# Lost In Translation? Challenges in Using Psychological Tests in the Philippines

Allan B. I. Bernardo
De La Salle University
Manila, Philippines

Filipino psychologists often use foreign-made psychological tests in English in their professional practice and in research. The paper raises questions on the validity of such tests for use with different Filipino respondents. Drawing from international standards for translating and adapting tests, the various levels of equivalence (qualitative and quantitative) between the original tests and their translations are discussed. The different types and sources of bias (construct, method, and item biases) that lead to non-equivalent translations of tests are also explained. The paper then reviews research on the equivalence of Filipino translations of tests with their original English versions and points to the strong possibility that the translations, as well as the English versions of the test used with Filipino participants, are not equivalent to the original tests used with the original target populations. The paper ends with a discussion of possible courses of action and the need for collective action from different sectors of the Filipino psychology community to address the concern.

**KEYWORDS:** psychological tests; translation; equivalence; bias; language; bilinguals; Philippines

Psychological tests have long been used in the Philippines—in the recruitment and selection of employees, in the admission of students in schools, in diagnosis of psychological and psychiatric conditions, among others. In recent years, there have been numerous developments that have institutionalized the use of psychological tests in legal and official government procedures

such as in the diagnosis of psychological incapacity in petitions for nullification of marriages, diagnosis of whether juvenile offenders are able to discern whether their actions are right or wrong, among others. There have been discussions on the possibility of using psychological tests in other domains like for screening applicants for oversees contract work, and perhaps half-seriously, for screening candidates for political positions. In one sense, these developments recognize the usefulness and validity of psychological testing as a measure of understanding of some facets of a person's experience, and the importance of psychological tests in various aspects of societal functioning. However, there are concerns regarding whether these tests are being used properly, and by people who are properly trained to use such tests.

One specific issue regarding the use of psychological tests in the Philippines relates to the cross-cultural and cross-linguistic validity of psychological tests in English that were developed in North America, Australia, and other English-speaking countries where most psychological tests are being constructed, validated, published, and sold. This issue is not specific to the Philippines, and indeed, is an issue all over the world. As such, there has already been quite a significant amount of work undertaken to safeguard the integrity of psychological assessment as a professional and scientific procedure. Moreover, various standards have been articulated for the use of psychological tests in cultures and languages other than where they originated (Hambleton, 2001; Van de Vijver & Tanzer, 2004). Unfortunately, however, such standards have not been the focus of much attention in the Philippine context.

In the Philippine context, another important consideration is the fact that most Filipinos taking psychological tests are either bilingual or multilingual. Research shows that bilinguals' responses to psychological tasks may vary depending on the language used in the task. In a study looking into perceptions of other people's personality Hoffman, Lau, and Johnson (1986) found that Chinese-English bilinguals from Hong Kong recalled different aspects of a target person's personality depending on whether the descriptions of the person were given in Chinese or in English. When it comes to retrieval of autobiographical memories, Marian and Neisser (2000) found that Russian-English bilinguals recalled life experiences differently depending on whether the elicitation of memories was done in Russian or in English. Moreover, the intensity of the affect expressed during the recall of autobiographical memories depended

on the language used as well (Marian & Kaushanskaya, 2004). Marian and Fausey (2006) account for these language effects in the memory-related psychological processes of bilinguals by proposing a language-dependent memory system for bilinguals, which is consistent with various studies showing language-dependent benefits and/or deficits in cognitive processing of Filipino-English bilinguals (e.g., Bernardo, 1996, 1998, 1999, 2001b, 2002, 2005; Bernardo & Calleja, 2005). Thus, in administering psychological tests and tasks among Filipino bilinguals, the language used may sometimes influence how the respondent performs in the test.

This paper addresses the various issues related to the use of English-language psychological tests developed in foreign countries with Filipino respondents, who are presumably at least bilingual and even possibly multilingual. The paper discusses the different issues and standards related to the translation of psychological tests, focusing on issues of equivalence and bias. In discussing these issues, examples shall be provided involving psychological tests used with Filipino participants. The paper ends with a series of recommendations regarding how Filipino psychologists can ensure the validity and integrity of psychological testing in the Philippines, as local psychologists continue to use foreign-made, English-language psychological tests.

## GENERAL OPTIONS FOR FILIPINO PSYCHOLOGISTS

To preface the discussion on the general options for Filipino psychologists who wish to use foreign-made psychological tests in the English language, let us consider this earlier version of the Beck Depression Inventory (BDI, 1995) which is readily available on the Internet (see e.g., http://www.ibogaine.desk.nl/graphics/3639b1c_23.pdf). The BDI is intended to measure whether the respondent is experiencing depression and to determine the level of depression (e.g., borderline clinical depression, moderate, severe or extreme depression). Because of its availability on the Internet, psychologists and psychology students have often used the BDI for diagnostics and research. The inventory consists of 21 items referring to different aspects of behavior (e.g., pessimism, self-dislike, past failures, etc.). For each item, the respondent is given four options referring to different experiences regarding the particular aspect of behavior, and each of the four options corresponds to a score. The respondent's total

score for all 21 items is summed and compared to normed levels of depression.

Now consider the following item (#11) that refers to agitation:

11. Agitation
    - 0 I am no more restless or wound up than usual.
    - 1 I feel more restless or wound up than usual.
    - 2 I am so restless or agitated that it's hard to stay still.
    - 3 I am so restless or agitated that I have to keep moving or doing something.

What options does a Filipino psychologist have when using this specific item in trying to assess depression in a client or respondent? The psychologist can use the item as it was originally developed and hope that the client understands the item. This option should be fine if the respondent is adequately proficient in English. However, a client who is not proficient in English may struggle with words like "restless," "wound up," and even "agitated."

