SA Journal of Industrial Psychology, 2004, 30 (4), 8-15
SA Tydskrif vir Bedryfsielkunde, 2004, 30 (4), 8-15

PLANNING A PSYCHOLOGICAL TEST IN THE MULTICULTURAL SOUTH AFRICAN CONTEXT

CHERYL D. FOXCROFT
Psychology Department
University of Port Elizabeth
cheryl.foxcroft@upe.ac.za

ABSTRACT

Given the dearth of psychological tests developed from a multicultural perspective in South Africa, this article contemplates how to set about developing a test plan for a test to be used in a multicultural context. Conventional components of a test plan are outlined and elaborated on in terms of the issues that need to be considered when planning a test from a multicultural perspective. Cross-cultural test design issues are illustrated by providing examples or by referring to research findings. Finally, some thoughts are provided related to who should develop the test plan when a multicultural test is to be developed.

OPSOMMING

In die lig van die feit dat baie min toetse in Suid-Afrika ontwikkel is vanuit 'n multikulturele perspektief, poog hierdie artikel om die stappe uiteen te sit vir die ontwikkeling van 'n toetsplan vir 'n toets wat in 'n multikulturele konteks gebruik sal word. Die gebruiklike komponente van 'n toetsplan word uiteengesit en uitgebrei in terme van aspekte wat in ag geneem moet word wanneer 'n toets vanuit 'n multikulturele perspektief beplan word. Kruis-kulturele ontwerp aspekte word geïllustreer deur gebruik te maak van voorbeelde of om na navorsing te verwys. Ten slotte word gedagtes gewissel aangaande wie die toetsplan moet ontwerp as 'n multikulturele toets ontwikkel word.

Despite the multicultural nature of our society, psychological test development in South Africa has historically been characterised by the development of tests for separate cultural and/or language groups (Claassen, 1997; Foxcroft, 1997). While past apartheid policies and legislation shaped the way in which test development was approached until the 1990s, it is somewhat disturbing to note that subsequent to apartheid's demise, very few new culturally relevant tests have been developed that can be applied to a diverse range of cultural and language groups in South Africa. Among the reasons for this is that there is a dire shortage of test development capacity in South Africa at present.

The development of psychological tests is highly specialised and should be undertaken by teams of experienced measurement experts (Robertson, 1990). Until the 1990s, the Human Sciences Research Council (HSRC) almost exclusively developed or adapted the psychological tests used in South Africa. However, while there was considerable test development expertise at the HSRC, little emphasis was placed on training test developers in postgraduate psychology programmes, which meant that test development capacity was not being built among younger researchers and psychologists. The situation was further compounded when, during the process of transforming and restructuring itself in the mid 1990s, the development of psychological tests ceased to be a prime focus of the HSRC. Many of the experienced test developers retired, were redeployed to other positions in the organisation, took up positions at academic institutions, or emigrated. Today there are only a handful of test developers employed by the HSRC. Fortunately, companies such as Psytech and SHL, among others, have taken up the challenge of adapting internationally developed psychological tests and norming them for the South African context, and pockets of test development expertise have developed among research teams at universities (often in collaboration with international test development and cross-cultural experts). However, it remains unfortunate that at this critical moment when psychological test development stands at the threshold of a new era in which new tests should be developed from a multicultural rather than a monocultural perspective, there is a critical shortage of experienced test developers in South Africa. While undergraduate and postgraduate programmes in psychological test development need to be instituted as a matter of urgency to ensure that there will be a supply of test developers in the future, efforts also need to be made to skill current psychologists and researchers in the art and science of psychological test development. With this in mind, the focus of this article will be on how to plan a psychological test that will be applied in a multicultural and/or multilingual context in South Africa.

The process of developing a psychological test is a complex and lengthy one, which has been well documented in standard psychometric texts (e.g., Foxcroft & Roodt, 2001; McIntire & Miller, 2000) but aspects related to the planning of a psychological test are not always sufficiently emphasised and sometimes not even mentioned (e.g., Kaplan & Saccuzzo, 1997; Murphy & Davidshofer, 1998). There are two reasons why a great deal of time and thought should be put into the planning phase of the test development process. First, when the test is to be used in a multicultural context, attention needs to be paid to the cultural relevance (and potential bias) of the test right from the planning and design phase instead of only being sensitive to cultural aspects from the item writing phase onwards. Second, given that we do not have a long history of developing culturally appropriate tests applicable to diverse groups in South Africa, test developers need to grapple with basic issues such as what methods of test administration might be appropriate or inappropriate for certain cultural groups and what language to develop the test in, for example. More time thus needs to be spent in the planning phase exploring and critically considering test design issues. It is for these reasons that the present article will attempt to unpack the aspects that need to be considered when planning a psychological test, which is intended for use in multicultural contexts.

