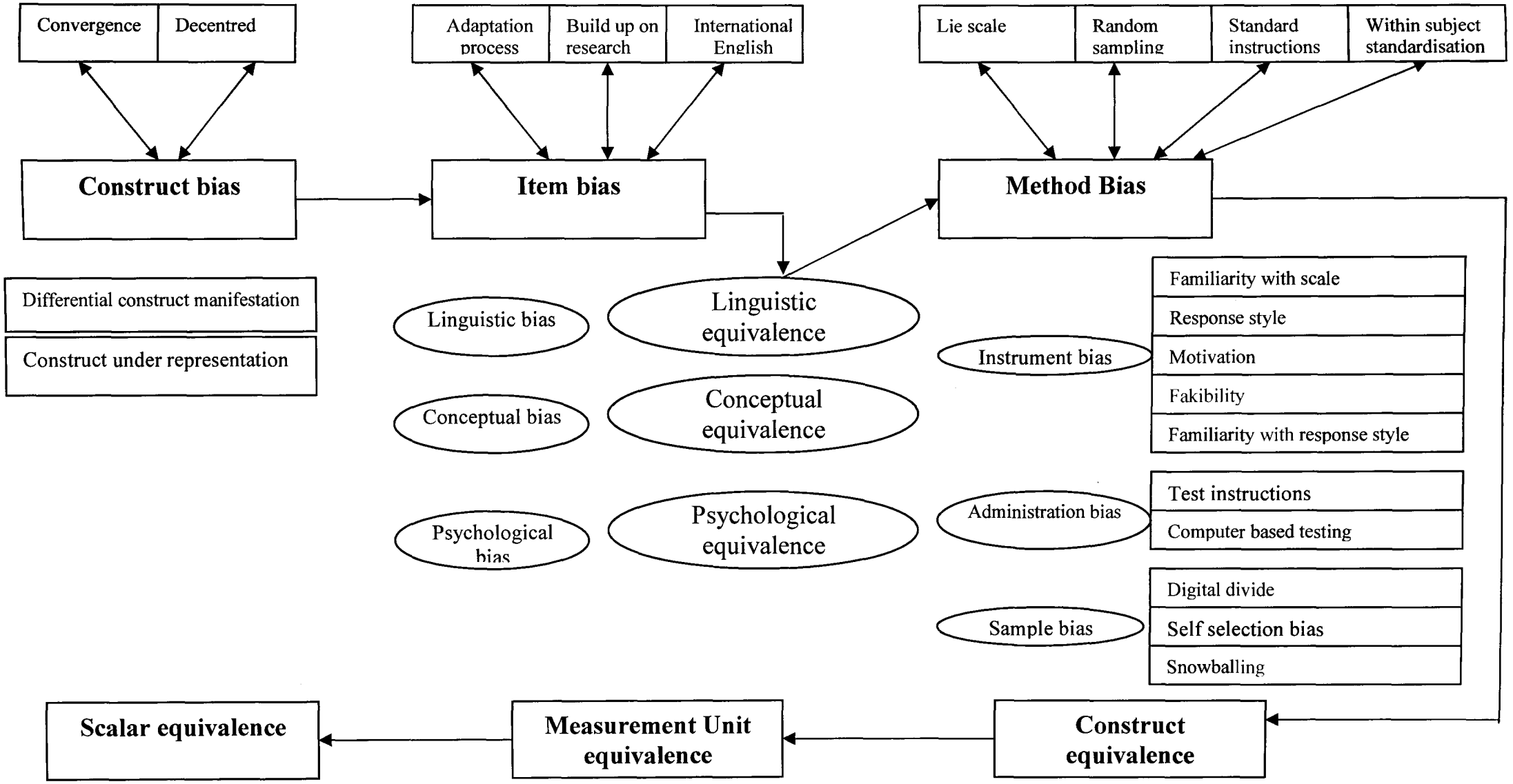


Figure 4.1: Theoretical Framework of Equivalence and Bias

Chapter 5:Overview of the process of test adaptation- summary of methods

Glossary

Abbreviation	Stands for	Definition
5.1.1 FT	Forward Translation	Translation from the Original language to the target language
BT	Back Translation	Translation from the Target language to the Original language
OL	Original Language	The language in which the questionnaire was developed
OC	Original Culture	The culture of the country where the questionnaire was originally developed
TL	Target Language	The language to which the questionnaire is being translated to
TC	Target Culture	The culture of the country where the questionnaire is being translated to
NS	Native Speaker	A native speaker of the Target language, also knowledgeable in the Original Culture
PM	Project Manager	A psychometrician who takes part of almost all the different stages of adaptation
V	Version	Language Version of Orpheus questionnaire

This chapter provides an outline of the adaptation procedure that was applied in this thesis to adapt Orpheus from English to Arabic, Chinese and Spanish and is divided into three main parts as follows:

Part one: Languages used and the challenges to reaching multi-lingual parallel versions in comparison to same language;

Part two: Methods for maximising equivalence and their limitations; and

Part three: The Test Adaptation Process.

We will start by explaining essential particularities of each language used in this research. We will then briefly highlight the main limitations of each test adaptation method with examples and how other methods can be employed to control for them and thus form the test adaptation process that we implemented. The last part of this chapter summarises the

adaptation process leading into the four studies that are discussed in the following chapters. This chapter is only an overview of the methodology and each of the topics discussed will be explored further in subsequent chapters.

5.2. Languages

The thesis focuses on four main languages, namely: Arabic, Chinese (Mandarin), English, and Spanish. Each one of them has its particularities, which are explained below.

5.2.1 Arabic

Arabic cannot be considered as the first language of any individual per se because of the difference between the *spoken* and *formal (*or *written)* Arabic. The latter is the common language shared by all Arabic speaking countries, and is the language used in books, news, official documents, or any written material. The spoken language is the dialect particular to every Arabic speaking country and is used in everyday conversations within each country. Individuals from Arab countries are native speakers of their own spoken dialect but they all learn formal Arabic at school. Some dialects are close to each other and could be understood by some neighbouring countries such as the gulf countries, but this is not the case for all Arabic speaking countries.

"The connotation Arab refers to a group of people whose behavioural pattern is unique because of their culture, language, religion, and even their nationalism" (Harris & Moran, 1996, p347). The Arabic language is the official language of 21 countries that constitute the Middle East and North Africa (MENA) region: Algeria, Bahrain, Djibouti,

Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tunisia, United Arab Emirates, and Yemen (Maamouri, 1998). Comoros, Chad, and Eritrea also have Arabic as their official language but are not part of the MENA region, which typically represents the Arab world in business, economy and academia.

Arab countries are different in terms of socio-economic status, Islamic heritage, culture and history, but they could be divided into two main groups: Mashrek and Maghreb (Maamouri, 1998). Mashrek consists of four subgroups: 1) Egypt, Sudan, 2) Syria, Lebanon, Palestine, Jordan, 3) Iraq, Saudi Arabia, 4) Bahrain, Oman, UAE, Kuwait, Qatar and Yemen, whereas Maghreb consists of: Mauritania, Morocco, Algeria, Tunisia, and Libya. Maamouri (1998) argues that this division suggested by the UNESCO is arguable and that Somalia and Djibouti are not part of Mashrek or Maghreb because their Arabic has a "reduced, though important, influence in their respective educational structures" (p 8). He also argues that although the Maghreb countries are referred to as Arabs, they are not similar speakers and users of Arabic as the Mashrek countries. Additionally, ethnic origins are usually overlooked when referring to Arabs and the focus is mainly on the fact that they share the same language (Maamouri, 1998). For the purposes of this study, the Mashrek region will be to focus of our research in order to minimise overgeneralizations to Arabic speaking countries that might be different ethnically among other things. Very influential research about the Arab world has also centred on either Mashrek or Maghreb (i.e. Hofstede's cultural dimensions and the Globe project) whereas other research that included both made the distinction very clear and incorporated information from both before generalising (i.e. PISA project).

Therefore we will use the term "Arab world" to refer to countries in the Mashrek.

5.2.2 Spanish

Similarly in Spain, the official language is referred to as Castilian or Spanish, and will therefore be used interchangeably. However, there are other languages that are co-official in different states and these are: Catalan (spoken by 74% of the population), Valenciana (17%), Galician (7%), and Basque (2%) (Bureau of European and Euroasian Affairs, 2004). This suggests that the country is bilingual, with one common language for all. For the purposes of this research, Spanish or Spaniards refer to any person from Spain who speaks Castilian.

5.2.3 Chinese

Mainland China or China is the common name used to refer to the People's Republic of China, which is different from Republic of China, also known as Taiwan. In most of China and Taiwan, Mandarin is the official language and derives from Sino Tibetan language families and is spoken by approximately 900 million people (DataMonitor, 2004). Mandarin is language taught in most schools, spoken on TV and radio stations and written in magazines and newspapers even in provinces where Cantonese (the second most common language in China) is spoken in everyday language. There are several dialects spoken in China and Mandarin could also be considered as a dialect spoken in some cities such as ShangHai. Mandarin is also one the official languages on the UN and due to its representativeness, Mandarin was chosen as the focus

of our research and throughout this thesis, we will use Chinese and Mandarin interchangeably and China to refer to the People's Republic of China.

5.3. Challenges to reaching equivalence between multi-lingual parallel versions

Test adaptation is cumbersome and time-consuming process and early detection of potential anomalies in the translation is highly recommended (Sireci, 1999). A good test adaptation ensures that differences measured between people using psychometric tests, within one culture or between cultures, are due to differences on the underlying construct being measured. Test adaptation is a procedure aimed at achieving validity and reliability of multilingual versions of tests through a cross-comparison with the original version. Multi-lingual versions of tests could therefore be considered as parallel versions and test adaptation as the procedure for validating those versions (Arffman, 2007). Typically, to establish parallel forms of a test in one language, item characteristics such as difficulty and discrimination of items in both versions should be equivalent (Rust & Golombok, 2001). The same applies to multi-lingual parallel versions; however, the challenges for developing equivalent multilingual items are greater than for equivalent items in one language. Difficulties in translation discussed in the previous chapter highlight some of these challenges. In the following section, we will draw attention to specific challenges with examples to illustrate.

5.3.1 Grammar

Each language is characterised by its own grammatical rules which is defined as

"the study of the way the sentences of a language are constructed: morphology and syntax" (Dictionary, 2007). Morphology relates to "the patterns of word formation in a particular language" whereas syntax relates to "the rules for the formation of grammatical sentences in a language" (Dictionary, 2007). Grammatical structures of sentences are not always equivalent because each language follows different grammatical rules, which may lead to differences in responding because of characteristics unrelated to the construct that the test purports to measure. To illustrate, suppose a test measuring grammatical knowledge in the following item:

"*Fill in the correct tense of the verb "to stay" in the following sentence:*

"Yesterday, I .... at my friend's house".

In this case, it is relatively easy to develop a parallel version in English that maintains the same level of difficulty and measures the same construct, for example:

Fill in the correct tense of the verb "to jump" in the following sentence:

"Last week, I .... over a fence".

Both verbs are regular and are as difficult/easy to conjugate, and both sentences follow the same structure. However, if a parallel version of this item was to be produced in another language, other considerations need to be taken into account. Firstly, it is important to find a verb that is equivalently easy or difficult to conjugate in the target language. Another point to bear in mind is whether the translation of this item will produce a sentence that follows the same structure. For example, if the item *" Yesterday I stayed at my friend's house"* is translated to Finnish, it will become: *"Olin eilen yötä ystäväni luona".* The verb in the Finnish version is no longer "stay" but it became the verb "to be" (Olla) meaning that "I spent the night at my friend's house". The verb is not

necessarily regular as in the English version therefore the same item is not necessarily testing the same grammatical concept and might be easier or more difficult in the target language.

5.3.2 Translation errors: alternative meaning

Other more common challenges that impede reaching equivalence between parallel multi-lingual versions relate to translation errors (Grisay, 2003). One of these challenges could be the result of using words that have alternative meanings. As an example from Orpheus, the item "if someone gave me too much change I will always tell them" means "if someone returned to me more money than they should have, I would always tell them". This item is one of several items designed to measure honest responding in people. However, during the adaptation of this item to Chinese, the item was initially translated as "if someone gave me too many changes (at work), I would always tell them". The word "change" has several meanings in English and was mistranslated into Chinese. The mistranslated item no longer measures honesty but perhaps it now measures assertiveness or a person's willingness to speak their mind. This consequently affects the whole meaning of the item in Chinese as well the underlying construct it is intended to measure, which in this case is honesty.

This is an example of translation error that affects individual items. However, the more items are affected by such problems, the more difficult it will be to reach linguistic, psychological, and conceptual equivalence between the test versions. This highlights the importance of minimizing the number of translation problems as early as possible in the adaptation process in order to maximise the likelihood of achieving the ultimate goal,

equivalence between parallel multilingual versions of tests.

5.4. Methods for maximising equivalence and their limitations

5.4.1 BT limitation: alternative meaning

There are several methods for assessing the quality of translation, one of which is back translation. In back translation technique, the test is first translated to from the OL to the TL, then translated back to OL by another independent translator. Typically, the back translation technique is characterised by a comparison of the original version with the back-translated one. Similarity between those two suggests a good translation. Back translation can help detect linguistic problems at item level (Hambleton, 1994; 2002); nevertheless many mistranslated items can pass through it unnoticed. Chapter 6 describes the strengths and limitations of back translation in detail. Reconsidering the *alternative meaning* example, back translation cannot account for such problems because 1) they could be translated wrongly into the TL and 2) whichever meaning they are translated to in the TL, when back translated to the OL the word will be exactly the same. Although there is a very close match between the original and back translated version, the mistranslation remains undetected. For example the "change" item can be mistranslated into Chinese into "change (in the workplace)", but the back translation could fail to detect that because both original and back translated English version will have the same word "change". A lot of words with alternative meaning will appear to be the same in the original version and the back-translated one although the meaning is different in the target language.

5.4.2 Dyads and triads limitation: non translation errors

Dyads and triads could be an effective tool that can be used in conjunction with back translation to control for some of its limitations. Dyads and triads, discussed in full detail in Chapter 6, consist of groups of two or three bilingual judges that scrutinise the quality of the translation in a panel discussion. Section 6.2.1 highlights the advantages and limitations of this method. With the addition of dyads/triads, back translation is followed by an in-depth qualitative assessment of the words in each item thus making the likelihood of alternative meaning words, and other translation inaccuracies, going unnoticed much slimmer. This combination renders the assessment of the quality of translation more accurate. It is important to emphasize that although dyads and triads can solve some of the limitations of back translation, they cannot detect all linguistic, cultural and psychological problems in test translation.

As a hypothetical example, let us consider the questions in figure 5.1 below, which is aimed at assessing a particular Maths competence in grade 3. When adapting this item to Arabic to be used in Lebanon, for example, it is important to change all the characteristics of the item that could be particular to one country or another. This could include the name of the character in the questions, the actual toy and whether it is used in the TC, and also the currency, which needs to be made understandable for children in the TC. The name Sourav could be changed to Maha, a more common one in Lebanon, the toy could stay the same if it is agreed that Lebanese children are familiar with it and the currency should be changed to Lebanese Liras (LL). This seems like a straightforward task, however, if the amount "48 cents" was changed to a realistic one in the target

Copyrighted information

Figure 5.1: example of an item measuring arithmetic ability

culture, it might be possible that this amount is "2350 LL", because the exchange rate of Lebanese Liras to British Pounds is very high. The item is therefore relatively easily adaptable into Arabic.