Translation would be the next option for the Filipino psychologist, and other scholars have reported the translation work that has been done in the Philippines on various psychological tests. But going back to the BDI item on agitation presented earlier, the translator would most likely have a difficult time translating the very same words, "restless," "wound up," and "agitated." The closest Filipino translation of "restless" that comes to mind are low frequency words such as *"balisa," "hindi mapakali"* and *"hindi mapalagay."* However, *"balisa"* is a low-frequency Tagalog word that may not be understood by most Filipinos who do not have a deep knowledge of Tagalog. The other two translations are higher-frequency terms, but may be understood to have a more "physical" element (i.e., similar to *"malikot"*), and thus, may not capture the full affective sense of "restless." In the main section of this paper, the various issues related to test translation will be discussed, as this seems to be the most viable option for Filipino psychologists who want to take advantage of the availability of foreign-developed tests with established reliability and validity.

However, there is still a third option for the Filipino psychologist, and that is not to use these foreign-made tests, and instead develop and validate new psychological tests for local use. Many Filipino psychologists have actually advocated this option as early as the 1970s. Enriquez (1992) criticized foreign-made tests, particularly those in the English language as being invalid for use among Filipinos,

and advocated the development of indigenous tests. Over the years, many such indigenous psychological tests of personality and intelligence, among others, have been developed (see Cheung, 2004; Church, 1987; Guanzon-Lapeña, Church, Carlota, & Katigbak, 1998). But interestingly, most practicing psychologists in the Philippines still prefer to use the foreign-made tests. This issue is another topic of discussion and is not addressed in this paper. Instead, the paper focuses on issues related to the first two options (i.e., use of original English tests, and use of translations of foreign made tests in English), as these seem to be the more common practice of Filipino psychologists who do testing today.

## ISSUES RELATED TO TRANSLATING ENGLISH-LANGUAGE TESTS INTO PHILIPPINE LANGUAGES

Historically, translation of psychological tests into different languages aimed to achieve a close linguistic translation, and these involved various forward and backward translation techniques (van de Vijver & Leung, 1997; Werner & Cambell, 1970). But not all items in a psychological test are translatable (i.e., the linguistic translation closely captures the psychological meaning of the item). Indeed, there are test items that are poorly translatable; that is, although the meaning is translatable, some conciseness is lost or some nuance in the meaning is lost (see the earlier example about using *"balisa"* to translate "restless"). In some cases, the item may actually be untranslatable, and these occur when there is absolutely no overlap between the linguistic and psychological features of the item in translation (consider for example, how the personality description of "happy-go-lucky" can be translated into Filipino). Thus, many practitioners and testing experts soon realized that a close linguistic translation could also involve many problems. Now test translators emphasize adaptation and localization in translation instead of close translations (Hambleton, 1994; Van de Vijver & Tanzer, 2004).

Adaptation involves literal translations of those parts of the items that are translatable and modifying other parts of the items (and even creating new items) based on the assumption that a close translation might lead to having a biased or non-valid psychological test. Instead of focusing on the linguistic translation, there is more consideration given to ensuring that the target psychological construct is adequately measured in the various languages. The State-Trait Anger

Expression Inventory (Spielberger, 1988) is one good example of a psychological test that has been translated into almost 50 languages, where the translations do not contain close linguistic translations of all the original English items. But in each language translation the psychological constructs of state and trait anxiety were assessed in a valid manner.

The work on test adaptation and translation has been quite extensive that important testing organizations have already drafted clear guidelines for doing so (see Hambleton, 2001). This paper shall not endeavor to repeat what has been said in these guidelines; instead, it shall emphasize some key concepts and elucidate on these concepts in relation to the experience of psychologists in the Philippines.

We have already noted that close linguistic translations may sometimes cause psychological tests or test items to become less valid measures of the intended psychological construct, and that the goal of translation is also to achieve psychological equivalence in the translations. But what does psychological equivalence mean? There are actually several levels of psychological equivalence that need to be considered, particularly as psychological test items are intended to be a quantitative measure of a psychological construct. The psychological testing literature (see e.g., Poortinga, 1989; Van de Vijver & Leung, 1997; Van de Vijver & Tanzer, 2004) refers to different hierarchically linked types of equivalence. In the broadest sense, equivalence between translations refers to a) similarity in the psychological meaning of the test, which is more of a qualitative equivalence, and b) similarity in the meaning of the scores of the tests and the items, which is more quantitative.

More extensive discussion of the types of equivalence can be obtained from other sources (Van de Vijver & Leung, 1997; Van de Vijver & Tanzer, 2004), but we should underscore some important distinctions about these types especially as they relate to how psychological tests and their translations are applied. The first type of equivalence is often called construct equivalence (but also functional or structural equivalence). When two translations of the test are equivalent, the same psychological construct is measured in the two language versions. Construct equivalence presupposes that the psychological construct is universally meaningful, or that the concept is understood in the same way in different cultures and languages (Van de Vijver & Tanzer, 2004). In other words, a psychological test that is supposed to measure fluid intelligence is supposed to measure fluid intelligence in the Philippines or in Italy, whether the test is in Filipino

or Italian. If the test measures fluid intelligence in the Philippines, but actually measured scholastic ability in Italy, then there is no construct equivalence. Moreover, in multidimensional constructs, construct equivalence also assumes that the construct measures the same dimensions and the same relationships among these dimensions in the different target populations. For example, if the psychological test is supposed to measure five interrelated personality dimensions, it should measure these five interrelated dimensions in the Philippines and in Argentina, whether the language of the test is Filipino or Spanish. If the test measures five personality dimensions in the Philippines, but only four dimensions in Argentina, then there is no construct equivalence.