DEVELOPING A MULTICULTURAL TEST PLAN

Typically, a test plan consists of the following aspects: (a) specifying the purpose and rationale for the test as well as the intended target population, (b) defining the construct (content domain) and creating a set of test specifications to guide item writing, (c) choosing the test format, (d) choosing the item format, and (e) specifying the administration and scoring methods (McIntire & Miller, 2000, Roberston, 1990). However, when a test is developed for a multicultural target population,

Requests for copies should be addressed to: CD Foxcroft, Department of Psychology, University of Port Elizabeth, PO Box 1600, Port Elizabeth, 6000

some expansion and elaboration of the typical aspects of a test plan is required to ensure that cross-cultural aspects are built into the fabric of its design. With this in mind, each of the aspects of the test plan identified by McIntire and Miller (2000) will be elaborated on from a multicultural perspective in this article.

Readers should take note that although the aspects of the plan are logically ordered, there is a dynamic interplay between the various aspects that will often result in the developer revising a decision made about one or other aspect. For example, the construct to be tapped is indicated when stating the purpose of the test, but its operationalisation and meaningfulness for different cultural groups is considered at a later stage. It might be that after a thorough investigation into how different cultural groups view the construct and for which cultural groups it is appropriate, the purpose of the test as well as the intended target population might have to be revised or refined.

Each of the aspects of the test plan will now be elaborated on.

Purpose of the test

McIntire and Miller (2000) assert that a statement of the purpose of a test should include an indication of the construct to be tapped (e.g., personality, self esteem, intelligence) as well as how the outcome (results) of the test will be used (e.g., to predict a performance criterion, to compare individuals to a norm group, to make a diagnosis). In addition, the fact that it is intended that the test will be used in the multicultural South African context should also be included in the purpose statement. The rationale for the latter is that in the same way as the nature of the construct to be tapped and the intended use of the test have implications for the development of the test specifications, so too will the fact that the test is to be used in multicultural settings guide the planning related to the design of the test.

It needs to be kept in mind that the South African society has a diversity of cultures in which appreciation for the culture of origin exists alongside variations in acculturation towards a Western norm (Claassen, 1997). In view of the varying cultural distances between cultures and subcultures in South Africa and the influence that culture exerts on behaviour (and hence test performance), Claassen (1997, p. 306) asserts that a "realistic objective in cross cultural testing is rather to construct tests that presuppose only experiences that are common to different cultures". To this, Retief (1992) adds that not only should multicultural tests yield an index of commonality but also an index of difference. By this is meant that a multicultural test could have two components. One that taps aspects of the construct that are common across cultures and one that taps aspects of the construct that are unique to each group. The former could be used when cross-cultural comparisons are made, while the latter can be used to get a fuller, more culturally contextualised picture of the individual being assessed. Consequently, if the fact that the test being developed for a multicultural context is written into the test plan, the test developer will be alerted to the fact that the test plan will also have to include ways of identifying aspects of the construct that are common to and unique to the various groups included. How this can be achieved will be discussed in a later section when ways to identify the meaning and meaningfulness of constructs across cultural groups are discussed.

Characteristics of the intended target population and design implications

The test developer needs to list the characteristics of the intended test-takers and especially those characteristics of test-takers that could impact on how they will respond to the test items as well as their performance on the test. Some of the more important characteristics that might need to be considered when developing a test for a multicultural South African context will be highlighted. Age is normally one of the critical aspects of the intended target population that needs to be stated because whether the test is to be developed for children, adolescents or adults will influence the nature of the test format and items, for example.

Educational status is another critical and complex variable to consider when delineating the target population in the multicultural South African context. Schooling experiences have an impact on the proficiency to read, write, and work with numbers as well as on higher order cognitive development in that they "influence how people think or the reasoning strategies they use, how they approach problems, their ability to deal with issues in an independent way, as well as to work accurately and quickly" (Grieve, 2001, p. 325). However, given the historical disparities in the provision of education among the various cultural groups in South Africa, it needs to be kept in mind that people who have experienced a poorer quality of education have not had the same opportunities to develop academic proficiencies and cognitive skills as those from more advantaged educational backgrounds. Furthermore, not only has the quality of educational provision differed for various cultural groups but educational provision in rural areas has been markedly inferior to that provided in urban areas. It is thus not surprising that urban-rural differences have been found on cognitive tests (e.g., Freeman, 1984; Schepers, 1974).