However, if the item was tested on a small sample, results might show that Lebanese children are performing poorly on the item compared to their English counterparts. The reason is that the adaptation rendered the item more difficult in Arabic than in English, due to differences in monetary values between the two cultures. It is easier to add coins to get to 48 cents, but much more difficult to add coins to get to 2350LL, because the number contains more digits. Such problems can go unnoticed when assessed qualitatively as they might be outside the scope of what problems qualitative methods can detect. The wording of the item in Arabic could be perfectly matching the wording in English and changing the currency made the item more accessible to Lebanese students. However, this might have affected the difficulty level of the item and made the parallel linguistic versions inequivalent.

These types of problems would be much more difficult to overlook if quantitative

techniques are also applied. Testing the item can reveal problems that are not easily detectible even by several bilingual judges. This is so because they provide evidence about how students respond to the item and hence that some underlying problems can be associated with it. Additionally, back translation and dyads/triads are qualitative techniques and tend to be subjective by nature even if they are conducted in a standardised manner, and thus cannot provide sufficient evidence for the linguistic equivalence between the sentences

5.4.3 Pre-Testing limitations: sample size

Pre-testing is discussed in detail in chapter 7, and consists of giving the questionnaire in both languages to a small group of people in each of the target cultures. The assumption is that if items are equivalent, people from the same culture should answer them relatively similarly whichever language they are presented in. However, differences in responding to the two language versions are not always indicative of translation problems.

The first obvious reason for such differences is participants' proficiency in the OL. Although samples in this kind of studies are usually bilinguals, it is not easy to assess whether they are equally proficient in both languages. Another source of variation in scores could be individual differences. Since the samples in pre-testing tend to be small, individual differences are not balanced between the comparison groups. That is, it is possible for one group to have more females than males or one age group and not another, or even be different on the construct being measured. Therefore statistical differences could possibly be detected although they are the result of real differences between the groups.

As an example, let us assume that most of the Chinese participants who filled out the Chinese version are on average older than the group who filled out the English version. Assuming that there is a negative correlation between age and the big five openness to experience scale, the younger group is therefore more likely to endorse items that reflect openness to experience in comparison to the other group. Although this can be detected statistically, it is not the result of linguistic problems in the item. Therefore the statistical differences cannot be properly interpreted until the item is scrutinised qualitatively.

These individual differences can usually be minimised by collecting data from large samples. Yet, the pre-testing is usually used in order to minimize the number of problems in the test before investing time in piloting the test and collecting data from larger groups. Therefore, pre-testing is a useful tool to the adaptation process though it should be followed by qualitative investigations such as cognitive interviews.

5.4.4 Cognitive interviewing

Cognitive interviews are in depth-interviews that aim to understand the cognitive process that goes into participants' mind when answering certain questions (Willis, 2005). These are discussed in detail in Chapter 7, and can be used to compare how participants think about the same item in English and in the TL. Therefore this could be a useful tool that can be applied on items that the pre-testing identified as behaving differently in the same culture. When cognitively interviewed, participants reveal information about how they understand and process the item in each language. Comparing the way participants think about the same item in different languages can help

identify whether there is linguistic problem in the item or whether the difference in responding is possibly random, and might disappear during piloting on a large sample. At this stage, a version can be produced with relatively high confidence that it is a well-translated version. This version can then be piloted to check it is measuring the same construct across cultures, and whether there are items that will not function well in one culture or across several, even if well translated. Those items can therefore be dropped out from the questionnaire without risking loosing items that could have functioned well if translated accurately.

5.5. The Test Adaptation Process

The following outlines the test adaptation process that we adopted in the form of four main studies. Each of these studies will be discussed in more detail separately in the following chapters.

5.5.1 Study 1: Translation and Monitoring

This study is presented in Chapter 6 and consists of the following 4 steps:

1- Forward translation

2- Dyads/Triads

3- Back translation

4- Dyads/ Triads

5.5.1.1. Forward translation

The adaptation process of each language version of Orpheus starts with a basic forward translation of the OL into the TL. A native speaker of the TL (NS1), also knowledgeable in both TC and OC, is briefed on the aims of the translation (appendix 3) and invited to translate the OL version of Orpheus (V1) into the TL to produce V2 of Orpheus.

5.5.1.2. Dyads/Triads

In each triad, V2 is revised in the presence of the PM and two native speakers of the TL. Dyads are only conducted when the PM is also a native speaker of the TL. The PM, NS2, and NS3 critically revise the forward translated version V2 item by item. Items are flagged when

1. The translation does not convey the meaning intended from the OL and/ or
2. The structure of the sentence needs revision and/ or
3. A word should be remove or added and/ or
4. A word or words need to be replaced by another more convenient one and/ or
5. Issues with grammar and/ or
6. Any other reason that NS2 and NS3 agree is affecting the meaning and/or equivalence.

A suggested translation is offered and later discussed with NS1 before developing V3- the revised version of V2.

5.5.1.3. Back-translation

Similarly to the forward translation, NS4 is briefed and presented only with V3 to back-translate it from the TL back into the OL to produce V4.

5.5.1.4. Dyads/Triads

Dyads and triads are run the same way as previously discussed in the presence of NS5 and NS6. However, they are now based on a comparison of the two versions in the OL, V4 and V1, as well as the translated version V3. Items are flagged whenever there is incongruence between any of the versions. The items are flagged as before and amended correspondingly.

5.5.2 Study 2 and 3: Pre-testing and Cognitive Interviewing

Chapter seven describes two very closely related studies and these are as follow.

5.5.2.1. Study1: Pre-testing

Approximately sixty participants in each target culture are invited to take part of the pilot study. Half of them are asked to fill out the original version V1 and the other half the translated version V5. Within each culture, the data is statistically tested for differences on the following two groups:

1. TC candidates filling out questionnaires in OL
2. TC candidates filling out questionnaires in TL.

Items are flagged and filtered to the next step when they show any significant differences in their endorsement of items.

5.5.2.2. Study 2: Cognitive interviewing

A semi structured cognitive interview is planned and run by the PM with twelve native speakers of the TL to review the items. The CI is planned as follow

1. introduction and framing
2. standardised instructions are presented to candidates
3. half the items are presented first in the TL and the other half in the OL
4. For each item, candidates rate their endorsement and described their thinking and feeling that lead them to agree, disagree, strongly agree or strongly disagree to the item
5. PM probe only when participants do not give enough information
6. Participants then rate the equivalence of multilingual items based on a *similarity rating scale* and suggeste amendments to the items accordingly.
7. The amendments are then agreed in a panel review session with two other native speakers.

At the end of each interview, a final version V6 is produced and piloted in the next study.

5.5.3 Study 3: Piloting

Study three is one study that is represented in two chapters 8 and 9. Each of the chapters covers different statistical analyses that are applied to this data set.

5.5.3.1. The pilot

The final target version V6 is given to approximately 200 candidates in each TC and 200 from the OC in order to investigate its validity in each, that is, the equivalence between the four language versions.

5.6. Conclusion

All the steps described in this overview chapter will be discussed in more detail in the corresponding chapters while drawing on the literature associated with each method. Additionally, each of the following chapters will discuss other techniques that can be used to serve the same purpose. However, a clear rationale will also be provided to explain the choice of methods in this study in specific. Figure 5.2 summarises the suggested Practical Framework of Test Adaptation

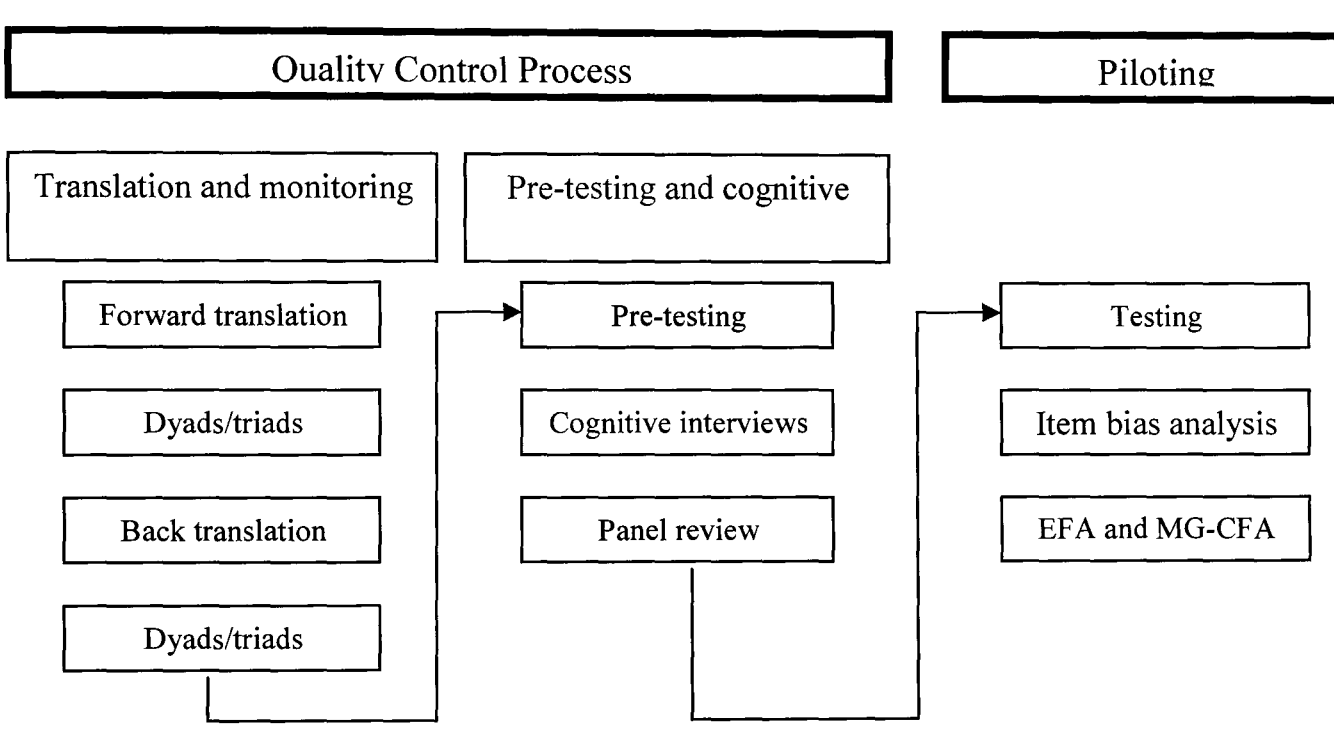


Figure 5.2: Practical Framework of test adaptation

Chapter 6:The translation phase- Using qualitative techniques to support the traditional back-translation method

"The language that enables us to communicate with one another also encloses us in an invisible web of sounds and meanings, so that each nation is imprisoned by its language, a language further fragmented by historical eras, by social classes, by generations."
Octavio Paz

6.1. Chapter summary

This chapter outlines the methodology employed in the first phase of the adaptation process, referred to as *Translation and Monitoring*. It is based on qualitative techniques and is designed to control the quality of the translation in the early stages of the process. However, this phase compliments the quantitative techniques that will follow in later stages to form the quality control process. In addition, the aim of this chapter is to uncover the problems commonly encountered in translating personality tests from English to Arabic, Chinese and Spanish.

Initially, we will review the literature on the common methods employed in test translation, such as back-translation and panel of judges' technique, while highlighting their strengths and limitations. Subsequently, we will discuss how this led to the development of the process of translation and monitoring that was adopted, which combines the use of bilingual judges (also native speakers of the target language) with the traditional back translation method. The analysis in this study is exploratory and revolves around the information that was gathered during the judgement process from the bilingual judges. During the analysis, the aim is to retrospectively underpin the item level problems commonly encountered in translating personality tests.

We conclude this chapter with implications for theory and practice. Theoretically, the findings support part of the Theory of Equivalence and Bias, mainly item bias and the sources it could possibly arise from. Practically, the findings provide important and

detailed information that could be used to implement guidelines and structure to the process of test translation and monitoring.

6.2. Introduction: Translation and Monitoring

6.2.1 Defining translation

"Translating consists in reproducing in the receptor language the closest natural equivalent of the source-language message, first in terms of meaning and secondly in terms of style" (Nida & Taber, 1982, p12). The "most natural equivalent" implies that the translation is not an act of word-by-word coding but rather an "act of communication" (Snell-Hornby, 1988, p 43). The meaning and style are thus the basic components that enable the reproduction of any sort of communication into another language. Newmark (1996) highlights that the only way to get to the synonym in the target language is through understanding the significance or purport of the word or text. The functionalist approach to translation also points out that the function or purpose of the translation is an important factor that affects the process of translation and choice of wording (Nord, 1997). That is, the purpose for which information is being translated dictates how equivalence can be established. When translating poetry for example, it is important to maintain the beauty of the sentences, and potentially the rhyming, which influences the choice of words and style of writing in the target language.

Translation originally gained attention as an essential tool for spreading religion and then poetry and literature, as a means of promoting and preserving a countries' cultural status (Delisle & Woodworth, 1995). Translation of different material depends

on the context of the material being translated. Newmark (1996) classifies translation into three types based on the most common contexts: a) non-literary, b) literary, and c) poetic. He rightly argues that the context facilitates reaching the meaning in the target language as the whole paragraph works together to convey the intended meaning. That is, the complexity of each sentence on its own is facilitated by the context of the material it belongs to.

6.2.2 Translation of psychometric instruments

In psychometrics and personality testing specifically, 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. Each group of items are designed to measure a certain personality construct (scale) and they are usually randomly mixed with other items measuring other constructs in order to make it difficult for test takers to guess what the items are aiming to measure. Therefore, items are necessarily listed somewhat out of context, unless the translator has knowledge about which items belong to which scale. This presents another complexity because knowledge in psychology becomes necessary for understanding the meaning of the scales and the underlying constructs they measure.

Additionally, in psychometrics, the psychological effect that a sentence or word conjures in test takers is an integral part of achieving equivalence in translation. *Meaning* and *style* are crucial to all types of translations, but in the context of cross-cultural test adaptation, the *psychological effect* of sentences is an additional criterion specific to personality tests and essential for achieving equivalent translation of these types of tests.

These three criteria, *meaning, style* and *psychological effect* are closely related to the concept of *linguistic, psychological* and *conceptual equivalence* discussed previously in chapter 4. The following section will illustrate the relationship between these translation concepts and the theoretical concepts of equivalence.

6.2.2.1. Linguistic, cultural and psychological equivalence

In the context of translating personality tests, sentences that are not equivalent in *meaning* are ones that have failed to achieve *linguistic equivalence*. Sentences that are not equivalent in *style* are ones that are appropriately written in one culture but not in the other, and are therefore *conceptually inequivalent*. An example of this in ability testing could be failing to adapt measurement systems, such as Kilograms and centimetres, into culturally appropriate ones. In personality testing, *conceptual inequivalence* results form language particularities such as sentence formulation, which may lead sentences to sound unnatural or "translated" in the target language.

Finally, words or sentences that have differential psychological effect in two cultures may lead to *psychological inequivalence* between items. For example the word "honour" and its literal equivalent in Arabic "شرف" (sharaf) have a different psychological effect in Arabic than in English. In individualistic cultures, such as UK or US, the word "honour" involves the individual's honour only. Whereas in collectivistic cultures, such as in Lebanon, this word tends to encompass one's own honour as well as one's close family and first cousins. Linguistic, cultural and psychological equivalence, as discussed in chapter 4, are essential in minimising item bias and are discussed later in the context of the findings in this chapter.

6.2.2.2. Back translation

The traditional back-translation method (Brislin, 1980) has been widely used in the area of cross-cultural assessment (Hambleton, 1993; Geisinger, 1994; Hambleton & van de Vijver, 1996; van de Vijver & Tanzer, 2004; Daouk, Rust, & McDowall, 2005). This method consists of two distinct parts, a forward and a back translation (Brislin, 1980). The forward translation is usually conducted by one translator who translates the material from the original language to the target language. The translated version is then presented to an independent translator, who has not seen the original language version, to translate it from the target language back to the original language. Although back translation has been widely used as a judgment technique to assess the quality of the translation and to detect potential item bias (Brilin, 1980; Hambleton, 1993), this method is fallible and is considered to be somewhat misleading when used on its own (Geisinger, 1994; Hambleton & Patsula, 2005). We will highlight these limitations, propose alternative methods from the literature that have been used to replace it, and conclude with the rationale for choosing to combine back translation and the alternative ones.