The quantitative equivalence of psychological tests relates to the scores and their interpretation, and there are two levels of this: measurement unit equivalence and scalar (or full scale) equivalence (Van de Vijver & Tanzer, 2004). With measurement unit equivalence, two translations of the test are assumed to reflect differences in the target construct in their respective target populations to the same degree. In other words, the two tests have the same measurement unit; for example a two-point difference in scores means the same thing within each test. However, measurement unit equivalence does not guarantee that scores from the two tests are comparable to each other, as there might be a constant offset compared to another measure, which is the case when there are different norms for each version of the test. As such, the scores from the two scales may not be comparable to each other, even if scores within each scale can be compared to each other in the same way for both scales. When there is no constant offset, or when the scores in the two versions have the same quantitative origin or intercept, the two tests are said to have full scalar equivalence. In such cases, the interpretation of scores is exactly the same in both versions and is comparable across the scores.

To summarize, when Filipino psychologists translate English-language psychological tests into any Philippine language, they should be concerned about the qualitative and quantitative equivalence of these with the original English versions. Otherwise, there might be no basis for interpreting the tests in the manner suggested in the original test manuals. Unfortunately, it is not safe to assume that a good linguistic translation would have these properties and thus make the translation equivalent to the original. Indeed, there are many factors that create non-equivalence between translations of psychological tests, and these are called different types of bias. These different types

of bias are discussed in the next section.

## BIAS IN TRANSLATIONS OF PSYCHOLOGICAL TESTS

In the psychological testing literature, the opposite or lack of equivalence is defined as bias. Different types of bias create different degrees of non-equivalence between the translation and the original. There are parallels between the types of bias and the levels of equivalence, but the parallels are not absolute. In this discussion, the three types of bias identified by Van de Vijver and Poortinga (1997, see also Van de Vijver & Leung, 1997) are referred to as: a) construct bias, b) method bias, and c) item bias.

***Construct bias.*** Construct bias is observed when the psychological tests measure different psychological constructs in two different cultural or linguistic groups, or when the psychological constructs measured are merely partially overlapping. Construct bias may also occur when different behaviors and consequences are associated with the psychological construct in different cultural or linguistic groups. Earlier we suggested that some tests that are supposed to measure intelligence may actually be measuring scholastic ability in other cultures, and this is possible in a host of other psychological constructs. Even seemingly simple psychological constructs can be given different meanings in different cultures.

Even a simple affective concept like "happiness" can actually be difficult to measure across cultures. The Happy Planet Index aimed to measure levels of happiness of people in different countries, and it found the people in Costa Rica to be the happiest people in the world; Filipinos were ranked 14th (www.happyplanetindex.org). But does "happy" mean the same thing across all cultures? Comparing North Americans with East Asian, Uchida, Norasakkunkit and Kitayama (2004) found different associations with the concept of happiness. In particular, in North American contexts, happiness tended to be associated with personal or individual achievement, positive affect, and self-esteem. In contrast, in East Asian contexts, happiness tended to be associated with interpersonal connectedness, attaining balance between positive and negative affect, and a perceived embeddedness of the self within one's social relationships. Even what is considered positive affect may also be different in these two cultural contexts. Tsai, Knutson, and Fung (2006) found that North Americans and East Asians seem to have different notions of what the ideal affective or

emotional state involves. While North Americans aspire more often to a high-energy elation, thrill, and excitement (or high arousal positive affect), East Asians aspire more often to a tranquil joy and calm (or low arousal positive affect). These differences also have distinct psychological correlates and consequences in both cultures (Tsai, 2007).

Given these cultural variations, we can see how specific psychological tests that are intended to measure affective states may have some problems related to construct bias. Consider another item in the Beck Depression Inventory (1995) that refers to "loss of pleasure":

4. Loss of Pleasure
   - 0 I get as much pleasure as I ever did from the things I enjoy.
   - 1 I don't enjoy things as much as I used to.
   - 2 I get very little pleasure from the things I used to enjoy.
   - 3 I can't get any pleasure from the things I used to enjoy.

Will the concepts of pleasure and enjoyment refer to the same thing in all cultures where this item is translated? Would the loss of pleasure be indicative of depression to the same extent across all cultures?

In the Philippine context, our research group has encountered similar instances of possible construct bias of some psychological constructs. One example involves the psychological construct "academic emotions," which is an important construct in the field of educational psychology, and for which there are readily available and validated psychological tests (Pekrun, 2006). In a study that explored the construct of academic emotions, Bernardo, Ouano, and Salanga (2009) observed that the term "emotions" could be translated as *"nararamdaman"* or *"damdamin"* in Filipino. The word "nararamdaman" does not refer exclusively to affective states, and may actually include even physical states. It is not surprising, therefore, that the study found Filipino students reporting *"pagod"* or *"kapuy"* (tired) or *"antok"* (sleepy) when asked about the emotions they experienced in the classroom. This is a case when the translation only partially overlaps with the intended psychological construct of academic emotion. The domain of emotions is likely to be an area where linguistic terms only partially overlap with the psychological construct when studied across cultural or linguistic groups (e.g., think of how "contempt" can be translated in the different Philippine languages).

But construct bias can also be found when assessing non-affective psychological constructs. One compelling example was actually found when Watkins and Gerong (1999) wanted to study the self-concept of Cebuano students. In their study, they used both the English and the Cebuano translation of a standard self-concept scale. They found that the students who answered the test in Cebuano responded to the self-concept scale using aspects of the self that related to their family and community roles and relationships. On the other hand, the students who answered the test in English responded to the self-concept scale by referring to aspects of the self that related to their being a student. Thus, two different components of the self-concept were being evoked by two language versions of one scale in one bilingual population. This is an interesting case where the methodological aspect of the psychological test (i.e., the language) seems to be associated with construct nonequivalence.

But before method bias is discussed in greater detail, we should underscore the importance of considering construct equivalence and bias in the translation, adaptation, and even use of psychological tests. Filipino psychologists should be mindful of the possibility that some of the psychological constructs that are measured in standardized psychological tests actually do not mean the same thing in the Philippine context. Moreover, given the heterogeneity in the cultural environments within the Philippines (i.e., large variations in level of urbanization, industrialization, availability of communication technology), it is possible that what we assume to be standard psychological principles relating to constructs may actually have functionally different meanings for different Filipinos.