Consequently, if the intended target population is to be defined in terms of covering a range of school grades as would be the case if a career maturity measure is to be developed for adolescents for example, test developers need to be mindful of the fact that the varied quality of the schooling that potential test takers have been exposed too could differentially impact on both their way of responding to the test as well as their test scores. In such an instance, test developers may, for example, want to consider whether their test plan needs to include a questionnaire to gather information on the quality of the schooling of test-takers who participate in the standardisation of the new test. Based on this information, it might be indicated in the plan that consideration will be given to exploring whether, in the case of a normative test, separate norms should be developed for test-takers from advantaged and disadvantaged schooling backgrounds and possibly even for urban and rural test-takers.

Nell (1994) argues that language is the most critical moderator variable of test performance, especially in our multilingual society. If a test is administered in a language in which test-takers are not proficient, it will be difficult to untangle whether poor performance on the test is due to language or communication difficulties or to the fact that test-takers have a low level of the construct being assessed. In this regard, for example, studies by Meiring, Van de Vijver, Rothmann, and Barrick (in press) and Abrahams and Mauer (1999) revealed that understanding English concepts in personality tests was problematic for black test-takers and impacted on the construct comparability of the tests across cultural groups.

According to the International Test Commission's Guidelines for Adapting Educational and Psychological Tests (Hambleton, 1994, p. 232), test "developers/publishers should provide evidence that language use in the directions, rubrics, and items themselves ... are appropriate for all cultural and language populations for whom the instrument is intended". Following from this, on the one hand, if a test will only be developed in one language but the intention is to use it with multilingual test-takers, the test plan must specify how the language proficiency of test-takers with respect to the test language will be determined and what level of proficiency will be required so as to ensure that the test results are not contaminated by language factors. On the other hand, test developers may prefer to develop a multilingual test and in this instance the test plan should indicate in which languages the test would be available. If multilingual versions of the test are to be developed, it should be specified in the test plan in which language the test content will initially be developed (source language) before it is translated to the other language versions. According to Hambleton (1994), one of the causes of the development of poor quality cross-cultural tests "is that the source language version is often unnecessarily complicated and therefore quite difficult to translate accurately" (p. 234). A further issue raised by Hambleton (1994) is that the concepts and idiomatic expressions used in the source language version do not have equivalents in the other languages into which the test is to be translated. A team of cultural, content, and language experts should thus be assembled right from the planning phase to scrutinise the content being developed so as to minimise potential translation difficulties.

Furthermore, consideration should be given during the planning phase as to whether the different language versions of the test will be produced separately or whether the final test will be produced in a bilingual or multilingual format in that the different language versions will be presented side by side. As regards which option is the better one, it should be noted that two recent South African studies, one with a multicultural, multilingual university sample (Mochela & Seymour, 2003) and one with black primary school Xhosa-speaking learners from a lower socio-economic background (Els, 2004), revealed that test-takers whose first language was not English nonetheless preferred to complete a group paper-based test either in English or by using a combination of the English and their home language versions. A relatively small percentage used only their home language version. It would thus seem that it might be wise for test developers to plan to produce tests that have various language versions in a bilingual (e.g., English and either Afrikaans or an African language) or a multilingual format.

Where more than one language version of the test is to be developed, the test plan also needs to specify which methodologies will be used to systematically gather judgemental and empirical evidence that the different language versions are equivalent (Hambleton, 1994). Readers are referred to Brislin (1970), Bracken and Barona (1991), Hambleton (1994), Kanjee (2001), Van de Vijver and Leung (1997) and Van Ede (1996) for a detailed outline of the issues and methods related to translating tests into different languages and determining the equivalence of the translations.

Test content is closely aligned to the cultural group for which the test is developed as well as the cultural background of the test developer. As this article is focusing on developing a test plan for a multicultural test, the assumption is that the test being developed should be applicable to at least the major cultural groups in South Africa, namely black, coloured, Indian, and white. Given the importance of the implications of culture for the test development plan, this aspect will be comprehensively addressed in the next two sections of this article.