Limitations of back translation

Some argue that the closer the match between the original version and the back translated one, the worse the quality of the translation ("Translation Myths and Misinformation", 2005). The reasoning is that a close match between the versions could be the result of a word-for-word translation, which often leads to nonsensical sentences in the target language. For example, if the idiom "everything is coming together" is translated literally into Arabic, the same sentence can be replicated in English during the back translation. However, the idiom "everything is coming together" in English implies

that a situation is working out well in every way. Whereas in Arabic, the literal translation of this item implies that everything bad is happening at the same time. Hambleton (1993) highlighted additional criticisms about this judgment technique:

1. Naturally, the back translation technique leads to assessing the quality of the translation using the original language only and therefore biases may arise. As discussed earlier, the original and back-translated versions are used to judge the quality of the translation without reference to the translated version. Words with double (or more) meaning in TL, will translate into the same word in the OL. Thus, by comparing the back-translated version to the original, it might be mistakenly assumed that the translated version is appropriately translated because double meanings can easily pass unnoticed. For example, the word "sense" has several meaning including "the meaning or gist of something", "sound practical intelligence", "feeling or perception produced through the organs" and many more (Cambridge dictionary, 2007). A sentence "Something makes sense" could be mistranslated to "Something makes feeling" in the target language but when back translated to English, it might come back to the same "something makes sense", thus hiding the mistranslation.
2. Problems may also arise from the discrepancy between the translation skills of the forward and back translators (Hambleton, 1993). That is, translators may use translation rules or techniques that would ensure the replication of the original version during the back translation. This can lead to sentences that are unnatural or misleading in the TL, yet perfectly fine in the back-translated target language. For example, a more proficient back translator might correct a grammatical

mistake committed by the forward translator, thus hindering a problem with the translated version.

Summary of advantages and limitations of Back Translation

In summary, the back translation procedure on its own is directed towards reaching the linguistic equivalence of items without taking into account the cultural aspects of the target culture (Geisinger, 1994) and the information it reveals is very limited (Hambleton & Patsula, 1999). The ITC guideline D.2 clearly states that when adapting an instrument from one language to another, not only the literal meaning of items should be preserved but also the connotative one (Hambleton, 2001). That is, the literal wording of the sentence is important but only if it conveys the same meaning in the target language. As demonstrated in the examples earlier, we can argue that the use of back-translation alone -even if adequately executed- does not ensure the production of an appropriately adapted target version (Brislin, Looner & Thorndike, 1973; Hambleton, 1993). Conversely, this method is still widely used but is agreed to be effective only when used as part of a sequential process of test adaptation or as a supplement to forward translation (Geisinger, 1994; Hambleton & Patsula, 1999). Perhaps one main reason for its popularity is the ease of administration and also the lack of awareness about its limitations when used on its own.

6.2.2.3. Alternative method: bilingual judges approach

As an alternative or additional method to back translation, Hambleton (1993) and several others (Gesinger, 1994; Hambleton, 1994; Sireci, 2005) suggested the use of

bilingual judges/ translators to check the quality of the transaltion. This is executed by a direct comparison of the original and translated versions by two or more judges/ translators. Therefore, the bilinguals base their judgments on the parallel versions in both languages rather than only on the original and back translated versions, which are in the same language. Also, judging the quality of the translation is assessed directly using the translated version, so there is no risk of the back translator correcting for the forward translator's mistakes remaining unnoticed.

Limitations of bilingual judges approach

Hambleton (1993) also draws attention to pitfalls associated with this method when used on its own

1. The difficulty in finding judges/ translators that are equally proficient in both languages
2. Unlike examinees who will be taking the test later, bilinguals make use of insightfully guessing when trying to understand the meaning of a translated item because they know that it is a translated version and because they have access to the original version
3. Bilingual translators might not understand the item similarly to monolinguals who will be taking the translated test later in the future.

6.2.3 Rationale for the translation method adopted in this study

Despite its limitations, backward translation can detect some problems of inadequate translation (Brislin, 1986; Geisinger, 1994). However, the problems

associated with it are serious and will affect both the validity of the translated version and the equivalence between multilingual versions. Therefore, back translation should be used carefully and in conjunction with other methods. In this study, we adopted the traditional back translation method mainly because it provides an opportunity for a non-speaker of the target language to check the quality of the translation, although not fully and comprehensively (Hambleton, 2005). This is an important point for practice because specialists is test adaptation will only have limited repertoire of languages that they personally master, yet their test adaptation responsibilities will undeniably surpass this repertoire. However, back translation will only be used as one part of the adaptation process. Based on the information discussed earlier, we combined the forward and back translations with dyads or triads.

6.2.3.1. Definition of Dyads and Triads

Dyads and triads involve a group of two (dyad) or three (triad) bilingual judges facilitated by the Project Manager (see section 6.3.2 for details of their characteristics), and seeking profound information about each item in the questionnaire. Dyads and triads are designed primarily to investigate the linguistic equivalence between the items in the different language version; an essential part in reaching full metric equivalence between tests (van de Vijver, 1998).

As discussed earlier and in chapter 4, *linguistic, psychological* and *conceptual equivalence* are key issues to be achieved at item level in order to minimise biases. Dyads and triads are designed to tap onto all three levels. Linguistic equivalence operates on a "meaning" level, which comprises of two distinct categories: semantic and connotative

meaning. The former refers to the content of a word; the latter refers to the emotional association a word carries with it. Psychological equivalence focuses on the psychological effect that the same item produces in different languages. Conceptual equivalence on the other hand, focuses on the cultural relevance of item wording and content.

The PM ensures that these three types of equivalence are highlighted throughout the dyads/triads. However, the following study in chapter 7 employs quantitative analyses to assess whether the *linguistic, conceptual* and *psychological equivalence* have been achieved. That is done by pre-testing the questionnaires on a small sample to detect the reaction and performance of test takers to the item in different languages. Pre-testing is a good technique for reviewing these types of equivalence, nonetheless, further evidence of inequivalence on these level might come to light during the pilot ona larger number of participantss. This will be discussed further in chapters 7, 8 and 9.

6.2.3.2. Level of analysis: word, phrase and sentence

During dyads and triads, the translation of each item is critically evaluated semantically, connotatively, culturally, and psychologically. Each item is assessed according to how well it caries the meaning of the original item and also according to the emotional association it portrays in the target language. Items are scrutinised on several levels: word, phrase/clause, and sentence level. These three levels are necessary because, as discussed earlier, the context in not obvious in the case of personality testing. Therefore, one word is likely to have more effect on participants' responses because each sentence stands on its own and its meaning cannot be supported in a paragraph or by

relying on other sentences. The distinction between the three levels is as follows:

1. A word is any single unit of language that carries meaning and is presented in any form possible. For example "employ" is a word that could take several forms such as "employed", "employer", "employing" and "employment". An adverb is also considered a word. (Cambridge advanced learner's dictionary, 2007)
2. Phrase or clause, on the other hand, refers to any group of words that act as one unit in a sentence whether it has a subject (Clause) or not (Phrase). For example in the sentence "I think that an employer should not delegate very important tasks before consulting with his employees", "very important tasks" is a phrase that could be treated as a separate entity and "an employer should not delegate" is a clause.
3. The final level of monitoring translation is the sentence level, which is any group of words that contains a verb and is meaningful as a separate entity. This, in most cases, is the whole item or statement in the questionnaire. For example, "I think that an employer should not delegate very important tasks before consulting with his employees" is considered a sentence.

6.2.3.3. Advantages of the approach adopted

Combining back and forward translation with dyads and triads could balance out the limitations associated with each and add validity to the translation and monitoring phase. In summary, the potential limitations that may be overcome through using the combination of these techniques are:

1. The translated version is included in the judgement analysis. Therefore the

likelihood of overlooking a mistranslation or any other inadequacy at item level is reduced

2. At least five different native speakers of the target language, with various degrees of proficiency in both languages, take part in the translation phase. Therefore translator bias and the effect of differences in their proficiency in the OL are minimised
3. The cultural relevance of items, which might be overlooked in back translation, is key to all dyads/ triads because they rely on native speakers of the target language but who are also knowledgeable in the target culture. Therefore, linguistic, cultural, and psychological equivalences are highlighted during the process.

Moreover, an important part of this project is the simultaneous adaptation of Orpheus into several languages. Therefore, I acted as a project manager and participated in all dyads and triads. This adds two main benefits to the process:

1. The PM ensures the standardisation of the process across all languages
2. Test adaptation consists of translating independent items that belong to different scales. The PM uses his/her knowledge of each scale and the items that belong to it, to guide the thought process in the dyads and triads.

Paradoxically, it could be argued that the PM can bias the dyads and triads by asking leading questions and probing when unnecessary. This was controlled by running a trial dyad and a trial cognitive interview in the presence of the thesis supervisor to ensure that the PM is not manipulating responses unconsciously.

In summary, the translation and monitoring phase relies on the use of forward and back translation in combination with dyads and triads, monitored by a PM. The rationale

for this is to control the limitations of each of these methods when used alone by including judges who are native speakers of the target language and who can contribute to ensuring the cultural relevance of items. Additionally, the PM plays a major role in standardising the procedure across languages, facilitating the dyads and triads and ensuring that the problematic issues are being dealt with as objectively as possible. However, the PM as well could contaminate the results due to some personal biases, which can include the languages he or she masters, their understanding of measurement scales and so on. The role of the PM should be monitored and appropriate feedback should be given at the beginning of the different stages. An example of this type of monitoring will be highlighted in the next section.

6.3. Methods

6.3.1 Summary of study 1

The following glossary summarises the terminology used in this chapter.

Glossary	
Original English version	V1
Target Language version	V2
Target Language version Revised	V3
Back translated version to English	V4
Pilot Version in TL	V5

Study one is the first phase of the adaptation process: Test Translation and Monitoring. It comprises of four main steps:

1) Forward translation of the 190 items of the original version V1 of Orpheus into 3 target languages to produce V2 in each TL
2) Monitoring the quality and accuracy of translation of V2 in *dyads and triads* using V1 in order to produce V3 in TL
3) Back translation of V3 into the original language to produce V4 in the OL
4) Monitoring the quality and accuracy of the translation of V3 in dyads and triads based on V1 and V4 to produce V5

6.3.2 Participants

We recruited native speakers (n=10) from a sample of convenience; 2 Arab speakers, 4 Chinese and 4 Spanish. All participants had lived in their home country most of their lives and in the UK for at least 2 years and had a higher education from a UK institution, to ensure an adequate level of understanding of the English language. Each

dyad or triad included:

a) One person knowledgeable in psychology/ psychometrics,

b) Two native speakers of TL, also knowledgeable in OC and TC.

The PM was the person knowledgeable in psychology/psychometrics and also a native speaker of one of the TL (Arabic). Therefore dyads were employed when PM was also a speaker of the TL but otherwise triads were the standard. The PM was also the common anchor to all dyads and triads, ensuring the standardisation of the procedure across time and languages. The inclusion criteria in the dyads and triads are depicted in the following table 6.1.

Phase	**Sample characteristics**
1. Forward Translation	A NS of the TL (NS1), also knowledgeable in the OC-have lived or are still living in the UK
2. Dyad and Triad I	Two NS of the TL (NS2 and NS3), also knowledgeable in the OC- have lived or are still living in the UK + PM knowledgeable in Psychology/Psychometrics
3. Back translation	A NS of the TL (NS4), also knowledgeable in the OC-have lived or still living in the UK
4. Dyad and Triad II	Two NS of the TL (NS5 and NS6), also knowledgeable in the OC- have lived or still living in the UK +PM knowledgeable in Psychology/Psychometrics

Table 6.1: Criteria for inclusion in dyads and triads

6.3.3 Material

The Original English version of the 190 items work-based personality

questionnaire, Orpheus® (Rust, 1996)[4] and dictionaries in four languages: Arabic, Chinese, English and Spanish.

6.3.4 Procedure

The dictionaries were used to clarify any ambiguity in the meaning of words. Also, they were used to ensure that the PM understood the exact meaning of words under discussion especially when the words are in a language unfamiliar to the PM. The overall procedure could be summarised in five main stages and these are:

I. The 190 items of Orpheus were concurrently forward translated by an independent NS1 in each culture into either Arabic, Mandarin (Chinese), or Spanish to produce V2 (appendix 4, 5, and 6)

II. Some items in Orpheus seemed less straightforward than others and were put together in a list of *potentially problematic items* a priori in order to establish structure to the dyads and triads (appendix 7). The ambiguity of items was judged based on criteria informed by Kline's (1986) item writing guidelines for personality testing, and as well as Brislin's (1985, in van de Vijver & Hambleton, 1996) guidelines. Kline's guidelines are designed to guide test developers in writing good items and consequently increasing the reliability and validity of personality tests (Kline, 1986) whereas Brislin's guidelines are aimed at guiding item writing in a way that would ensure ease of translation (van de Vijver &

[4] Orpheus® is published by Harcourt Assessment.

Hambleton, 1996). Our aim was to flag items that could be problematic in translation due to their content. Therefore, we selected the most relevant from Kline and Brislin's guidelines to the adaptation of personality tests and adapted them into criteria that could be used to evaluate existing items retrospectively. As a result, appendix 7 was generated based on the following criteria:

1.Items that include idioms or colloquialism

2.Items that follow a complex grammatical structure

3.Items that include more than one term of frequency (i.e. a little, a lot) and preposition telling time (i.e. soon, often)

4.Items that cannot be understood the first time they are read by the PM

5.Items written in the passive voice

6.Items that use general rather than specific terms (i.e. member of family instead of mother)

Such problems in item writing are likely to confuse the test taker and contaminate his or her responses (Kline, 1986). If this is likely to happen in the English language, then we argue that will have a similar effect on the item translator and might therefore be problematic in the adaptation process.

III. The quality and accuracy of all the items in V2 were then discussed in a dyad or triad, the PM and two NS of the target language (NS2 and NS3). Dyads were only used in the case where the PM is a NS of the TL to make sure that at least two NS of the TL are present. Dyads and triads involve seeking profound information about each item independently. These were facilitated by the PM, who probed the

judges about specific components within the items, and took the form of a conversational back-translation of the 190 items of Orpheus.

As shown in figure (5.2, the Practical Framework of Test Adaptation), there are two sets of dyads/triads in the first phase of our suggested model of test adaptation. Each set refers to three dyads/ triads, one for each language and was conducted simultaneously across languages. The dyads/triads that followed the forward translation were conducted within the same month across all languages and each of them lasted for approximately 3 hours. Similarly, the dyads/triads that followed the back translation were also conducted within the same month for all languages. Judges were briefed (appendix 8) about the nature of the exercise, without being given any details about how it will actually run.

In this first set of dyads/triads, which followed the translation of the English versionV1 into V2, the material available was the original version V1 and the translated version V2. The PM presented each item in the TL separately to the judges and asked them to explain the meaning of the item in English. This is referred to as a simultaneous back-translation and is applied to avoid contaminating the judges' mind with the wording used in V2. The simultaneous back translation is aimed at flagging and discussing any discrepancies in item wording to identify its causes. The PM placed more weight on potentially problematic items (appendix 7), which were expected to suffer from translation problems.