***Method Bias.*** The second type of bias relates to different aspects of the psychological testing process, and has three different sources: [a] sample bias, [b] instrument bias, and [c] administration bias. When Filipino psychologists use psychological tests, they have to be mindful that there are vast differences in educational level, language proficiency, and cultural experiences among Filipinos coming from different sectors of society. Thus it is possible that how a score of a respondent differs from the norm reflects differences in experiences relative to the norm group, and not actual psychological differences. In such cases, we have a form of item bias. When test norms and interpretation of scores are derived from a homogenous population, it is difficult to rule out sample bias, especially in a very diverse country like the Philippines.

Related to sample bias is instrument bias, which refers to

characteristics of the instrument, such as the nature of the response options, demand characteristics of the tests, and other factors that relate to response biases in different cultural groups. For example, an American study found that Hispanic respondents tended to use the extreme values of the five-point scale more than Caucasian respondents (Marin, Gamba, & Marin, 1992). Interestingly, bilingual Hispanics showed the bias to use extreme values when the questionnaires were in Spanish, but not when these were in English (Hui & Triandis, 1989). Closer to home, Smith (2004) found that the respondents from the Philippines are among the highest in showing an acquiescent bias, or the bias for more extreme responses at the positive end of the response scales (but see Grimm & Church, 1999). Watkins and Cheung (1995) studied various types of response biases (e.g., positivity and negativity bias, low standard deviation, inconsistency of related items, and consistency of unrelated items) in various cultures including the Philippines and suggested that these response biases may reflect differences in academic ability, and perhaps, intellectual ability, among other variables. But at this point, we do not know whether there are differences in levels of these various response biases for different subgroups in the Philippines.

Other than response bias, there are other forms of instrument bias. For example, the language of the psychological test may also be a source of bias, when Filipino translations are given to Cebuano, Ilocano, or Waray speakers. Using standardized computerized testing procedures may create bias for individuals who are not accustomed to using the computer.

The last form of instrument bias relates to the administration of the test, starting with how instructions are communicated by the test administrator to the respondent. There may be differences in how instructions are understood if there are language differences between the administrator and respondent (Gass & Varone, 1991). Even very subtle factors like if the test administrator is perceived to be violating cultural norms of communication in the test administration process (Goodwin & Lee, 1994) can be biases in the testing process. In the Philippine context, attempts to help the respondent who may not be familiar with psychological testing or with the English language can also create method bias. Test administrators in some Philippine testing centers have shared how they sometimes provide additional explanations or elaborations on the items if it seems that the respondent does not understand the question. These procedures are ad hoc and are not standardized for all respondents, creating so much

room for method bias.

***Item Bias.*** Even if we assume that problems related to construct and method bias are addressed, there is still the more subtle form of bias, which is the bias at the level of the item. This type of bias has also been called differential item functioning (DIF) in the psychological testing literature (Holland & Wainer, 1993). An informal definition of DIF would be that there are specific problems at the item level that make the item incomparable across translations or test versions. Operationally DIF is observed when two persons coming from two cultures and answering two versions of the same psychological test with the same level of the psychological construct do not score in the item in the same way. For example, imagine a group of Filipino adults and a group of Dutch adults all attain exactly the same score in a psychological test measuring anxiety (but in the Filipino and Dutch languages, respectively), so they can all be presumed to have the same levels of anxiety. But for one specific item of the psychological test of anxiety, the Filipinos score the item significantly higher than their Dutch counterparts. This would be a case of DIF.

Why would specific items function differently in different versions or across cultures? There are many different sources of DIF, the most basic of which may be a poor translation of the item, which may have led to an ambiguous or even incorrect statement of the item in one version of the test. But there could also be cultural specifics related to the connotative meaning or cultural significance of certain words, items, or concepts. Extending the Filipino and Dutch examples, consider a hypothetical item from a hypothetical Dutch test of crystallized intelligence borrowed from Van de Vijver (personal communication): *"Hoe heet de koningin van Nederland?"* The literal translation of the item in English is: *"What is the name of the queen of the Netherlands?"* and in Filipino it is: *"Ano ang pangalan ng reyna ng Netherlands*?" The first problem with the translation of the item is that although knowing the name of the queen of the Netherlands may represent crystallized intelligence for Dutch children, it does not do so for Filipino children, and hence the latter would find this item much more difficult than their Dutch counterparts. The item may be modified and adapted in the Philippine context to *"What is the name of the president of the Philippines?"* Now both items refer to the head of state of the respective countries. However, there are still cultural differences in how salient the head of state is in the lives of children in these countries. In the Netherlands, the queen is head of state for life, and her name and image are everywhere including the coins,

stamps, most buildings including schools, among others, and is thus, quite a common sight for most children in The Netherlands. In the Philippines, the president changes every few years, and names and images of present and past presidents can be found in various places in different parts of the country. Thus, knowledge about the name of the president of the Philippines may not be as well established in the knowledge schemes of Filipino children, compared to their Dutch counterparts. This can be a possible source of DIF: given that a group of Dutch children and a group of Filipino children have the same level of crystallized intelligence, the Dutch children might get this item correctly more often than the group of Filipino children.

Another important source of item bias or DIF could be the appropriateness of item content. Still following through with the Dutch and Filipino comparison, consider the following item in the Beck Depression Inventory (1995) which refers to loss of interest in sex:

4. Loss of Interest in Sex
   - 0 I have not noticed any recent change in my interest in sex.
   - 1 I am less interested in sex than I used to be.
   - 2 I am much less interested in sex now.
   - 3 I have lost interest in sex completely.

A group of Filipino and Dutch adults with the same level of depression might respond to this item differently because the Filipinos might find disclosing their true responses to the item as inappropriate, whereas their Dutch counterparts may not.