Sometimes tests are developed for special populations (e.g., mentally handicapped, hearing impaired, visually impaired, learning disabled, physically disabled). Given the specialised nature of such tests, it is not possible to discuss how a test plan would be developed for such target populations in this article. Readers are referred to Anastasi and Urbina (1997) and Luiz and Jansen (2001) for thoughts related to the development and adaptation of tests for special populations.

Defining the construct and its cross-cultural meaningfulness

Traditionally, when a construct is defined in a test plan, test developers consult a variety of sources to assist them in concisely defining and operationalising it in terms of observable, measurable behaviours. A few examples of sources are provided here by way of illustration. In clinical settings, an extensive literature review is normally conducted and any available tests related to the construct of interest are reviewed. In organisational contexts, a job analysis can assist in identifying the knowledge, skills, abilities and other characteristics (KSAOs) required to perform a job successfully. These KSAOs represent the constructs to be tapped by the criterion-referenced test to be developed. As the job analysis also describes the tasks performed on the job in terms of observable and measurable behaviours associated with the KSAOs, it helps to operationalise the constructs to be tapped (McIntire & Miller, 2000). In educational settings, the learning outcomes and curricula provide the basis for identifying and delineating the construct(s) to be tapped and to operationally define them. Furthermore, when an educational or ability test is to be developed, the Taxomony of Educational Objectives (Bloom, Engelhart, Furst, Hill & Krathwohl 1956; Krathwohl, Bloom & Masia, 1956), which specifies and categorises behaviours associated with higher order cognitive processes and affective functioning, is often a useful source to consult when trying to operationally define cognitive and affective constructs (Robertson, 1990).

When planning to develop a test for use with a variety of cultures and language groups, however, there is a crucial step that needs to be undertaken before the construct is defined and operationalised using the sources and resources outlined above. In view of the differences that exist between various cultural and language groups with respect to their traditions, customs, values and different world views, the same construct could be interpreted and understood in very different ways in various cultural and language groups (Hambleton, 1994). One of the often quoted examples in this regards relates to the construct of intelligence, which is associated in Western cultures with being mentally sharp and quick thinking while in Eastern cultures it is associated with being thoughtful and reflective (i.e., wise and slow to respond). It is thus critical that the meaning and understanding of the construct in the various cultural groups for whom the test is intended should be explored in the planning phase otherwise construct bias may be built into the test from the start.

Furthermore, according to the International Guidelines for Test Use (ITC, 2001), when a test is to be used with test-takers from different groups, it should be ensured that the constructs being assessed are meaningful for each group. Thus, the construct to be tapped should not only be explored with respect to how different cultural and language groups conceptualise it, but also in relation to whether the construct is meaningful for them. The wisdom of developing a test for a group who do not perceive the construct to be important or of value for them would be questionable.

What should be specified in the test plan related to how test developers could explore the meaning and meaningfulness of the construct to be tapped in various cultural groups? The meaning, meaningfulness and cross-cultural appropriateness of the construct that the test intends to tap could be explored through focus groups and individual interviews with key informants from the various cultural and language groups. For example, if the intention is to develop a personality test, interviewees could be asked to share their understanding of what the term "personality" means and also to describe a good friend, relative, boss, and so forth in terms of the characteristic way in which they behave, think, and respond to complex situations. Furthermore, interviewees could be asked to describe how two public figures (e.g., ex-President Nelson Mandela and President Thabo Mbeki) are similar and different. A question could also be posed as to whether interviewees believe that it is important to be able to evaluate a person's personality and to what use such information can be put. By synthesising the interview information, test developers should be able to ascertain whether there is a shared (common) understanding of the construct, which could form the basis of the definition of the construct in the test plan, whether common dimensions of the construct emerge across the groups that could be built into the test specifications, and whether the construct appears to be a meaningful one for all or some of the groups.

Furthermore, cross-cultural studies could be consulted to see whether any of the findings shed light on whether or not there is a shared understanding of the construct to be measured by the various cultural groups that comprise the target population for the test, whether existing tests that tap the construct have been found to demonstrate cross-cultural construct equivalence, and, if not, what explanations were offered. For example, Meiring, Van de Vijver, Rothmann, and Barrick (in press) found that various scales of the 15PQ+ revealed construct bias for a sample of black, coloured, Indian and white participants. To explore this finding, they consulted an expert group of black psychologists and African language experts. The experts found that there were difficulties in understanding some of the English words and idiomatic expressions used in the item content and difficulties in understanding the meaning of qualifying words used (e.g., "rarely", "generally", "less"). The experts further pointed out that some of the construct dimensions could have stronger political connotations for black respondents than for respondents from other groups (e.g., "Conventional-Radical"), which could account for some of the variability in the responses among the groups. These findings provide valuable pointers for test developers who wish to design a test plan for a new multicultural personality test in South Africa with respect to both the language in which the content is developed as well as the differential meaning attached to some of the construct dimensions.