Until that moment, only the PM had seen the original version, and by comparing it to V2 and the simultaneous back-translation, the PM probed if

1- Any of the wording simultaneously translated is different from the original version V1 (except for the use of synonyms)

2- The item was listed under the "potentially problematic items" developed a priori

3- The two judges disagreed on the simultaneous back translation

The probing occurred before presenting the judges with V1 to make sure their clarifications were based on their opinion and not influenced by V1 or V2. In some cases, the differences between V1 and the simultaneous back-translation were due to the fact that the translation was happening on the spot. These differences were clarified before the judges had access to V1 or V2.

The judges were then presented with the original item and were asked to rate the similarity between V1 and V2 on a 4 points Likert scale:

1- Not similar at all

2- Not very similar

3- Very similar

4- Exactly the same

This scale was developed to standardise the procedure according to which the judges were rating their opinion of the translation. The judges were asked to provide an alternative to any word, phrase/clause, or sentence if their rating was between 1 and 3. The first three response options indicate that the judges detected some differences in the multi lingual items; hence they did not rate it as exactly the same. In some instances, the similarity was rated 3, however, the judges did not have an alternative for the item due to language idiosyncrasies or because it

was difficult to come up with an alternative on the spot. In these cases, the reasoning was noted and the items were sent to the judges to think about it and try to come up with an alternative for it.

In the case of any *amendments,* these were noted for each item based on agreement between the two NS judges and the PM, indicating the part or parts of the items that needed changing and the reasons for subjecting these items for changes (appendix 9, 10 and 11 respectively for amendments in Arabic, Mandarin and Spanish). These changes constituted the first set of data for this study, which is reported in the results section below. The outcome of dyads/triads was a revised translation in the target language V3 in the target language (appendix 12, 13, 14 for Arabic, Mandarin and Spanish).

IV. The translated and revised version V3 was back-translated to English (V4) by an independent speaker of the Target Language (NS3).

V. The quality and accuracy of V3 were reassessed using V1 as well as V4 in a dyad/triad as described before, facilitated by the PM in the presence of two judges NS of the target language (NS5 and NS6). The materials available at this stage were: the original version V1, the translated version V3, and the back-translated version V4. This set of dyads/triads was run in a relatively similar fashion to the previous one.

The PM presented each item in the TL separately to the judges and asked them to simultaneously back-translate it to the original language. The PM

compared the simultaneous back-translation to the original version V1 and the back translated one V4 and probed if

1- Any word in the back translated version V4 was different from the original one V1 (except for synonyms)
2- The judge presented a simultaneous back translation different to V4 and/or V1
3- The item was listed under the "potentially problematic items"
4- The two judges disagree on the simultaneous back translation

Any of the issues 1 to 4 hinted to potential discrepancies between the versions and were therefore given special attention.

As discussed earlier, the probing occurred before presenting the judges with V1 and V3. The judges were then presented with the V1 and V3 and were asked to rate the similarity between them on the same Likert scale as before.

Again, the judges provided an alternative to any word, phrase/clause, or sentence if their rating is between 1 and 3. Amendments, based on agreement between the PM and the judges, were also noted for each item as well as the reasoning behind them. All the changed items in Arabic, Mandarin (Chinese), and Spanish are listed respectively in appendix 15, 16, and 17 with the reasoning behind them. These changes, in addition to the ones produced from the first set of dyads and triads, were the data that formed the basis of the analysis in this chapter. The outcome of the second set of dyads/triads is the amended version V5 in the target language (appendix 18, 19, and 20 for the Arabic, Mandarin and Spanish

versions). The resulting V5 was labelled *pre-test version* and served as the material for the second study presented in Phase Two of the Adaptation process.

Although the final product of this process (versions V5) is the main material for the second study in chapter 7, the dyads and triads generated a lot of data that fed into the Theoretical Framework of Test Adaptation. The reasoning behind the amendments of the items from the dyads/triads is the main material analysed qualitatively for the present study. It is also important to mention that practice dyads and triads were conducted in the presence of the PM's supervisor in order to give feedback on the performance and to monitor any potential personal biased that the PM bring to the procedure.

6.4. Review of Analysis Technique

6.4.1 Theoretical background of the analysis: Template Analysis

The reasons for each amendment were transcribed during the six dyads/triads (appendix 9, 10, 11, 15, 16 and 17) and formed the basis of this qualitative analysis. These reflect the nature of the challenges encountered in reaching linguistic as well as psychological and conceptual equivalence between multi lingual items. Thus the retrospective analysis of this data could help better identify these challenges.

Each amendment derives from a particular item and will be specific to it. Therefore, a standardised procedure should be employed in order to minimize subjectivity and to ensure that the most crucial ideas from the amendments all analysed. *Template analysis* is a qualitative method of analysing any form of textual data, including

interviews, personal correspondence, focus groups etc. (King, 2006). The result of template analysis is a *coding template or coding manual*, "which summarises themes identified by the researcher(s) as important in a data set, and organises them in a meaningful and useful manner" (King, 2006). A coding template condenses the text into *codes*, *broad codes* and *themes* hierarchically (Crabtree and Miller, 1999; King, 2006).

6.4.1.1. Units of analysis

Coding is a hierarchical process, which means that smaller order codes are more specific than higher order ones, referred to as broad codes. Codes, the most specific unit of analysis and sometimes referred to as sub-codes, are labels used to index relevant segments of the text (King, 2006) and are grouped together to form broad codes. Broad codes are sometimes referred to as broad themes and are more general than codes, but together with other broad codes constitute a theme. A theme is a "feature of participants' accounts characterising particular perceptions and/or experiences that the researcher sees as relevant to the research question" (King, 2006). In this analysis, *codes* are the third level of analysis followed by *broad codes* then *themes* in first level.

King, Thomas and Bell's (2003) study on carers' experiences of out of hours palliative care services provides a good example to illustrate the differences between higher and lower order coding. One of the themes that emerged from their study was "drugs and equipment out of hours". One of the general challenges that carers face when working out of hours related to drugs and equipment. The second level of coding is more specific and contained two broad themes "access to drugs" and "access to equipment". Therefore, the main issue with drugs and equipment was access to them. Finally, on the thirds more

engrained level of analysis, "access to equipment" contained four codes, two of which are "use of syringe drivers" and "equipment delivered out of hours" as illustrated in table 6.2 below. Some studies require more levels of coding but three to four levels are considered as manageable (King, 2006). The more levels there are, the more difficult it will be to add new data being analysed in the right level on analysis.

Level 1: Theme	**Level 2: Broad code**	**Level 3: codes**
1. Drugs and equipment out of hours	1.1 Access to equipment	1.1.1 Use of syringe drivers
		1.1.2 Equipment delivered out of hours

Table 6.2: Example of three levels of codes from King, Thomas and Bell's (2003) study

6.4.1.2. Exploratory and confirmatory approaches

In template analysis, themes could usually be designed a priori, but could also be left to emerge form the text (King, 2006). A priori themes usually emerge from previous research, a certain theory or the literature that the researchers can argue should be present in their data. From that respect, template analysis strikes a balance between Grounded theory (Glaser & Strauss, 1967), where themes are left to emerge, and Content Analysis (Weber, 1985) where themes are usually developed a priori. Developing themes a priori can be desirable as it makes the process of coding simpler (King, 2006). However, King recognises that a priori themes can contaminate coding and bias the analysis. On one hand, he argues that important parts of text that do not fit under any of the a priori themes may be overlooked. On the other hand, some a priori themes may not be the most appropriate even if it was previously thought they were.

A priori themes are suitably employed within the analytic strategy when the approach adopted is a confirmatory one. That is, if the researcher is looking to confirm the existence of certain themes, then these could be defined in advance and the data can be examined to prove or disprove their existence. However, a priori themes are less useful when the approach adopted is an exploratory one whereby the researcher is interested in exploring what the text is hiding. In general, the research question usually influences the template analysis approach. As a hypothetical example, researchers may wish to investigate the reasons behind a politician's proposal about a new health care system in the country, an exploratory approach might be more suitable in this case and themes could be left to emerge from the data.

6.4.1.3. Development of the coding template

As discussed earlier, the aim of analysing the reasoning behind the amendments that materialized from the dyads/triads is to uncover translation problems that may be encountered when adapting personality questionnaire into other languages and cultures. In this exploratory type of analysis, developing a priori themes could then constrain the breadth of the results. Another approach that could be adopted in template analysis is the development of preliminary codes after an initial exploration of the material (Crabtree and Miller, 1999). This is referred to as *initial template*, *initial coding template* or *initial coding manual* and is developed from a small sample of data to be analysed and then applied to the rest of the text and amended as necessary (Crabtree and Miller, 1999; King, 2006). The initial coding template facilitates coding because it makes it a systematic approach and codes could be added or removed from it depending on how well they

apply to the rest of the data. The advantage of this approach is the structure that it adds to the process of analysis. King (2006), however, argues that early initial templates might bias and limit the coder's coding technique.

Producing an initial template based on one transcript from the current study would mean developing one based on the amendments from one language only, and after either the forward or the back translation. This could lead to a premature template biased from two different angles:

1- The "reasons for amendments" are specific to one culture or language but not another.

2- The "reasons for amendments" for a version that has been translated by one person (V2) ought to be different from those arising from a version that passed through a translator and two other judges (V3).

Alternatively, a sample of amendments could be taken randomly from each data set (the 6 amendment documents appendix 9, 10, 11, 15, 16 and 17)) in order to produce an initial template that is more representative of the whole data.

6.5. Analysis

6.5.1 Development of Initial Coding template

Twelve amendments were randomly drawn from the dyads and triads, two from each of the six (appendix 21) and were used as the basis for developing the *initial coding template*. The initial coding template draws on samples of data from all the dyads and triads conducted in all the languages, thus creating a more comprehensive initial

template. Each amendment was given one or more codes (appendix 22) and as a result, an initial coding template was developed comprising of seven codes as shown in appendix 22.

The total number of items amended across all languages in the first set of dyads and triads was 222 (19 Arabic, 84 Chinese and 119 Spanish). Part of the discrepancy in the number of items changed may have been due to the translators' skills. This became evident when we closely reviewed the reasons behind the amendments, and found that nearly 55% of the item changes in the Spanish version were due to missing or wrongly added words. The Arabic amendments were much less perhaps because the PM is a native speaker of Arabic. However, this might also be attributed to the translators' skills. The 222 items changed constitute 40% of the 570 items translated across languages (190 items x 3 languages). After the second set of dyads and triads, this number was down to 87 (17 Arabic, 37 Chinese and 33 Spanish), constituting 15% of the total number of items translated across languages.

6.5.2 Development of Final coding template

The approach adopted is bottom up whereby specific codes were developed first, followed by broad codes then themes. Typically, a top down approach is used in template analysis (King, 2006) but due to the nature of the data, a bottom up approach seemed more suitable in this case. The data in this study has a very specific content, so the initial coding is based on data that is very detailed. For example, the amendment "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." This amendment has two issues that can be coded, both of which are *wrong meaning*. However, wrong translation is a very specific code that can be grouped with other codes, such as *grammatical mistake* to form a more general broad theme *lingo-syntactic mistake* (appendix 24).

The initial coding template was applied to the rest of the transcripts as a guide, but more codes were added if the initial ones did not encompass certain segments of the transcripts. That is, each amendment was coded first using one or more of the initial codes from appendix 22, or given new ones when it could not be ascribed to any of the initial codes. After coding each transcript separately, the new codes were crosschecked with the initial ones to avoid redundancies. For example the codes "Better Sentence Structure" and "Clumsy Sentence Structure" were collapsed together to form one code because they could not be distinguished from each other. This process continued until reaching the point of saturation; that is the minimum number of codes that could be generated from parts of the data that most relevantly contributed to answering the research question. At the end of the analysis, eighteen codes were needed to fully capture the content of the data. These are listed in appendix 22.

Each of the eighteen codes was written on small cards then grouped with other ones according to broader codes that put a more specific meaning into their emergence. The same process applied to the development of themes. The development of broad codes and themes are discussed separately in the two following sections.

6.5.3 Development of Broad Codes

The 18 codes above were grouped further into 12 broad codes. Each broad code combined codes that represented a similar challenge to translating personality tests. For example, the codes "Wrongly Omitted Word", "Wrongly Added Word" and "composed words in TL" were grouped together to form the broad code "wrongly omitted or added word" because they are similar types of translation challenges. For example, words that are composed of several others in the TL require additional wording that is not present in the original version. Similarly, a wrongly omitted or added word has a comparable effect on the equivalence between languages than that of a wrongly translated composed word. Therefore, the addition or omission of words from the item is the broad theme that demonstrates the similarity between these codes. Some codes that did not combine with others to form broad codes were moved up one level, leaving some broad codes without codes under them.

6.5.4 Development of Themes

Building on the example above, the broad codes "wrongly omitted or added word" and "sentence, phrase/clause, or words grammatically inequivalent" are grouped together under the same theme "accuracy of translation" because both affect the precision of the translation. Three main themes emerged from the data analysis and these are:

1- Accuracy of Translation
2- Language Idiosyncrasies
3- Connotative Meaning

The final template consisted of three levels, the first of which is themes, the second level

broad codes and the third level codes. The number of broad codes and codes were not equal across themes, a common phenomenon in template analysis (King, 2006). Also, some themes had two levels only (themes and broad codes) whereby others had three levels (themes, broad codes and codes). Table 6.3 below illustrates the final template, which consisted of 3 themes, 12 broad codes, and 18 codes

Theme	Broad Code	Code
1. Accuracy of translation		
	1. Literal translation more appropriate	
	2. Sentence, phrase/clause, or word grammatically inequivalent	1. Sentence Grammatically non-equivalent
		2. Word(s) Grammatically Nonequivalent
	3: Wrongly omitted or added word	1. Omitted Word(s) 2. Composed Words in TL 3. Wrongly Added Word
	4: lingo-syntactic mistake	1. Wrong Meaning 2. Grammatical Mistake
2. Language idiosyncrasies	1.Context Dependent Synonym	
	2. Sentence Formulation	1 Better Wording or Structure 2. Unnatural or informal wording
	3. Words Nonexistant in TL	
	4. Idiosyncratic omissions or additions	1. Elaboration 2. Shrinking
	5. Idioms	
3. Connotative meaning	1. Leading Literal Translation 2. Different Magnitude 3.Literal Translation not most appropriate	

Table 0.1: *Final Coding Template: Summary of all Codes, Broad Codes, and Themes from the dyads/triads study*

6.5.5 Inter-rate reliability and independent scrutiny of analysis

Independent scrutiny of analysis is usually employed to assess the quality of the coding in template analysis (King, 2004). This entails getting another person to code the same data to check whether there is consistency in coding. This could be coupled with statistical analyses to provide a more objective measure of agreement between raters. This is referred to as inter-rater (or inter-coder) reliability, which is usually applied to assess the degree of objectivity by measuring the rate of agreement between two independent raters (Rust & Golombok, 1999).

Cohen's Kappa was used as the coefficient of reliability because not only does it compute the degree of agreement between the raters, but also the degree of agreement by chance. The first rater developed the codes, broad codes, and themes with examples and description and the second rater received the final coding template and a brief detailing the coding procedure (appendix 23). The coding was based on the themes and included 312 units of coding in total. Cohen's Kappa coefficient was calculated as follow:

$$kappa = \frac{\sum a - \sum ef}{N - \sum ef}$$

whereby $\sum a$ is the sum of agreement between rater 1 and rater 2 and is computed by adding the diagonal cells; $\sum ef$ is the expected frequency by chance and is calculated as follow: $ef = \frac{rowtotal * coltotal}{overalltotal}$; and N is the total number of codes. Kappa was 0.50, which reflects moderate agreement between the two raters (Landis & Koch, 1977).