Thus far, the paper has pointed to the possible sources of bias that results in the possible nonequivalent translations of English-language psychological tests into Filipino and other Philippine languages. For practical purposes, it should be underscored that there are profound implications of this bias depending on how the test is used. For diagnostic purposes, there can be internal bias in terms of how an individual respondent's personality, intellectual ability, or some other individual difference variable is interpreted. Problems with any of the various forms of bias can result in incorrect assessment of an individual person in any of the possible psychological variables, thus making the assessment invalid. There can also be external bias, when scores from biased tests are used to make decisions—for example to make decisions as to whether an applicant for a school will be admitted, whether a scholarship will be given to one student and not another, whether to hire a job applicant or not, whether a juvenile offender is

able to discern what is right from wrong, or whether a spouse will be judged as being psychologically incapable of sustaining a marriage. For Filipino psychologists, the implications of such biases in the use of psychological tests are far-reaching. But what has been done so far to understand and address these issues?

## UNDERSTANDING AND ADDRESSING EQUIVALENCE AND BIAS IN PSYCHOLOGICAL TESTS IN THE PHILIPPINES

In reviewing what has been done in the Philippines related to equivalence and bias, we can refer to the published studies on these two topics. Unfortunately, there are very few published research studies to be found that address these issues of equivalence and bias. Although there are numerous research studies that involve validating psychological tests, and the number of these has increased markedly in recent years (see e.g., Magno, 2011; Olvida, 2010; Villavicencio, 2010), studies that explicitly address the issue of equivalence are few.

***Construct equivalence research in the Philippines.*** Among the various issues and concerns raised in earlier sections of the article, the area where there has been much work is the topic of construct equivalence (structural or functional equivalence). One of the earlier studies on this concern involves a very popular personality test, the Revised NEO Personality Inventory (Costa & McCrae, 1992). In a cross-cultural study that included a Filipino sample, McCrae, Costa, Del Pilar, Rolland, and Parker (1998) confirmed the five-factor personality structure of the test, thus providing evidence for the construct equivalence of the Filipino translation of the test in the Philippines and the other test versions. Another example involves comparisons of the Sense-of-Self Scales developed by McInerney, Yeung, and McInerney (2001) to measure various self-concept related constructs among students. Ganotice and Bernardo (2010) analyzed English and Filipino versions of the scale and found structural equivalence between the two versions.

These translation studies do not always find construct equivalence. When Bernardo, Posecion, Reganit, and Rodriguez-Rivera (2005) studied the translation of the Social Axioms Survey (Leung et al., 2002), they found support for the five-factor structure, but only after removing numerous items from the original scale. The Epistemological Beliefs Questionnaire (Schommer, 1998) was also translated into Filipino by Bernardo (2008), and the results of

the study found construct non-equivalence. Whereas the original questionnaire was intended to measure four interrelated constructs, the translated and English language versions only suggested two interrelated constructs.

Other researchers have explored the construct validity of English language tests with Filipino samples, without translating the tests. Some of these studies validated the original structure of the constructs intended by the original scales. For example, Ganotice and Bernardo (2010) validated the structure of constructs of three scales: Facilitating Conditions Questionnaire (McInerney, Dowson, & Yeung, 2005), Sense of Self Scale, and Inventory of Student Motivation (both by McInerney et al., 2002). All the scales were in the original English language but administered to Filipino respondents, and the respective constructs of the scales were all confirmed in the study. King, Ganotice, and Watkins (2011) also validated the hypothesized four-factor structure of Inventory of Student Motivation with Filipino students and showed that this structure was invariant with a cross-cultural sample from Hong Kong. King (2010) also studied a short version of the Academic Emotions Questionnaire (Pekrun, 2006) in English with a sample of Filipino students and likewise confirmed the intended constructs of the scales. Finally, King and Watkins (2011a & b) also confirmed the intended four-factor structure of the constructs within the Goal Orientations and Learning Strategies Survey (Dowson & McInerney, 2004) with various samples of Filipino students.

But similar investigations do not always confirm the structure of the original scales. De La Rosa (2010) administered the original English version of the achievement goals questionnaire of Elliot and McGregor (2001), which had a 2 x 2 factor structure. Instead of four interrelated constructs, he found only three factors when the psychological instrument was used with his Filipino sample. Bernardo (2001b, Bernardo, Zhang, & Callueng, 2002) also found that with a Filipino sample, the higher order constructs of the Thinking Styles Inventory (Sternberg, 1997) was not equivalent to the higher order constructs found with the original English test, but was more similar to the higher order constructs found in the Hong Kong Chinese version (Zhang & Sternberg, 1998). Zhang and Bernardo (2000) likewise found that the structure of the constructs of the Learning Process Questionnaire (Biggs, 1987) seemed to be valid with students of higher scholastic ability but not for those with low scholastic ability.

Taken together, all these studies suggest that we cannot safely assume that the English language tests are measuring the intended

constructs when used with Filipino samples, and that we can also not assume the same with Filipino translations of the scales.

This brief review of studies may suggest that there has been quite a significant amount of work done related to studying the construct equivalence of English-language tests and their translations in the Philippine setting. But if one considers the very wide range of psychological tests used in various settings like schools, companies, hospitals, psychological clinics, and testing centers, the proportion of tests that have been studied in this way probably comprises a very small percentage of the total number of tests currently in circulation.

***Measurement unit and full scalar equivalence research in the Philippines.*** If we consider studies that inquire into the measurement unit and scalar equivalence of scales, there are even fewer to refer to. Perhaps, this is because the research and statistical techniques for doing so are more complex. Indeed, to study construct equivalence or structural equivalence, a researcher only needs to establish exploratory and confirmatory factor analytic procedures, and some other basic multivariate techniques. However, more complicated procedures such as multigroup confirmatory factor analytic techniques and procedures for measuring differential item functioning are required to study the quantitative equivalence between versions of a psychological test. It should be noted that in the educational measurement field, there has been more research work done on this topic (see e.g., Pedrajita, 2009; Pedrajita & Talisayon, 2009).