Content development and test specifications

Once the construct to be assessed has been defined and operationalised from a multicultural perspective, a decision needs to be reached regarding what approach will be employed to guide the development of the test content and specifications. Generally, the development of test specifications and content is guided by whether theory-based, empirical, or criterion-referenced methods will be used to develop the test. This section will particularly address these issues in test development and the subsequent implications of each for the development of the test plan.

Theory-based approach. According to Murphy and Davidshofer (1998), a theory-based (or rational) approach has traditionally been used to guide the development of test content and specifications, especially for tests used in clinical settings. Here test developers draw on an existing theory to guide the development of the test. For example, when Das and Naglieri (1994) constructed the *Cognitive Assessment System* (CAS), they based its development on a model (or theory) of cognitive functioning called Planning, Attention, Simultaneous, and Successive processing (PASS) (Naglieri & Das, 1990). Consequently, the structure of the CAS is such that it consists of four scales (i.e., Planning, Attention, Simultaneous Processing, and Successive Processing) so as to tap the four main elements of the theory on which it is based.

It should be noted that when a test is developed on the basis of a theory, test developers usually subject the test items to a factor analysis at a later stage during the test development process to empirically verify that the structure of the test closely matches the theoretical model on which it is based. The advantage of a theoretically grounded test is that assessment practitioners can draw on the theory to make predictions about behaviour. For example, researchers have found that reading performance can be predicted by performance on successive, simultaneous, and planning tasks (e.g., Das, Bisanz & Mancini, 1985). Consequently, it could be predicted that if a child performs poorly on the successive, simultaneous, and planning tasks of the CAS, the development of reading problems could be anticipated and early remedial intervention could be instituted. The latter represents a further advantage of basing a test on a theory in that there is often a close link between the test results and suggestions for intervention. Knowledge of how successive processing develops and can be stimulated, for example, could lead to concrete intervention possibilities being suggested if a child performs poorly on successive processing tasks, and so forth.

However, rationally or theoretically derived tests also have severe shortcomings (Murphy & Davidshofer, 1998). The most obvious one being that the validity of the test is closely linked to the validity of the theory on which it is based. If the theory is not substantiated, then the validity of a test based on it will be doubtful. For example, although some empirical validation has been found for the PASS theory, it appears as if planning and attention processes are so linked to and dependent on each other that they load onto one factor when tasks that tap these processes are factor analysed. Consequently, it is not surprising that factor analytic studies of the CAS have suggested that the CAS probably taps three as opposed to four cognitive processing dimensions, which has implications for its construct validity.

Furthermore, when basing the development of a multicultural test on a theory, evidence first needs to be gathered that the theory is appropriate and relevant for the various cultural groups. This presents something of a problem as the majority of the theories that have been generated in psychology have been generated from a Western perspective. Consequently, it cannot simply be assumed that they will be applicable in the multicultural South African context. Test developers will thus firstly have to consult research studies, or given the lack of theory building research in South Africa, conduct their own studies to investigate whether the theory can be substantiated here, or whether it first needs to be modified. If no applicable, substantiated theory can be found, test developers will have to generate a theory from scratch and the test plan will have to make provision for this. The information gathered from the various cultural groups concerning the meaning and meaningfulness of the construct, could also serve a theory-generating purpose. By way of illustration, in the previous section ways in which the meaning and meaningfulness of the personality construct could be explored across cultures were provided. By applying qualitative content analysis methods to the personality descriptions gathered, descriptions that share something in common could be grouped or clustered and the cluster groupings could then be synthesised to develop an implicit South African personality theory. Taylor and Boeyens (1991) support the use of such an approach and add that the Repertory Grid Technique or a variation of it might be a useful technique to use to generate an implicit theory of personality.

While South African test developers might be intimidated at the thought of contextually modifying or generating a theory on which to base a test, it is hoped that they will accept the challenge. There is an urgent need for tests to be developed on the basis of valid, applicable theories for our multicultural context. For too long we have been content to develop or adapt tests based on Western theories that have not been verified here, which has reduced the accuracy of the test results obtained and the quality of the resultant decisions made on the basis of the results.