6.6. Results

The findings presented in this section are based on the analysis of the reasons behind amending items across three languages simultaneously. We opted for analysing the data across three languages together in order to generate a list of item level translation problems that could be applied to several languages. Furthermore, the types of translation problems encountered across the three languages were similar so it was possible to group them together while respecting the particularities of each language. Three main themes emerged from the analysis and are outlined below but are discussed in the section 6.7.

6.6.1 Themes description

6.6.1.1. Theme 1: Accuracy of Translation

Accuracy of translation is the most basic and straightforward item level source of inequivalence between multi-lingual versions of tests. Accuracy of translation symbolizes the aspects that hinder the reproduction of equivalence in *meaning*, especially in cases where the meaning can be reproduced correctly. This theme is a representation of the challenges to linguistic equivalence, which, as defined in chapter 4, encompasses the reproduction of the literal as well as the connotative meaning of the sentence. Linguistic equivalence revolves around the wording of sentences and its effect on the equivalence in meaning of multi-lingual versions of items. Problems such as wrongly omitting or adding words, using adjectives instead of nouns, and committing grammatical mistakes are examples of challenges that affect the equivalence between sentences at the linguistic level. This is one of the lower order levels of equivalence discussed in chapter 4, that are

prerequisites for the higher order levels: construct, measurement unit, and scalar equivalence.

Accuracy of translation contains broad codes that refer to linguistic inequivalence in the meaning of a word, phrase/clause, or sentence level. The codes and broad codes that make up this theme are grouped on the basis that they all affect the linguistic equivalence of sentences; that is the basic meaning. For example, in the item "I find clerical work somewhat tedious", the word "tedious" was translated into "aburrida" which means boring in Spanish (appendix 6). Although boring and tedious are close in meaning, the word "tediosa" can be used in Spanish and actually means tedious. The summary of theme 1 is presented in appendix 24 with its constituents and examples to illustrate them. The implications of these findings and the ones that follow are explored further in the discussion section.

6.6.1.2. Theme 2: Language Idiosyncrasies

Language idiosyncrasies represent translation problems that are due to characteristics directly related to particularities of the target language. For example, the broad code "Context Dependent Synonyms" represents words in the target language that have several synonyms that are used in different contexts. The word "写作风格" and "工作作风" both mean style in Chinese but the first one refers to the style of writing whereas the latter refers to the style of working (appendix 5). Using one or the other does not affect the meaning of the sentence but the correctness of saying or writing it in the target language. Similarly, the broad theme "sentence formulation" relates to the style of

writing sentences, which is another language idiosyncrasy that affects the cultural relevance of the item to the target language.

Language idiosyncrasies dictate the way sentences are written in a specific language. A sentence translated into another language could be capturing the meaning of the original one while sounding artificial or translated. Such sentences satisfy the linguistic equivalence, but at the expense of cultural equivalence due to stylistic reasons. As a hypothetical example, the English sentence "in the journal of Psychology, professor Smith points out that..." can be translated easily to Arabic and the meaning will be intact even if the sentence followed the exact structure as the English version. However, it is more natural in Arabic to day "Professor Smith points out in the journal of Psychology that...". Changing the structure of the sentence will make it equivalence on a linguistic as well as on a cultural or conceptual level.

These broad codes grouped under the umbrella of language idiosyncrasies all affect the conceptual equivalence between items, that is, their cultural relevance in each language. This is detailed in appendix 25.

6.6.1.3. Theme 3: Connotative Meaning

Connotative meaning relates to words or sentences that produce a difference in the psychological effect of the item in both cultures. For example, the broad code "Different Magnitude" refers to the use of words that have differential psychological effect in different languages. As an example from this data, the Spanish words "nunca" and "jamas" both mean never but the latter has a stronger connotation, maybe closer to "never ever" (appendix 6). Having one or the other in a sentence will undeniably create a different psychological feeling and reaction to the item. A participant could Strongly Agree to the item " I never do anything without good reason" if this is a common thing they do. However, they will not necessarily Strongly Agree with the item "I never ever do anything without good reason" because "never ever" is stronger and more definite.

Similarly, the broad code "leading literal translation" implies that the wording in the target language diverges test takers' thinking, thus creating a psychological impact different from the impact of the original version. The word "manipulate" in English means influence but either smartly or by unfair means (Cambridge online dictionary, 2007). In Mandarin, 操纵，影　means manipulate unfairly whereas 操作, 使用 means manipulate smartly or handle or use properly; there is no word that implies both meanings. Therefore, the connotative meaning that the words hold may result in a different psychological effect in participants in the UK vs. those in China. The theme *connotative meaning*, the broad codes (leading literal translation, differential magnitude, and literal translation not most appropriate) are summarised in appendix 26 with examples to illustrate them.

6.7. Discussion

The aim of this study is to pinpoint the item level challenges encountered in translating personality tests into other languages and cultures. The findings showed that there are three main categories that translation problems arise from and these are: accuracy of translation, language idiosyncrasies, and connotative meaning.

6.7.1 Identification of three sources of item bias

Three types of bias were identified in the previous chapter 4 and these are: construct, method, and item bias. Each one was associated with several sources based on discussions in the literature. The purpose of this study is to empirically identify the sources of item bias. The translation and monitoring phase is carried out on the item level where *item bias* can be found. As a result, three sources of item bias were identified: accuracy of translation, language idiosyncrasies and connotative meaning. However, each one of these sources can be understood in detail using the broad codes that fall under it. For example, accuracy of translation is one source of item bias that could take the form of grammatical inequivalence, wrongly omitted/added words, using words that are not the literal equivalent in the target language. Similarly, item bias could arise from language idiosyncrasies, which could be due to using words in the wrong context or formulating a sentence in a non-idiosyncratic way.

6.7.2 Effect of these sources of bias on responding

These sources of item bias affect candidates' responses from two angles:

through face validity or psychologically. The broad themes "context dependent wording" and "structure of sentences", for example, do not directly affect the meaning of the sentence. That is, the item will be understood correctly when it is read in either language. As argued earlier, however, these broad themes do not affect the linguistic equivalence between items but the conceptual equivalence between them. Therefore the item in the target language may sound as if it has been translated and therefore affect the way it is viewed by test takers. Filling out a questionnaire that contains items that are not idiosyncratic to the target culture undeniably diminishes respondents' faith in the test. The test taker may question whether the test developers, who were not able to cross the language barrier accurately, will be able to cross the cultural barrier and provide an adequate assessment of the test taker's personality.

On the other hand, broad themes such as differential magnitude of words, wrong translations, and wrong additions/omission of words will create a difference in the psychological impact of the item in either culture. For example, the item "I sometimes prefer being at work to being at home" is not the same as the item "I prefer being at work to being at home" though the only difference between them is the omission of the adverb "sometimes" in the second item. The psychological reaction the same person can have towards the first item can be very different to the one they could have to the second item. In the same way, the connotative meaning a word carries with it has a strong psychological effect on participants. For example, the word "discouraged" can be literally translated in Arabic to " محبط" (pronounced as "mouhbat" in Arabic, see appendix4). However, since the concept of "depression" has only recently started to be applied in the Arab world, the same word used to refer to discouraged" is also

used to refer to "depressed". Therefore the item "I am never discouraged by failure" has a different connotation in Arabic when the literal equivalent is used, which creates a different psychological reaction from participants towards it in English than in Arabic. During one of the dyads, discouraged was agreed to be replaced by "my determination decreases" in order to maintain the same magnitude the item has in both languages.

6.7.3 Implications for future research and practice

As discussed earlier, dyads and triads are a useful addition to the basic process of back translation. In deed both the methodology and the findings of this study could help for further improve the process to be more structured and standardised. For example, the broad codes could be used as a guide during subsequent dyads and triads to monitor the quality of the translation. These could be useful for probing, guiding and training the judges, who might not have insight into the importance of maintaining the linguistic, psychological and conceptual equivalence between items. Moreover, these could also be useful for the PM especially when they do not speak the target language. The findings can be used to assist the PM monitor the personal biases that each individual judge might be holding, by limiting the amendments to the broad codes, unless the judges can give a reasonable argument for amendments that are not covered in this list. As for the dyads and triads, the number of judges included in this process depends on the test being adapted. In the case of personality testing, it is important, as discussed earlier, to have at least two native speakers of the language and to have a person knowledgeable in psychometrics and the scales being assessed. Although it is important to have people who are fluent in both

target and native language, it is not necessary to have professional translators because a) there are several people who take part at different stages of the process (i.e. first and second sets of dyads and triads) and b) the translation and back translation is done by qualified translators. Moreover, a larger number of participants in each dyad and triads might be counter productive because the more opinions there are, the more difficult it will be to reach an agreement. Additionally, there are at least seven different people examining the adapted test before it goes into pre-testing.

The broad codes could also be used earlier in the process as a brief for the forward and back translators. This can help increase the trans-cultural accuracy of the translation and provide them with a clear *purpose* for the translation, as the functionalist approach suggests (Nord, 1997). By doing so, the purpose of the translation is focused from early stages, and the forward and back translators as well as the judges and the PM will all be working towards the same goal (achieving linguistic, psychological and conceptual equivalence) based on common understanding of how it could be achieved.

6.7.4 Conclusion

In summary, when reproducing an equivalent translation the three basic levels of equivalence are: linguistic, conceptual and psychological. The achievement of these three can be monitored and standardise using dyads and triads. These meetings could be designed in a way that emphasizes the achievement of linguistic, conceptual and psychological equivalence. However, it is important to acknowledge that some items developed in one language may

be impossible to translate effectively into another while retaining the same factor structure. The adaptation process is aimed at achieving equivalence as closely as possible, though this might not always be feasible. Perhaps another way of creating trans-cultural/linguistic questionnaires is by simultaneous item reduction across samples. The items that do not achieve equivalence can be dropped out from all the versions, making the questionnaire shorter. Although this might lower the reliability of the test, it might increase the validity of it across cultures. This study has begun to address issues with adapting personality tests and the subsequent chapters will explore this framework in more details.

6.8. Study Strength and Limitations

6.8.1 Strengths

The strength of this study lie in two areas: 1) the combining of several qualitative techniques and judges to form a tool for monitoring the quality of the translation and 2) the further analysis conducted on the data.

Firstly, adding the dyads and triads to the back-translation technique resulted in an improved method for monitoring translation. As discussed in the introduction of this chapter, the limitations of back translation technique are: 1) not including the translated version in the process, 2) discrepancy in forward and back translator skills, and 3) focus on *linguistic equivalence* only. The dyads and triads revolve around the translated version using the knowledge and expertise of several judges rather than relying on only two. Additionally, as the findings revealed, dyads and triads focus on the linguistic equivalence as well as the psychological and conceptual equivalence leading to a more comprehensive

control of the quality of translation even from early stages of the adaptation.

Secondly, the qualitative analysis approved to capture some of the sources of item bias. Several types of bias have been identified in the literature as well as their sources. Poor translation and connotations of words are listed as two sources of item bias (van de Vijver & Poortinga, 2005). The sources of item bias found in this study (accuracy of translation, language idiosyncrasies and connotative meaning) empirically confirm the existence of these sources of item bias, and even add one more to the list (language idiosyncrasies). The findings add valuable empirical knowledge and structure to the Theory of Equivalence and Bias proposed by van de Vijver and Leung in 1997.

6.8.2 Limitations

It has been recognised however that this study may have several limitations. Although the combination of techniques resulted in a more accurate way of monitoring the translation process, this does not guarantee a production of an equivalent version in other languages. This stage is only one part of the adaptation process and cannot substitute the rest of the procedure. It is aimed at monitoring the translation closely in order to minimise any linguistic, psychological and conceptual equivalence but further studies are necessary to reach equivalence on all levels.

Another limitation of this study is the use of four languages only in order to pinpoint the sources of item bias. Although the four languages used in this study are not representative of all the world's languages around the world, they represent Indo-European (romance and Germanic), Sino-Tibetan, and Semitic language families. It is also important to highlight the fact that the findings were

based on data from all the language families together, and therefore are not specific to any one language due to their generic content and could be applied to other languages. Further research replicating this study with another instrument and perhaps other languages is crucial for validating the themes and generalising the findings to other languages and cultures.

Chapter 7:Item pre-testing and cognitive Interviewing – Piloting the multi-lingual versions in the target cultures using small sample size

"Colourless green ideas sleep furiously"

Noam Chomsky

7.1. Chapter Overview

This chapter builds on the previous one to outline the methodology employed in the second phase of the adaptation process, referred to as *Pre-Test and Cognitive Interviews*. The three multilingual versions of Orpheus were pre-tested on a small sample in every culture and the results were analysed using cognitive interviewing technique with native speakers of the target languages. Therefore this phase first assessed the quality of the translation from stage 1 quantitatively then qualitatively to elicit participants' view on problematic items. In the first part of the chapter, we discuss the literature supporting these methods. We first focus on pre-testing, how it links strongly to cognitive interviews, and its relation to the next phase of the process, piloting. Subsequently we examine cognitive interviewing as a technique, which, although popular in the social sciences in the adaptation of survey methods (Willis, 2005) have been used less so in test adaptation taking a psychometric perspective. Cognitive interviewing provides invaluable information that increases the accuracy of the interpretation of empirical results. This is achieved through the investigation of the thinking process of participants to facilitate the understanding of statistical data. Finally, we present the research design and findings as two studies, study 2 for the pre-testing and study 3 for the cognitive interviewing.

Together, phases one (*Translation and monitoring*) and two (*Pre-testing and cognitive interviewing*) of the process, outlined in chapter 6 and the present chapter, form the Quality Control Process for the translation in each language before piloting. In addition to providing an empirical angle to the quality control

process, this chapter contributes to the literature on test adaptation by evaluating the combination of t-tests and cognitive interviews as a filtering technique for flagging items as potentially inequivalent. This will be achieved by comparing the number of items flagged as statistically different in both languages to the items that were agreed to be different after the cognitive interview.

7.2. Introduction

7.2.1 Rationale for qualitative quality control check

It is essential to have a robust *Quality Control Process* that minimises the potential problems, such as translation errors, particular to test adaptation. In other instances, the problem might be related to cultural biases or curricular differences (Grisay, 2003). These problems are more difficult to address by back-translation or even experienced panel of judges since they relate to the item content rather than the linguistic equivalence between items. Hambleton and Patsula (1999) argue that field-testing can help detect problems with adaptation that back translation and translators will fail to identify. Statistical techniques can detect differences in performance between groups that test developers cannot anticipate. This is best illustrated in an example from the field testing of the ITC guidelines based on the adaptation of the National Assessment Educational Progress (NAEP) maths exam items into Chinese (Hambleton, Yu, & Slater, 1999, p274). One of the items translated to Chinese was based on the mathematical concept of estimation and presented the students with pictures of a plane for $4.99, glue for $1.29 and paint for $2.19 and asked students: "Chen had $10 to buy a model plane, glue, and paint as shown above. At which

of the following times could an estimate have been used instead of exact numbers?

A. When Chen tried to decide whether or not he had enough money to buy the place, glue, and paint.
B. When the clerk entered each amount into the cash register.
C. When the clerk told Chen how much he owed.
D. When Chen counted his change."

Translators judged the item to be well translated and did not anticipate any problems with it. However, when the item was field tested, the analysis showed that American students outperformed their Chinese counterparts on it. More in-depth investigations revealed that "estimating when planning to make a purchase is not a habit of Chinese students" (Hambleton, Yu, & Slater, 1999; p274) rendering them less likely to identify the correct answer (A).

7.2.2 Pre-testing

Although Hambleton and Patsula (1999) rightly suggested that field-testing will expose measurement problems that could not have been identified qualitatively, field-testing involves collecting data from a large number of participants, which is a challenging and time-consuming process. It is therefore important to have flagged and corrected as many faulty items as possible before reaching this stage. Plus, the smaller the number of potentially flawed items the greater is the chance of overall equivalence between multilingual versions (Grisay, 2003).