The few studies that attempt to inquire into psychometric equivalence between English language tests and their Filipino translations actually do not find full scalar equivalence. For example, as part of their theoretical investigation, Bernardo and Ismail (2010) looked into the equivalence of the mastery and performance achievement goals scales in English with Filipino and Malaysian students. They found only partial invariance of the scales and not full scalar invariance or equivalence. Similarly, Ganotice, Bernardo, and King (in press) looked into the equivalence of English and Filipino versions of the Inventory of Student Motivation (McInerney, et al., 2002) and also found only partial equivalence. Note, however, that the absence of full-scalar equivalence and measurement-unit invariance is only a problem when comparing scores across cultures. If tests are being used only within one culture, it is sufficient to construct equivalence, and perhaps some degree of measurement unit equivalence, especially if the scores are being interpreted within the culture, with reference to norms within the culture, and in relation to

variables also measured within the same culture.

***Practical approaches to reducing bias.*** Other than doing research establishing the equivalence (or nonequivalence) of psychological tests, there are actually some very practical things that could be done to reduce bias. For example, when it comes to method bias, the standardization of test administration and scoring procedures is important to keep in mind. Thus, more effort should be taken to ensure that test administrators and scorers are sufficiently trained and are guided by very detailed manuals and/or protocols for test administration, scoring, and interpretation. Instructions for such tests should be as detailed as possible, anticipating possible misinterpretations by the respondents.

When it comes to translations, several experts in psychological testing have already provided (see e.g., Hambleton & De Jong, 2003; Van de Vijver & Hambleton, 1996) practical guides for managing the translation processes. These experts now propose what is called an integral management approach in translation. Test translators should not only aim to reproduce the items and instructions of one test in another language but also make judgments about whether the test is suitable and appropriate to the target culture in which the translation will be used. As it is impossible to be completely certain about issues of suitability based merely on theoretical assumptions, test translation experts now suggest that translators do pilot studies on early versions of the translation and closely document and validate these translations. Harkness (2003) suggests that the translation procedures follow several stages: a) the translation stage, which may involve translation and back translation procedures, the committee approach, or a combination of these approaches, b) qualitative pre-testing of the translations using feedback from monolingual and bilingual judges in focused group discussions or think-aloud interviews, c) reviewing the revisions based on qualitative pre-testing feedback, d) quantitative pre-testing with actual administration and tests of equivalence and bias, and e) final adjudication or decision on a final version of the test. More careful and deliberate steps in translation processes could help prevent biases at different levels.

## THE STEPS AHEAD FOR FILIPINO PSYCHOLOGISTS USING FOREIGN ENGLISH-LANGUAGE TESTS

After discussing what has been done and what can be done to ensure

that foreign made English-language psychological tests are valid for use with Filipino samples (that is, that the tests used with Filipinos are equivalent to those developed in other countries for their intended samples), we come to the question of how do we ensure that these tests are used appropriately by Filipino psychologists? Of course, this question cannot be answered with a simple step that will be a panacea to this rather huge concern. But there are a number of points that need to be emphasized in order for there to be a consensus and collective action on this issue. I list ten propositions, which vary in terms of complexity and feasibility.

***Proposition 1***. We should stop assuming that these tests are valid for all Filipino respondents. Indeed, we should not take the word of the test distributors that the tests are universally valid. Instead, we should inquire into whether the tests have been validated and tested for construct and measurement equivalence with a Filipino sample. In cases when the tests have not been validated, the psychologist using the tests should bear in mind the possible limitations associated with interpreting the scores of a test that has not been validated with the target population and seek additional convergent information using a combination of other validated tests and/or relevant assessment tools (e.g., clinical interview, etc.).

***Proposition 2.*** Teachers of psychological testing and assessment courses should give emphasis to the topics related to bias and equivalence in using psychological tests in different cultures and linguistic groups. Of course this would require that they update themselves on such topics, and teachers should be provided support and resources for doing so.

***Proposition 3.*** Researchers should do more systematic and sustained investigations on construct bias, method bias (e.g., on response biases among different Filipino samples, etc.), item bias (e.g., differential item functioning). There should be more research on establishing construct equivalence or structural validity of foreign made tests with Filipino respondents, whether these tests are translated or not. If possible, there should also be research investigating measurement unit and full scalar invariance, which would allow comparability with norms in other countries or cultures.

***Proposition 4.*** Researchers should prioritize investigations on widely used personality and intelligence tests used by human resource recruiters & personnel, schools and admissions officers, the legal system (annulment cases, juvenile justice system), medical practitioners, and so on. Indeed, much of the research done so far

involve psychological tests used for research. These types of research should be directed at the tests that are actually more widely used by practitioners.

***Proposition 5.*** Researchers should work on more translations for different Philippine ethnolinguistic groups and do research on the bias and equivalence of the translations. Filipino is just one of the many languages spoken in the Philippines. It may not be practical to do translations in all the Philippine languages, but it may be necessary to develop and study translations in the major languages such as Cebuano, Ilonggo, Ilocano, Bicol, Waray, and others.

***Proposition 6.*** Researchers should publish and disseminate their results and also ensure that their findings and test translations/adaptations are made available to various users and practitioners. It is not necessary to make these translations available for free; instead, researchers and translators should find ways to provide practitioners and other researchers access to these materials, even for a fee.

***Proposition 7.*** Practitioners, particularly those working with rural and/or less-educated clients, should deliberately seek for test translations and adaptations for their own practice. They should not rely on what is easily available, unless they are certain of their validity. Indeed, applying foreign made tests in an uncritical manner could be the foundation upon which many invalid diagnoses and recommendations are made.