Pre-testing provides a good solution to this as it involves administering the test to a small group of individuals then interviewing them to monitor the clarity of instructions and the quality of the overall translation (Geisinger, 1994;

Hambleton & Patsula, 1999; Sireci, 1999). Pre-testing provides a cost and time effective solution for revising any translations as well as empirical evidence to support or elicit revisions (Hambleton & Patsula, 1999; Sireci, 1999). Arguably, the length of the test could render this process time consuming.

There are several methodological approaches that can be applied to collect pre-test data (Sireci, 2005). For example, the two language versions of the test can be given to the same group of bilingual native speakers of the target language. Statistical comparison is then conducted to monitor whether the same item behaves differently in when presented in two different languages. This way the scores of the same participants can be compared on both tests, thus controlling for individual differences that usually exist between groups. The most obvious challenge to this approach is that bilinguals are not necessarily equally proficient in both languages (Sireci, 2005). That is, if bilinguals respond differently to the same item in different languages, this might be due to faulty adaptation but also due to inequality in language proficiency. Another fault associated with this approach is a representation problem because bilinguals are likely to be very different from monolinguals (Sireci, 2005) from the same culture in terms of socio-economic status, education, etc. Additionally, fatigue and practice effects are common drawbacks in such within subject designs. Alternatively, a two-group design can be implemented whereby two randomly equivalent groups of bilinguals take a different language version (Sireci, 2005). Although the problem of representation is not solved, this design is more time effective because the 2 versions can be administered at the same time. Fatigue and practice effects are also eliminated and random equivalence between the groups should reduce group differences (Sireci, 2005). This has informed the

design adopted here as described and discussed further in the methods section below.

Pre-testing increases the likelihood of reaching linguistic equivalence during the pilot and the field testing stages of the adaptation process. In a paper presenting the procedures implemented in the PISA project (Programme for International Student Assessment), Grisay (2003) argues that after pre-testing and getting rid of flawed items, even if further few flawed items were detected, "they will be unlikely to affect the overall estimate of a country's mean in any significant way" (p 2). That is, pre-testing minimizes the number of items that are likely to cause inequivalence and only a small number might escape this procedure, which is not likely to have a drastic effect on the final overall characteristics of the test. Therefore pre-testing is typically used to loose malfunctioning items, which implies that test developers have a large item bank and can loose items without affecting the reliability of the test. However, when adapting an existing test into another language, it is important to preserve as many items as possible so that reliability and validity are not jeopardised. Some items will undeniably not function well across all the target languages, as they might be too culturally dependent for example. However, malfunctioning of items during the pre-test could be attributed to translation problems or to using small sample sizes, which do not control for individual differences between the groups.

Statistical tests can detect difference between multi-lingual versions of items within one culture but they fall short in that they cannot provide an interpretation for it. Statistically significant differences for a small sample size are only meaningful when items are scrutinized qualitatively to uncover the

source and reasons for the differences in responding. One way doing so is by investigating the thinking process that test takers go through when faced with the items in either language. Therefore we will use pre-testing to flag items as potentially malfunctioning, either wrongly or correctly, then analysing these in standardised *cognitive interviews* to determine whether the difference is caused by rectifiable linguistic, psychological, or cultural discrepancies or indeed other factors. Accordingly, items can be revised but dropping of malfunctioning items will be left for the pilot study, where a larger sample is involved and less individual differences effects are present. Cognitive interviews will be discussed in the next section and their outcome in the results and discussion later in this chapter.

Reiterating from a previous point, pre-testing can help flag some linguistic problems but also some others that are not easily detectable by judges and translators. Using an example from Orpheus to illustrate, in the item "on some occasions I have behaved very improperly", "on some occasions" could mean "a significant event or happening" but could also mean "sometimes". In the English version, it is understood from the sentence that "on some occasions" refers to "sometimes". However, the equivalent word in Arabic "في بعض المناسبات" hints more to "a specific event" though it could also mean "sometimes". There is a big difference between behaving very improperly at an event compared to in one's own time. Yet, the distinction is not easily detected by translators and judges because it is a somehow correct translation that is hard to be picked up. However, respondents in the pre-test might respond differently to this item in the different languages therefore facilitating the detection of such linguistic problems.

Finally, one of the challenges for using a panel of judges discussed in the previous chapter points that the judges are *different* from test takers that the test is designed for. Hambleton, (1993) argues that "bilingual translators do not necessarily think about test items the same way that unilingual might" (p 9). Bilingual judges might approach items with a critical view due to their extensive knowledge and focus on linguistic equivalence. This might dissociate the item from the psychological impact that it can produce when it is encountered as part of an assessment for the first time. Pre-testing offers the opportunity to scrutinise the test from the test takers' perspective and balancing the differences between bilingual judges and test takers.

7.2.3 Rational for cognitive interviewing

Cognitive interviewing has been applied as a method for detecting errors in questionnaires and increasing their quality for the past 15 years (Redline, Smiley, Lee, DeMaio, & Dillman, 1998; Rothgeb, Willis, & Forsyth, 2001; Willis, 1999; Snijkers, 2003). Typically in cognitive interviews, participants are asked to describe verbally the thought process they go through while answering each item in the questionnaire (DeMaio, Rothgeb & Hess, 1998; Redline, Smiley, Lee, DeMaio, & Dillman, 1998). Cognitive interviewing is thus directed to studying the way *specific* audiences understand, process, and respond to questions (Willis, 2005) making it possible to examine the effect of the item on respondents. It is important to highlight that cognitive interviewing relies on participants who resemble the specific audience for which the test was developed in order to incorporate both experts' knowledge as well as participants' perspective into questionnaire development.

7.2.4 The cognitive interviewing technique

Cognitive interviewers probe test takers to reveal clues about their thinking process in order to identify "sources of miscommunication between survey designer and respondent before a survey instrument in fielded" (Hughes & DeMaio, 2002, p 1535). The idea is that understanding the test takers' thinking process when attempting to answer an item makes it possible for test developers to detect problems at item level. However, cognitive interviewing has been mainly associated with test development rather than test adaptation (DeMaio, Rothgeb, & Hess, 1998; Rothgeb, Willis, & Forsyth, 2001; Hughes& DeMaio, 2002). We will now discuss the different types of cognitive interviews and the way they function then highlight how they could be implemented to support statistical item analysis in the pilot study.

7.2.4.1. Types of cognitive interviews

Generally, there are two main types of cognitive interviewing: think aloud technique and verbal probing technique (DeMaio, Rothgeb, & Hess, 1998; Willis, 1999; 2005). However, another distinction can be made between concurrent and retrospective approaches. Therefore, there are four possible types of cognitive interviews, think aloud concurrent, think aloud retrospective, verbal probing concurrent and verbal probing retrospective.

Think aloud technique

In think aloud technique, participants are instructed to think aloud while attempting to answer questions. The interviewer's role centres on encouraging

the participants to say what they are thinking with minimal interference (Willis, 2005). Using this technique assumes very little interviewer bias and the information revealed could be extremely valuable if the participants are outgoing and articulate (Willis, 1999). However, this technique could be impractical in that participants need training to learn how to think aloud and to make sure they stick to relevant information instead of extrapolating to other topics (Willis, 1999).

Verbal probing technique

Alternatively, the verbal probing technique is based on an interaction between the interviewer and the interviewee whereby the former asks questions that the latter answers to. The interviewer probes based on predefined or semi structured set of questions that can help the interviewees verbalise their mental processes (Willis, 1999, 2005). An obvious advantage of using this method is that there is no need for training participants, which could be a very difficult task. Instead, the probes are designed and structured in a way that facilitates the retrieval of information from the participants. Additionally, in think aloud technique, interviewers do not have much involvement in the process, which gives the interviewee control over the amount of information they wish to reveal (DeMaio, Rothgeb, & Hess, 1998). This is problematic in situations where the interviewee is not very open in responding, and answers questions with very minimal information. With verbal probing, the interviewers can probe spontaneously if they felt they did not get the full information about the interviewees' thinking process (Willis, 1999). Although probing can be very useful, it could also be a source of bias if the interviewer unintentionally leads

the respondent. This, however, could be controlled by carefully designing the interview protocol with non-leading questions to standardise the cognitive interview and minimise interviewer bias (Willis, 1999; 2005). The cognitive interviewing method most commonly applied in practice is a combination of both techniques, whereby interviewers ask open-ended questions that can encourage to interviewee to think aloud while probing when necessary (DeMaio, Rothgeb, & Hess, 1998).

Concurrent and retrospective

Another distinction to be made is between the approaches to probing: concurrent and retrospective (DeMaio, Rothgeb, & Hess, 1998; Willis, 1999, 2005). Think aloud, verbal probing, or the mixed approach can be conducted either concurrently or retrospectively. When run concurrently, the cognitive interview takes place while participants encounter the questions for the first time. So the information is captured right when it is available for the respondent. However, there is a risk of contaminating the following questions because test takers become too focused on their thought process and might not answer the questions as they would normally, that is without having to think about them so deeply (DeMaio, Rothgeb, & Hess, 1998). When conducted retrospectively, the participants take the test first and then are cognitively interviewed about the thinking processes that lead them to give the answers they did. Although there is no contamination of participants' responses in this case, it might be difficult for them to accurately remember the thought process they went through that lead them to answer the questions as they did (DeMaio, Rothgeb, & Hess, 1998).

7.2.4.2. Approach adopted in this study

For the purposes of this study, cognitive interviewing could be a useful tool for analysing the potential reasons for significant differences between items presented in different languages, through understanding the thought process that participants go through while attempting to answer the questions. To clarify, when the same item is presented in the same culture in two different languages, similarity in responding should indicate a strong resemblance between the multi-lingual versions of the item. However, as highlighted earlier, discrepancies in performance could be the results of:

1) Linguistic, cultural, or psychological inequivlance on item level which could be rectified, which are the reasons why this method was implemented in the first place
2) Linguistic, cultural or psychological inequivalence, which cannot be rectified and would potentially be dropped from the item if they continue to cause nuisance in the pilot and field testing phase, and
3) Individual difference due to the small sample size in the pre-testing phase.

By understanding how participants think about the item in each language, in-depth cognitive interviews could therefore provide the possibility of distinguishing which of the preceding three points might have caused the statistically significant difference.

In conclusion, the combination of the Pre-testing and Cognitive Interviewing could create a robust quality control process due to the amalgamation of qualitative and quantitative techniques and also the reliance on several test takers as well as other native speakers of the TL. Figure 7.1 below

illustrates the interaction between the different parties involved in the process of adaptation in order to maximize the likelihood of reaching equivalence.

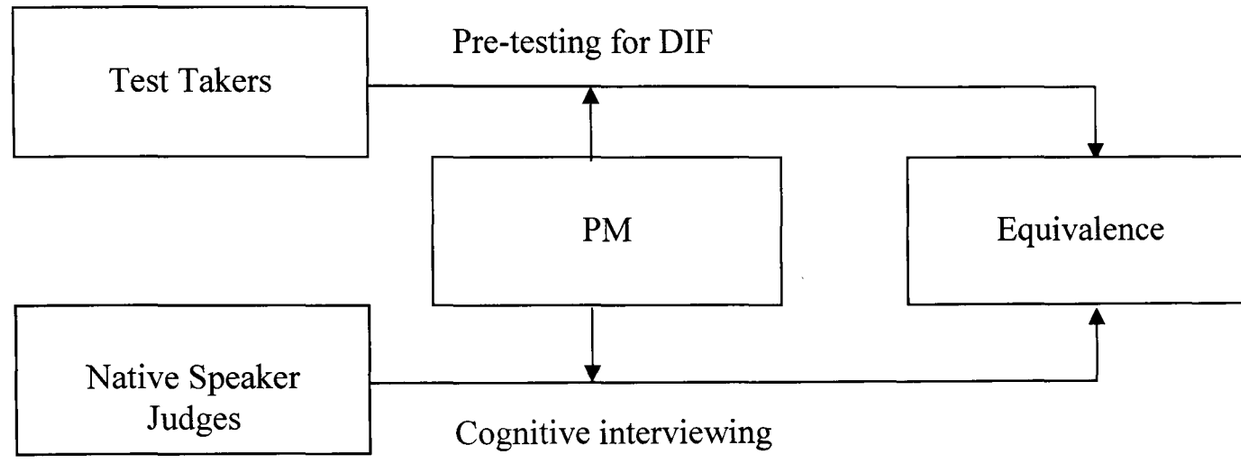


Figure 7.1: interactions during the adaptation process

7.3. Methods

7.3.1 Summary of study 2 and 3

The following glossary summarises the terminology used in this chapter.

Glossary

Original English version	V1
Pre-testing Version in 3 TL	V5
Piloting version in 3 TL	V6
Native Speakers of OL	NS

The methods and results sections in this chapter are divided into two parts: Pre-testing (study 2) and Cognitive interviewing (study 3). We will present these sections consecutively then discuss the findings of both studies in one discussion section because the findings support each other. Study 2, the quantitative part of the Quality Control Process, comprises of two main steps:

1- The original version V1 and the pre-test version V5 of Orpheus are administered to approximately 60 participants in each of the three target cultures. Half of them fill out the V1 and the other half fill out V5
2- Each item from V1 is tested for statistically significant differences with its parallel from V5 using t-test.

Study 3 follows up on Study 2 and consists of 2 steps:

3- Items that are considered to behave significantly differently in the same culture are reassessed qualitatively in cognitive interviews with a NS1 and NS2 separately
4- Changes from the cognitive interview are agreed in a panel discussion with NS3 and NS4 to agree on a final field-testing version V6 to be tested ready on 200 participants in each culture in the third phase of the adaptation process.

7.3.2 Study 2: Pre-Testing

7.3.2.1. Design Study 2

As discussed earlier, a single groups design, which involves giving the two versions of the test to the same group of bilinguals has its advantages (controlling for individual differences) and disadvantages (practice effects) (Sireci, 2005). In order to control for practice effects, counterbalancing is usually employed, whereby half the sample takes the OL version first and the second half takes the TL version first. This means that, if conducted based on this data design, the data collection for the pre-testing phase should take place in two stages. This would not be a problem if participants can be given incentives and testing can take place on two specific dates with, say, one month in between. However, when data is collected using snowballing technique, there should be a reasonably extended period of time between the two administrations to make sure that 1) enough participants have taken the first test in either TL or OL and 2) practice effects are minimised. More importantly, this procedure can suffer from attrition if participants, for any reason, withdraw from taking the second test. Particular to this approach is the practical challenge to pre-assess the bilinguals' proficiency in both languages first before giving them the same questionnaire in both languages, thus testing them 3 times. Therefore in most cases researchers might rely on the assumption that participants are equally proficient in both languages, that is that they are equal on the independent variable level.

To counteract the practical and financial limitations of the single group design, we resorted to a two-group design where two independent groups of individuals from the same cultures are given the questionnaire in two different languages,

English and TL (Sireci, 2005). This was considered the best available option because:

1) Orpheus consists of 190 items, which makes it a very challenging to get the same participants to fill out the same questionnaire twice especially with limited funding available to offer incentives other than feedback on the test itself
2) Within subjects design in this context does not have an advantage over between subjects design because of the challenge to pre-assess participants' proficiency in both languages
3) Even if differences on the item are wrongly flagged due by individual differences or differences in sample characteristics between the two groups, all items that are flagged as DIF will be assessed qualitatively using cognitive interview during which any wrong assumptions could be rectified.

7.3.2.2. Participants Study 1

Participants in this study (n=194) were sampled through a snowballing technique. This was done by sending emails to friends and colleagues (appendix27), sending emails to staff in foreign language departments in the UK, and by posting notes at City University. Participants belonged to one of the following three groups, Arab world (n=62), China (n=68), and Spain (n=64). Nine participants were excluded from the sample because either they scored 2 or 3 on any of the response audit scales (which meant that they manipulated their responses) or their age group and/or gender information were missing (2 Arabic and 7 Chinese). The final number of participants included in the analysis was

185 with a gender ratio of 63.24% females and 36.76% males, presented in figure 7.2 below.

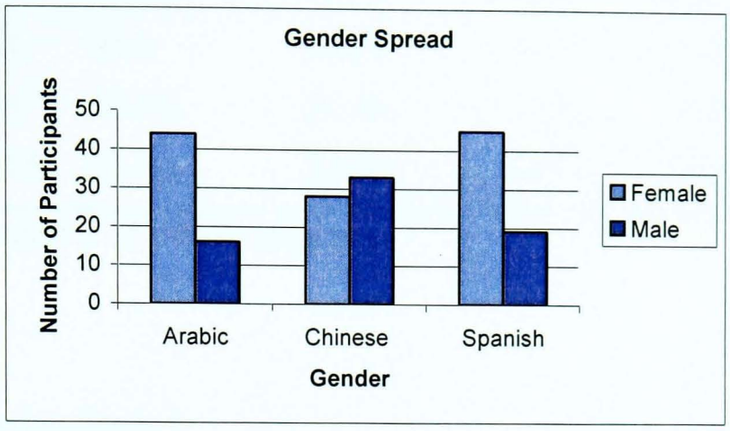


Figure 7.2: Gender

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 in line with the customary administration of Orpheus as shown in figure 7.3.

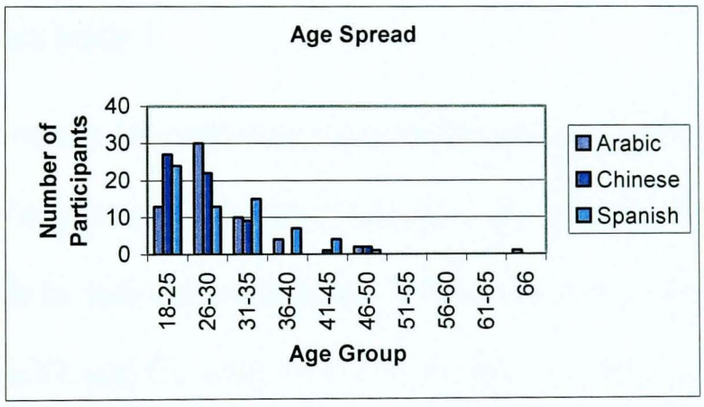


Figure 7.3: Age

Participants across the three cultures were predominantly between 18 and 35 (Arab world 86.7%, China 82%, and Spain 81.2%). Refer to table 7.1 for a summary of sample statistics.

Group	N	% Females	% Males	Age 18-25	Age 26-30	Age 31-35
Arab	60	73.3%	26.7%	20%	45%	18.3%
China	61	45.9	54.1%	47.5%	34.4%	13.11%
Spain	64	70.3%	29.7%	37.5%	23.4%	20.3%
Total	185	63.2%	36.8%	35.6%	34.6%	11.9%

Table 7.1: Summary of sample statistics

7.3.2.3. Materials Study 1

Four multi-lingual electronic versions (V4) of Orpheus in: Arabic, Chinese, English and Spanish comprising of 190 items each (appendix 18, 19, and 20).

7.3.2.4. Procedure Study 1

Approximately 60 participants in each culture group took part in this study, half of which filled out the questionnaire in the original language, English, and the other half in their native language (Arabic, Chinese, or Spanish). The questionnaires in TL and OL were sent as an electronic version, except in China where data was partly collected in paper and pencil format, and the information in the introductory email were used to explain confidentiality issues and the purpose of the research (appendix 27). As discussed in chapter 4, differences between paper and pencil versus computer based testing have been shown to be negligible (Bartram & Brown, 2004). All questionnaires included detailed instructions with an example to illustrate how the test should be completed. The instructions clearly states that there was no time limit associated with the test and

that participants could drop out at any point of the process. The instructions also encouraged participants to answer as honestly as possible as the questionnaire contains an honesty check. After completing the test, participants received a thank you email with a feedback report describing their personality preferences at work. There were 6 groups of participants in total, two language groups in each culture as follows:

1) Native Arabic speakers taking English version
2) Native Arabic speakers taking Arabic version
3) Native Chinese speakers taking English version
4) Native Chinese speakers taking Chinese version
5) Native Spanish speakers taking English version
6) Native Spanish speakers taking Spanish version

7.3.3 Results Study 1

7.3.3.1. Analysis

First of all, we converted the scores of each participant using within subject standardisation, also known as ipsative rescaling (Cheung & Rensvold, 2000) using the following equation (Rust & Golombok, 1999):

$$z = \frac{x - \bar{x}}{sd}$$

whereby $\bar{x}$ is the participant's average on all items, x is the participant's score on a specific item and sd is the participant's standard deviation based on his or her scores. Within subject standardisation is an essential part of Orpheus norm development, and is employed mainly as a way of controlling for acquiesce (Rust

& Golombok, 1999). Chapter 4 provides a more detailed discussion of within subject standardisation.

The items were then divided according to the major or minor scale they belonged to, 5 major and 7 minor scales (appendix 1 and 2). Furthermore, within each culture we analysed the data of each language group separately, so there were 6 groups in total, as shown above, two from each culture. Item analysis investigated two item statistics: facility and discrimination indices based on classical test theory (Kline, 1993; Sireci, 2005).

Comparing item facility

Within each culture, the facility index ($\bar{x}$) was computed for each "questionnaire language" group. The two means were then compared using t-tests with language version as the IV with two levels (English and TL) to assess whether there is an effect of language on item mean scores. Conducting multiple t-tests increases the likelihood of Type I errors, that is, rejecting the null hypothesis when in fact it is true. In this particular case, the null hypothesis assumes that there are no differences between the items in the two languages; therefore items are assumed to be equivalent. A type I error will lead to mistakenly rejecting this assumption and accepting the alternative hypothesis, which in this case assumes that there is a difference between some items in the two languages (Field, 2005; Fife-Shaw, 2006; O'Sullivan, 2006). Therefore, type I error will lead to flagging more items as functioning differently, whereby they actually are not.

However, since all statistically different items are scrutinised in cognitive interviews, type I error is not considered as a threat to the adaptation process at this stage. Items mistakenly flagged as functioning differently can be left as they

are to be discussed in depth during the interviews. Unlike other areas of research, such as experimental psychology, t-tests are not used to gather evidence and assume causality. Rather, they are merely a tool in the adaptation process for filtering items instead of taking them all (570 in total, 190 in three languages) into cognitive interviews because this could be an extremely time consuming activity.

The number of items that were significantly different at $p<0.05$ is: 40 in Arabic, 43 in Chinese, and 30 in Spanish. Tables 7.2, 7.3 and 7.4 list all the significant differences, and their effect size Cohen's *d* calculated as follow (O'Sullivan, 2006):

$$d = \frac{\bar{x}_{1-}\bar{x}_2}{\sqrt{(sd_1^2 + sd_2^2)/2}}$$

whereby $\bar{x}_1$ and sd_1 are the average and standard deviation of the first comparison group (TL group taking English test), and $\bar{x}_2$ and sd_2 are the average and standard deviation for the second group (TL group taking TL test) (O'Sullivan, 2006). It is important to state that significant differences between means were used as the criterion for including items in to cognitive interviews and not effect sizes. However, these significance levels were later compared with effect size in the second part of the result section to assess if either one of these is more accurate in detecting inequivalence.

Table7.5 presents all significantly different items across the three cultures, where only two items were significant simultaneously across the three cultures:

139- On some occasions I have behaved very improperly.

170- I will always take extra time to do a job well even when it's relatively unimportant.

Item number	Mean of English	Mean of Arabic	t(df)=x, all at p>0.05	Effect size d
114	0.913	0.4863	t(58)=2.014	0.26
167	0.5042	0.0707	t(58)=2.054	0.26
92	0.1095	0.65	t(58)=-2.07	0.26
128	0.4997	-0.0542	t(58)=2.083	0.26
182	-0.0885	0.3253	t(58)=-2.134	0.29
98	-0.3912	-0.8781	t(58)=2.149	0.28
146	0.105	-0.3173	t(58)=2.149	0.27
157	0.0758	0.5243	t(58)=-2.185	0.28
158	0.3082	-0.1551	t(58)=2.221	0.28
48	-0.536	0.0049	t(58)=-2.238	0.3
117	-0.2304	-0.2579	t(58)=-2.3	0.29
134	0.3028	-0.2332	t(58)=2.321	0.29
55	-0.4583	0.0571	t(58)=-2.324	0.29
185	0.3481	-0.2015	t(58)=2.349	0.29
97	0.3214	-0.3476	t(58)=2.444	0.31
101	-0.3842	0.1929	t(58)=-2.47	0.31
166	0.039	0.4916	t(58)=-2.48	0.31
104	0.167	-0.3981	t(58)=2.493	0.31
50	0.1154	0.7036	t(58)=-2.512	0.32
139	0.7846	0.1425	t(58)=2.519	0.31
107	1.3789	0.7747	t(58)=2.528	0.32
67	0.0043	-0.5241	t(58)=2.54	0.36
153	-0.5996	0.0714	t(58)=-2.719	0.34
132	-0.5726	0.001	t(58)=-2.816	0.35
2	-0.0802	0.6025	t(58)=-2.822	0.35
71	-1.0426	-0.3031	t(58)=-3.037	0.37
84	0.2263	-0.555	t(58)=3.07	0.37
6	-0.0811	0.762	t(58)=-3.116	0.38
89	0.7294	-0.0344	t(58)=-3.234	0.39
20	-0.1359	0.492	t(58)=-3.325	0.41
15	0.1377	-0.7035	t(58)=3.395	0.43
135	0.3182	0.9901	t(58)=-3.466	0.41
14	1.3216	0.5836	t(58)=3.57	0.44
136	-0.408	0.8267	t(58)=-3.647	0.43
142	0.0833	1.0602	t(58)=-3.896	0.46
162	0.7605	0.0454	t(58)=3.91	0.46
170	0.0328	-0.7296	t(58)=4.114	0.48
112	-0.3878	-0.7019	t(58)=-4.548	0.51
40	0.1779	-0.7049	t(58)=5.266	0.58
90	-0.3495	0.9775	t(58)=-5.889	0.64

Table 7.2 : Means for standardised items, t-values, and effect sizes for the Arabic pre-test study

Item number	Mean of English	Mean of Chinese	t(df)=x, all at p>0.05	Effect size d
8	0.25	1.0529	t(46)=-3.837	0.49
84	-0.0967	-0.7259	t(49)=-2.812	0.37
178	0.1176	-0.529	t(51)=-2.076	0.28
72	-0.3385	0.2413	t(51)=2.341	0.31
103	0.3522	1.1227	t(54)=4.274	0.5
52	0.3037	-0.9396	t(54)=-4.332	0.51
66	0.0002	0.5714	t(55)=2.187	0.28
25	0.0053	0.6199	t(55)=2.474	0.32
170	-0.1449	0.3927	t(55)=2.522	0.32
18	0.0667	0.8486	t(55)=3.232	0.4
48	-0.7428	0.2189	t(55)=3.866	0.46
120	0.3175	0.7874	t(57)=2.09	0.27
44	0.3034	-0.5609	t(57)=-4.034	0.47
98	-0.6464	-0.1753	t(59)=2.053	0.26
111	-0.2056	0.2621	t(59)=2.066	0.26
187	-0.3839	-0.9674	t(59)=-2.07	0.26
1	0.5634	1.0268	t(59)=2.093	0.26
90	-0.1347	0.4516	t(59)=2.136	0.27
92	0.3439	-0.1177	t(59)=-2.157	0.27
79	-0.3524	0.1189	t(59)=2.169	0.27
159	-0.1666	0.3591	t(59)=2.177	0.27
14	0.3915	0.9655	t(59)=2.213	0.28
95	0.4632	1.0402	t(59)=2.215	0.28
23	0.4827	-0.3197	t(59)=2.231	0.28
148	0.1393	0.652	t(59)=-2.233	0.28
67	-0.4781	0.0323	t(59)=2.248	0.28
124	0.3324	-0.1528	t(59)=-2.269	0.28
87	0.5178	0.0122	t(59)=-2.298	0.29
7	0.1176	-0.529	t(59)=-2.364	0.3
12	0.4749	1.0347	t(59)=2.421	0.3
97	-0.3629	0.1947	t(59)=2.53	0.31
153	-0.4151	-0.6106	t(59)=-2.545	0.31
139	0.276	0.8584	t(59)=2.557	0.32
82	-0.352	0.2788	t(59)=2.713	0.33
93	0.0992	-0.5851	t(59)=2.717	0.33
158	0.1489	-0.4665	t(59)=-2.846	0.35
46	0.1118	0.8904	t(59)=3.146	0.38
128	-0.6529	0.935	t(59)=3.192	0.38
108	-0.1456	-0.7601	t(59)=-3.214	0.39
112	-0.1673	-0.844	t(59)=-3.5	0.41
117	-0.1935	0.8458	t(59)=3.886	0.45
165	-0.1976	0.5005	t(59)=4.049	0.47
63	0.1739	-0.7901	t(59)=-4.053	0.47

Table 7..3 : Means for standardised items, t-values, and effect sizes for the Chinese pre-test study

Item number	Mean of English	Mean of Spanish	t(df)=x, all at p>0.05	Effect size d
88	0.4544	-0.4352	t((61)=-5.566	0.58
111	-0.104	-0.6656	t(44)= -2.870	0.4
175	-0.1697	-0.5437	t(50)=-2.13	0.29
118	0.0408	0.6071	t(54)=2.578	0.33
2	-0.1752	0.4185	t(60)=2.41	0.3
130	0.6179	0.1552	t(61)=-2.019	0.25
147	-0.7221	0.0218	t(61)=4.356	0.49
190	0.5279	0.0579	t(62)= -2.042	0.25
109	0.0914	-0.3437	T(62)=-0.21	0.25
123	-0.3297	0.0842	t(62)=2.00	0.25
122	0.14	-0.27	t(62)=-2.03	0.25
73	-0.1356	-0.5569	t(62)=-2.031	0.25
189	0.1558	-0.3184	t(62)=-2.07	0.25
100	1.0459	0.6621	t(62)=-2.079	0.26
126	-0.368	0.1367	t(62)=2.09	0.26
180	-0.1265	0.3909	t(62)=2.12	0.26
116	0.4003	-0.0785	t(62)=-2.123	0.26
94	-0.0919	0.4369	t(62)=2.130	0.26
170	-0.11	0.34	t(62)=2.170	0.27
131	-0.2255	0.278	t(62)=2.236	0.27
19	-0.4866	0.0842	t(62)=2.282	0.28
99	0-0.3144	0.2398	t(62)=2.47	0.3
58	0.3639	0.8506	t(62)=2.732	0.33
161	0.34	-0.23	t(62)=-2.86	0.34
139	0.2298	-0.4748	t(62)=-2.889	0.34
95	-0.2921	0.4167	t(62)=3.02	0.36
146	-0.3485	0.5126	t(62)=3.029	0.36
37	-0.5479	0.3323	t(62)=3.61	0.42
50	0.4541	-0.5559	t(62)=-5.70	0.59

Table 7.4 : Means for standardised items, t-values, and effect sizes for the Spanish pre-test study