***Proposition 8.*** Practitioners should systematically observe and document their observations and experiences involving English language versions of the test with less educated populations. The experiences of these practitioners are actually rich resources upon which further improvements can be made on the psychological tests being used; but these experiences need to be systematically documented, shared, and analyzed in ways that aim to develop better tests for use with particular clients.

***Proposition 9.*** Practitioners should systematically observe and document their observations and experiences involving translations of the tests. For those who actually translate the tests or use translations made by other people, the qualitative validation is also as important. Again the need to systematically document, and share these experiences should be emphasized.

***Proposition 10.*** As there will always be non-translatable constructs, Filipino psychologists should also invest more effort, or at least provide support for developing indigenous psychological tests. This support should extend to the validation, marketing, distribution,

and application of the tests.

There are a lot of challenges facing Filipino psychologists who use psychological tests in their profession. The actions required to face these challenges need to involve the collective action of the whole community of psychological researchers, practitioners, and educators. Practitioners need researchers to develop translations that are valid for use with different Filipino clients. Researchers need to collaborate with practitioners for the qualitative and quantitative validation of translations. Practitioners can provide researchers with rich personal observations on how the tests are applied and how they work. Researchers can guide practitioners on how to reflect on and systematically document their experiences using the test. Educators should create awareness and knowledge related to the issue of validity of using foreign made tests with Filipino clients, especially since their students will be the future psychology practitioners. Indeed, Filipino psychologists need not lose their clients in bad translation if they work together in ensuring that attention and effort is invested on guaranteeing the validity of tests and their translations.

## REFERENCES

Bernardo, A. B. I. (1996). Task-specificity in the use of words in Mathematics: Insights from bilingual word problem solvers. *International Journal of Psychology, 31* (1), 13-28.

Bernardo, A. B. I. (1998). Language format and analogical transfer among bilingual problem solvers in the Philippines. *International Journal of Psychology, 33* (1), 33-44.

Bernardo, A. B. I. (1999). Overcoming obstacles to understanding and solving word problems in Mathematics. *Educational Psychology, 19* (2), 149-163.

Bernardo, A. B. I. (2001a). Asymmetric activation of number codes in bilinguals: Further evidence for the encoding-complex model of number processing. *Memory & Cognition, 29*, 968- 976.

Bernardo, A. B. I. (2001b). Do Asians and Americans think differently? Thinking styles among Filipino, Hong Kong, and American college students. *Philippine Journal of Psychology, 34* (2), 27-43.

Bernardo, A. B. I. (2002). Language and mathematical problem solving among bilinguals. *The Journal of Psychology, 136*, 283-297.

Bernardo, A. B. I. (2005). Language and modeling word problems in Mathematics

among bilinguals. *The Journal of Psychology, 139*, 413-425.

Bernardo, A. B. I., & Calleja, M. O. (2005). The effects of stating problems in bilingual students' first and second languages on solving mathematical word problems. *The Journal of Genetic Psychology, 116*, 117-128.

Bernardo, A. B. I. (2008). Exploring epistemological beliefs of bilingual Filipino preservice teachers in the Filipino and English languages. *The Journal of Psychology, 142*, 193-208.

Bernardo, A. B. I. & Ismail, R. (2010). Social perceptions of achieving students and achievement goals of students in Malaysia and the Philippines. *Social Psychology of Education, 13*, 385-407.

Bernardo, A. B. I., Ouano, J. A., & Salanga, M. G. C. (2009). What are academic emotions? Insights from Filipino bilingual students' emotion words associated with learning. *Psychological Studies, 54*, 34-42.

Bernardo, A. B. I., Posecion, O. T., Reganit, A. R., & Rodriguez-Rivera, E. (2005). Adapting the social axioms survey for Philippine research: Validating Filipino and English versions. *Philippine Journal of Psychology, 38*(2), 77-100.

Bernardo, A. B. I., Zhang, L.F., & Callueng, C. M. (2002). Thinking styles and academic achievement among Filipino students. *The Journal of Genetic Psychology, 163*, 149-163.

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

Church, T. A. (1987). Personality research in a non-Western setting: The Philippines. *Psychological Bulletin, 102*, 272-292.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Dowson, M., & McInerney, D.M. (2004). The development and validation of the Goal Orientation and Learning Strategies Survey (GOALS-S). *Educational and Psychological Measurement, 64*, 290-310.

Enriquez, U.G. (1992). *From colonial to liberation psychology*. Quezon City: University of the Philippines Press.

Gass, S. M., & Varone, E. M., (1991). Miscommunication in normative speaker discourse. In N. Coupland, H. Giles, & J. M. Wiemann (Eds.), *Miscommunication and problematic talk* (pp. 121-145). Newbury Park, CA: Sage.

Ganotice, F., & Bernardo, A. B. I. (2010). Validating the factor structures of the English and Filipino versions Sense of Self Scales. *Philippine Journal of Psychology, 43*, 81-99.

Ganotice, F. A., Bernardo, A. B. I., & King, R. B. (in press). Validating two language versions of the Inventory of School Motivation among Filipinos. *Journal of Psychoeducational Assessment*.

Goodwin, R., & Lee, I. (1994). Taboo topics among Chinese and English friends: A cross-cultural comparison. *Journal of Cross-Cultural Psychology, 25*, 325-338.

Guazon-Lapeña, M.A., Church, A.T., Carlota, A.J., & Katigbak, M.S. (1998). Indigenous personality measures: Philippine examples. *Journal of Cross-Cultural Psychology, 29* (1), 249-270.

Grimm, S. D., & Church, A. T. (1999). A cross-cultural study of response biases in personality measures. *Journal of Research in Personality, 33*, 415-441.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10* (3), 229-244.

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

Hambleton, R. K., & De Jong, J. H. A. L. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20*, 127-134.