Arabic		Chinese		Spanish	
Item	**Mean**	**Item**	**Mean**	**Item**	**Mean**
2	**0.6025**	1	1.0268	**2***	**0.4185**
6	0.762	7	-0.529	19	0.0842
14	**0.5836**	8	1.0529	37	0.3323
15	-0.7035	12	1.0347	**50**	**-0.5559**
20	0.492	**14**	**0.9655**	58	0.8506
40	-0.7049	18	0.8486	73	-0.5569
48	**0.0049**	23	-0.3197	88	-0.4352
50	0.7036	25	0.6199	94	0.4369
55	0.0571	44	-0.5609	**95**	**0.4167**
67	**-0.5241**	46	0.8904	99	0.2398
71	-0.3031	**48**	**0.2189**	100	0.6621
84	**-0.555**	52	-0.9396	109	-0.3437
89	-0.0344	63	-0.7901	111	-0.6656
90	**0.9775**	66	0.5714	116	-0.0785
92	**0.65**	**67**	**0.0323**	118	0.6071
97	**0.3476**	72	0.2413	122	-0.27
98	**-0.8781**	79	0.1189	123	0.0842
101	0.1929	82	0.2788	126	0.1367
104	-0.3981	**84**	**-0.7259**	130	0.1552
107	0.7747	87	0.0122	131	0.278
112	**-0.7019**	**90**	**0.4516**	**139**	**-0.4748**
114	0.4863	**92**	**-0.1177**	**146**	**0.5126**
117	-0.2579	93	-0.5851	147	0.0218
128	**-0.0542**	**95**	**1.0402**	161	-0.23
132	0.001	**97**	**0.1947**	**170**	**0.34**
134	-0.2579	**98**	**-0.1753**	175	-0.5437
135	0.9901	103	1.1227	180	0.3909
136	0.8267	108	-0.7601	189	-0.3184
139	**0.1425**	111	0.2621	190	0.0579
142	1.0602	**112**	**-0.844**		
146	**-0.3173**	117	0.8458		
153	**0.0714**	120	0.7874		
157	0.5243	124	-0.1528		
158	**-0.1551**	**128**	**0.935**		
162	0.0454	**139**	**0.8584**		
166	0.4916	148	0.652		
167	0.0707	**153**	**-0.6106**		
170	**-0.7296**	**158**	**-0.4665**		
182	0.3253	159	0.3591		
185	-0.2015	165	0.5005		
		170	**0.3927**		
		178	-0.529		
		187	-0.9614		

Table 7..5 : All significantly different items across the three languages

* items highlighted in bold are significantly different across more than one culture

Comparing item discrimination

Within each culture, we computed the item discrimination index "corrected item total correlation" of each language group. A Fisher transformation (Kanji, 2006) of the correlation coefficients to z-scores was used to test whether the two correlations were significantly different from one another using the following formula:

$$z = 0.5 * \log \frac{1+r}{1-r}$$

The analysis showed that 15 Spanish items, 3 Chinese items, 11 Arabic items had significantly different correlations at $p<0.05$. Table 7.6 lists the items in the Arab world, China and Spain that showed significant differences between the discrimination indexes.

Arabic	P value	Chinese	P value	Spanish	P value
8	0.013	61	0.003	3	0.028
10	0.026	160	0.016	34	0.002
43	0.003	165	0.035	41	0.018
53	0.036			44	0.014
82	0.000			**53**	**0.012**
84	0.007			55	0.035
89	0.035			63	0.033
107	0.021			65	0.013
108	0.039			78	0.022
152	0.031			85	0.045
168	0.023			97	0.009
				109	0.017
				153	0.045
				168	**0.039**
				188	0.044

Table 7.6: Items with significantly different corrected item total correlation across the three languages

Items84, 89 and 107 in Arabic, item 165 in Chinese, and item 109 in Spanish had significantly different difficulty indexes as well as discrimination indexes, therefore the total number of items that were taken to the cognitive interview was

48 in Arabic, 45 in Chinese and 43 in Spanish as shown in appendix 28, 29 and 30.

7.3.4 Study 2: Cognitive Interviewing

7.3.4.1. Participants Study 2

We recruited participants in this study (n=12; 4 from each target group) from a sample of convenience. All participants had lived in their home country most of their lives and in the UK for at least 2 years and had a higher education from a UK institution and good proficiency in English and target language. Participants' age ranged from 23 to 61 ($\bar{x}$ =30.42; *sd*=10.37). In each culture group, there was one male and three females.

7.3.4.2. Materials Study 2

The material used in this study consisted of the following:

1) Items in TL and in OL written separately on white cards (10.5x29.7 cm)
2) Confidence rating scale on white card (10.5x29.7cm) (appendix 31)
3) Cognitive interview protocol (appendix 32)
4) Same item in English, TL, and Changed TL written in a table on A4 paper

7.3.4.3. Procedure Study 2

We adopted a combination of think aloud and verbal probing techniques (Willis, 2005). Therefore we developed an interview protocol consisting of main questions and additional scripted probes prior to the interview targeted at

understanding the thinking process of interviewees (appendix 32).

At the beginning of the interview, participants received a consent form and a debriefing form (appendix 33 and 34) explaining how the data was used to encourage them to respond freely and honestly by stressing on confidentiality, and asking their consent for audio recording the session. Since the number of items showing significant differences were large (48 Arabic, 45 Chinese and 43 Spanish as shown in appendix 28, 29, 30), four interviews in each culture. Each cognitive interview lasted for approximately 2 hours during which between 12 and 16 items were discussed in depth. Two participants were recruited from each culture, one male and one female, to take part in a cognitive interview, but each one was presented with different items.

During the first half of the interview, the interviewee was presented with all items on a card, one at a time in alternating languages (TL or English). However, each item was presented in one language only. The participant was asked to read the item and paraphrase it when it was presented in English, or translate it to English when it was presented in the TL. In case the participant did not understand a word in the English versions of the item, he or she was given a dictionary to look up the meaning. The participant then rated whether he or she Strongly Disagree (SD), Disagree (D), Agree (A), or Strongly Agree (SA) to the item and explained with examples the reason behind their choice. The interviewer probed when the participant did not give enough information about the item.

During the second half of the interview, participants received the same items but in the other language. Therefore, by the end of the interview, both language versions of each item were reviewed by the same participant but at

different times. This approach was adopted in order to minimise practice effect and to give the participant enough time to forget the exact phrasing of the item in the other language. As before, participants were asked to paraphrase or translate the item then to explain the reason for their choice of answer option. The interviewer probed as before, when participants did not reveal all the information necessary to aid the interviewer in comprehending their thinking process. For example, if the participant was presented with the item:

2- I enjoy talking to my friends about work.

And explained that they chose "agree" because "they like to talk to their friends about work", the interviewer would probe further because the participants answer in this case a merely a repetition of the item rather than an explanation of the reasoning behind their choice.

Participants were then presented with the item in the language they were presented with in the first half of the interview, and asked to rate the similarity between the two language versions of the items on a 5 point Likert scale: Not similar at all, not very similar, similar, very similar, and exactly the same. They were then asked to explain their choice and provide amendments for the items if they rated the similarity anything but "exactly the same".

As all cognitive interviews were conducted by the project manager, there was an element of subjectivity in the probing and evaluation process. Additionally, using two native speakers from each culture to take part of the interview also adds subjectivity to the process. Therefore, it was important to cross check the suggested changes suggested by the cognitive interviewees in a panel discussion before piloting. All suggested changes to the items were therefore discussed in a panel with two native speakers of the target language

simultanuously in order to confirm the suggested amendments from the cognitive interview were appropriate and necessary.

7.3.5 Results Study 2

The total number of items analysed in cognitive interviews was 136 (48 Arabic, 45 Chinese, and 43 Spanish) out of which 67 were changed (33 Arabic, 10 Chinese and 24 in Spanish) as shown in table 7.7 below. All changes were suggested in the cognitive interview and then agreed with a panel of reviewers.

Arabic (33 items)	**Chinese (10 items)**	**Spanish (24 items)**
6	12	34
8	18	41
10	25	53
15	46	55
20	63	63
40	67	65
43	90	73
50	112	85
53	117	88
55	158	95
67		97
71		100
84		116
89		118
90		126
97		130
98		131
101		139
104		146
108		147
112		153
114		168
117		170
132		175
136		
139		
146		
152		
153		
158		
168		
170		
182		

Table 7.7: Items analysed in cognitive interviews

The changes varied in nature, same as in the dyads and triads. For example, some words were used in the wrong context such as the following item in Chinese:

8- I sometimes wish I was more able to speak my mind

To which the participant being interviewed explained that "I think this is not very correct. We do not normally use this word in this context".

Very commonly, the changes were due to wrong or inaccurate translations such as the case of item 73 in Spanish.

73- It's often necessary to break the rules in order to get things done.

The item was first translated to "Often, it is necessary to break the rules to do things", and then rectified during the cognitive interview to "Often, it is necessary to break the rules to get things done". The difference between the two sentences is that, one might break the rules in order to meet deadlines, but they would not necessarily break the rules all the time to "do things". Here is another similar example from the Chinese interviews:

44 – it always pays to tell the truth

When the Chinese participant was presented with the English version of the item and was asked to explain it in her own words, she said "there's a price to pay if you try to promise to tell the truth". The PM explained that the she understood it wrong and provided her with the correct explanation to which she replied: "Ah Ok. I did not notice the structure". Interestingly, when she was presented with the Chinese version, it turned out that the item was translated wrong as well: "Not similar at all because (…) the Chinese version is totally opposite, it says you need to pay a price if you tell the truth".

Some other items were changed because of missing words such in the case of the

following item in Arabic:

71- My work is more important to me than almost anything else.

The participant explained that the English and the Arabic versions were "not very similar because here it says then *almost anything else* and here it says *anything else*. You know what I mean, my work in more important to me that almost anything else but not more important than everything else".

However, there were situations where the changes were done for stylistic reasons rather than for affecting the meaning of the items. For example:

7- I find that my day-to-day work performance varies with my mood.

The Chinese participant argued that "I would put a word here. A Chinese person would understand it even without this word but you need it for grammar".

As a final example, items were no equivalent and needed changes but no alternative was available due to idiosyncratic language issues such as:

167- I am sometimes too rash in making decisions.

There were differences in the magnitude of the word "rash" between the English and the Arabic versions but could not be changed because no alternative was found. The participant disagreed to the item when it was presented in English "I don't agree. I usually think my decision especially if they are big." However, she explained that the two versions were "not very similar, the Arabic version is lighter and I would be more likely to say yes to it". When asked to provide an alternative that would be closer to the English version in magnitude, she could not do so even with the use of the dictionary.

7.4. Discussion

Seventy percent of the significantly different items in Arabic were changed after the cognitive interview, 22% of Chinese items and 53% of Spanish items. The fact that items were changed across all the languages reflects the importance of the pre-testing in detecting problems of equivalence that have previously been overlooked. Conversely, since not all items needed changes, it seems reasonable to use t-tests as a filtering technique and not to run all items through the time consuming process of cognitive interviewing.

The percentage of changed items during the cognitive interview is not a good representation of which language was most problematic. Some items were not changed because there was no better alternative that can be used to increase the linguistic, psychological, and cultural equivalence between the items. For example, the Arabic version of item 167 "I am sometimes too rash in making decisions" was agreed to be different in magnitude to the English equivalent. The word "rash" was seen as milder in Arabic, yet, there was not alternative word that could replace it to make it as strong as the English equivalent. Such problems are due to particularities of the target language and might have to be dropped after the pilot study if they continue to exhibit differential functioning between cultures. This highlights the point that developing equivalent versions across cultures can only be accomplished by reducing the number of items of the final questionnaire. Some items are idiosyncratic and can never achieve equivalence especially that, as argued in chapter 2, the Big Five Model is highly dependent on language, wording and the connotative meaning of words.

As for the comparison between p values and effect size as a judgement method, the results showed that this could not be established in this study. In

order to decide which technique is less sensitive to error, it is necessary to investigate the p values and the effect sizes of items that were changed. However, some items needed changing but this was not possible due to other factors, such as language idiosyncrasies discussed above. Additionally, the effect size *d* of the items changed in Arabic ranged between 0.26 and 0.64 in comparison to 0.26 and 0.46 for unchanged items. Similarly, for Chinese items the effect size ranged between 0.27 and 0.47 for changed items and 0.26 to 0.50 for unchanged. The Spanish data also mirrored these results whereby changed items had d between 0.25 and 0.58 and 0.25 and 0.59 for unchanged. Therefore, both changed and not changed items had similar effect sizes. Therefore, no conclusion can be drawn from this study about the effectiveness of either effect size of p value in flagging problematic items. Although not ideal, p value remains useful in filtering items into the cognitive interview.

Results from this study mirror some of the results from the previous study. Cognitive interviews revealed that linguistic, cultural, and psychological changes prompted most of the changes to the items, although some of the changes were not possible. The example discussed above, item 167, and illustrates a case of differential magnitude that was elicited from the previous study. As another example, item 44 – it always pays to tell the truth, discussed above was *translated wrongly* as "you pay a price if you tell the truth". This item was not problematic in Arabic neither in Spanish. Further discussions about this item revealed that this might have occurred because there is famous Chinese saying that "做个老实人容易吃亏" (being an honest man, very likely others would take advantage from you), which might have lead translators to assume that this is the meaning intended from the English item. The Chinese participant

in the cognitive interview explained that the idea from the proverb "is a common thought in many Chinese people's mind". However, it could also be argued that "it pays" might be syntactically confusing for Chinese participants because it is not clear what "it" refers to, and there is no *you* (it pays *you*) to indicate that who the truth pays back to. Moreover, since this type of sentence structures does not exist in Chinese, "it pays" could be understood as "it pays price" (negative) or "it pays prize" (positive). As indicated by one of the Chinese participants.

Another issue relating to this item is the contradiction scale. As mentioned earlier, participants who scored 2 or 3 on any of the audit scales were disregarded from the analysis. However, the audit scores are computed based on participants' responses to specific items such as this one. Participants who agree to

item 44 "it always pays to tell the truth"

and agree to

item 16 "There are times when it's not sensible to tell the truth"

are contradicting themselves and this counts towards their audit score on contradiction. Since item 44 was wrongly translated in the Chinese version, participants who agreed to both item 44 and 16 were not actually contradicting themselves but their answers might have lead to this assumption. These two items do not constitute the whole contradiction score, however, the answers to this item might have pushed some participants' scores on contradiction to become higher and some others lower. In this particular study, only seven respondents were disregarded from the Chinese sample, and this did not affect detecting flagging this item as problematic. Maybe this was because this was the only item that was translated wrongly from the items that measure contradiction.

So participants who scored 2 or 3 must have contradicted themselves on other items from that scale as well. However, if there were more items translated wrongly from that scale, more participants would have been wrongly removed from the analysis thus hindering the flagging of these items as problematic. Therefore, it is advised that audit scales are not used to make any decisions about disregarding or keeping participants in the pre-test until the questionnaire has been fully validated in the other culture.

Significant differences between the means of the two samples indicate that the patterns of responding of participants from the same culture to the same item in different languages are inconsistent. While this could be the result of linguistic differences between items, we cannot rule out the effect of individual differences on responding. For example, in the Arabic sample, 19 out of 30 participants who filled out the Arabic version were women compared to 25 out of 30 who filled out the English version. Difference in the item means between the group that filled out the English version and one that filled out the Arabic version could be significant because of gender differences rather than linguistic differences. Additionally, the majority of the participants in all the samples and sub samples were under the age of 35 Arab world 86.7%, China 82%, and Spain 81.2%). This indicates that the samples used in the pre-test study do not fully represent the target population. Having differences in sample characteristics in the pre-test stage is not problematic because no items are being dropped at this stage. Additionally, variation resulting from individual differences should be much less in the pilot study where the number of participants is larger. The most fundamental point here is to be aware of such group differences and their implication on results and also to incorporate this knowledge in the interpretation

of DIF analysis. That is, it is important to take each significantly different item independently and examine which source affected this item the most.