Harkness, J. (2003). Questionnaire translation. In J. Harkness, F. J. R. Van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35-56). Hoboken, NJ: Wiley.

Hoffman, C., Lau, I., & Johnson, D. R. (1986). The linguistic relativity of person cognition. *Journal of Personality and Social Psychology, 51*, 1097-1105

Huang, C. D., Church, A. T., & Katigback, M. S. (1997). Identifying cultural difference in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology, 28*, 192-218.

Hui, C. H., & Triandis, H. C. (1989). On the empirical identification of dimensions for cross-cultural comparison. *Journal of Cross-Cultural Psychology, 20*, 296-309.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

King, R. B. (2010). What do students feel in school and how do we measure them? Examining the psychometric properties of the S-AEQ-F. *Philippine Journal of Psychology, 43*, 161-176.

King, R. B., Ganotice, F. A., & Watkins, D. A. (2011). Cross-cultural validation of the Inventory of School Motivation (ISM) in the Asian setting: Hong Kong and the Philippines. *Child Indicators Research.* Published Online First 10 June 2011 doi: 10.1007/s12187-011-9117-3.

King, R. B., & Watkins, D. A. (2011a). Cross-cultural validation of the five-factor structure of social goals: A Filipino investigation. *Journal of Psychoeducational Assessment,* Published Online First: 1 August 2011. doi:10.1177/0734282911412542.

King, R. B., & Watkins, D. A. (2011b). The reliability and validity of the Goal Orientation and Learning Strategies Survey (GOALS-S): A Filipino investigation. *The Asia-Pacific Education Researcher, 20*, 579-594.

Leung, K., Bond, M. H., De Carrasquel, S. R., Muñoz, C., Hernandez, M., Murakami, F.,Yamaguchi, S., Bierbrauer, G., & Singelis, T. M. (2002). Social axioms: The search for universal dimensions of general beliefs about how the world functions. *Journal of Cross-Cultural Psychology, 33*, 286–302.

Magno, C. (2011). Validating the Academic Self-regulated Learning Scale with the Motivated Stategies of Learning Questionnaire (MSLQ) and Learning and Study Strategies Inventory (LASSI). *The International Journal of Educational and Psychological Assessment, 7*(2), 56-73.

Marian, V. A., & Fausey, C. M. (2006). Language-dependent memory in bilingual learning. *Applied Cognitive Psychology, 20*, 1025-1047.

Marian, V., & Kaushanskaya, M. (2004). Self-construal and emotion in bicultural bilinguals. *Journal of Memory and Language, 51* (2), 190–201.

Marian, V., & Neisser, U. (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General, 129* (3), 361–368.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among Hispanics. *Journal of Cross-Cultural Psychology, 23,* 498-509.

McCrae, R.R., Costa, P.T., del Pilar, G.H., Rolland, J.P. & Parker, W.D. (1998). Cross cultural assessment of rhe Five-Factor model. *Journal of Cross-Cultural Psychology, 29*, 171-188.

McInerney, D. M., Yeung, S. Y., & McInerney, V. (2001). Cross cultural validation of the Inventory of School Motivation (ISM). *Journal of Applied Psychological Measurement, 2*, 134-152.

Olvida, C. F. (2010). Development of Ethnic Identity Measure (EIM) for Cordillera indigenous people. *Educational Measurement and Evaluation Review, 1,* 78-89.

Pedrajita, J. Q. (2009). Using logistic regression to detect biased test items. *The International Journal of Educational and Psychological Assessment, 2*, 54- 73.

Pedrajita, J. Q., & Talisayon, V. M. (2009). Identifying biased test items by differential item functioning analysis using contingency table approaches: A comparative study. *Education Quarterly, 67*, 21-43.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. E*ducational Psychology Review, 18*, 315-341.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24*, 737-756.

Schommer, M. (1998). The influence of age and schooling on epistemological beliefs. *British Journal of Social Psychology, 68*, 551–562.

Smith, P. B. (2004). Acquisecent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology, 35*, 50-61.

Spielberger, C. D. (1988). *State-trait anger expression inventory* (Research ed.). Professional manual. Odessa, FL: Psychological Assessment Resources.

Sternberg, R. J. (1997). *Thinking styles.* New York: Cambridge University Press.

Tsai, J. L. (2007). Ideal affect: Cultural causes and behavioral consequences. *Perspectives on Psychological Science, 2*, 242-259.

Tsai, J. L., Knutson, B., & Fung, H. H. (2006). Cultural variation in affect valuation. *Journal of Personality and Social Psychology, 90*, 288-307.

Uchida, Y., Norasakkunkit, V., & Kitayama, S. (2004). Cultural constructions of happiness: Theory and empirical evidence. *Journal of Happiness Studies, 5*, 223-239.

Van de Vijver, F., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89-99.

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

Van de Vijver, F., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*, 29-37.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de psychologie appliqué, 54*, 119-135.

Villavicencio, F. T. (2010). The development and validation of the Emotionality in Teaching Scale (ETS). *Educational Measurement and Evaluation Review, 2,* 6-23.

Watkins, D., & Cheung, S. (1995). Culture, gender, and response bias: An analysis of responses to the Self-Description Questionnaire. *Journal of Cross-Cultural Psychology, 26*, 490-504.

Watkins, D., & Gerong, A. (1999). Language of response and the spontaneous self-concept: A test of the cultural accommodation hypothesis. *Journal of Cross-Cultural Psychology, 30*, 115-121.

Werner, O., & Campbell, D. T. (1970). Translating, working through interpreters, and the problem of decentering. In R. Naroll, & R. Cohen (Eds.), *A handbook of cultural anthropology* (pp. 398-419). New York: American Museum of Natural History.

Zhang, L.F., & Bernardo, A. B. I. (2000). Validity of the Learning Process Questionnaire with students of lower academic attainment. *Psychological Reports, 87*, 284-290.