Empirical approach. In this approach, a set of possible items is administered to clearly defined criterion or contrasting groups and items that differentiate between the groups statistically are included in the final version of the test (Murphy & Davisdhofer, 1998). The Minnesota Multiphasic Personality Inventory (MMPI) is an example of a test that has been developed using the criterion group strategy. Items were included in the final version of the MMPI on the basis of their ability to discriminate between psychiatric patients and normal people. The drawback of this test development approach is that it is often difficult to understand the psychological significance of an item included on the basis of its discriminatory power as well as the theoretical reasons for the differences between the criterion groups. A further drawback of the empirical approach to scale development in the multicultural South African context relates to the difficulty of defining criterion or contrasting groups in such a way that the groups do not differ on other factors that could impact on performance on the item. For example, if the aim is to identify cognitive processing items that differentiate between good and poor readers, the test developer will not simply be able to delineate the groups in terms of their reading performance at school. As was pointed out previously, the quality of the schooling received will have to be considered and criterion groups may have to be formed on the basis of good and poor readers who have received a high, acceptable, or poor quality of schooling. In addition, whether the reading performance was achieved in the learners' home language or in their second language would also have to be taken into account when delineating the criterion groups. Thus, if the empirical approach is to be used to develop a multicultural test in South Africa, much consideration will have to be given as to how to delineate the criterion or contrasting groups so that they only differ on the criterion (e.g., reading performance) of interest and not on other variables that could impact on test performance.

A variation of the empirical approach to test development entails the use of factor analysis in the item selection and scale development process. While Murphy and Davidshofer (1998) comment that tests developed on the basis of this method usually have psychometric properties that are superior to those derived from theories, the essence of what is measured by a factor analytically derived scale is difficult to capture and its relation to theoretical and clinically usefully concepts may be unknown. If factor analysis is used to develop tests in a multicultural context, a two-step procedure is followed. In the first step the covariance matrices of all the cultural groups are combined in order to make a single, pooled data matrix. Factors derived from this pooled covariance matrix define the global solution, with which the factors obtained in the separate cultural groups are compared (after target rotation to the pooled solution) in the second step. The agreement between the solution for a cultural group and the global solution is then evaluated by means of a factor congruence coefficient. Only if the factor structures of the cultural groups are found to be essentially similar to each other and to that of the total sample, can the empirically derived scales be applied with confidence cross-culturally.

Criterion-referenced approach. According to Hambleton and Zenisky (2003), criterion-referenced tests (CRTs) provide information about a test-taker's performance with respect to a clearly defined domain of content (or construct) and/or behaviours. The scores obtained on CRTs are compared with established performance standards associated with the content domain or behaviour evaluated. This is in contrast to norm-referenced tests where test score norms are used to compare or rank-order test-takers on the construct being assessed. CRTs are widely used in education (e.g., to evaluate and monitor learner performance in relation to learning outcomes, curriculum goals and instructional approaches) and in industry (e.g., to evaluate job competence, for selection purposes, and to identify training needs).

If a test developer decides that, given the purpose of the test, a CRT should be developed, cognisance should be taken of the fact that this has implications for the development of the test plan as well as the process of developing the test (Hambleton & Zenisky, 2003). When a CRT is developed the content domains or behaviours of interest need to be rigorously defined as the resulting test scores are referenced back against the appropriate content domain or behaviour. This usually entails surveying curriculum frameworks, specific learning outcomes and associated assessment criteria, or undertaking a job analysis, for example. Based on this, a final set of content standards or behavioural objectives can be selected. Thereafter, item specifications need to be prepared for each content standard or objective, as this will enable the test developer to clearly lay out the content or behaviours to be covered by the test. An item bank is then developed and items are reviewed in terms of whether they meet the content specifications and whether they are well written. Flawed items are removed from the item bank before it can be used to generate CRTs. Readers are referred to Hambleton and Zenisky (2003) for a detailed discussion of the 12 steps involved in developing a CRT.

From a cross-cultural perspective, the panel of experts who develop curriculum frameworks and learning outcomes or who perform job analyses should represent a mix of cultures. This will ensure that the framework used to guide the development of the content specifications would have included input from various cultural groups. More importantly, however, a review panel that will qualitatively evaluate whether the items meet the content specifications should be widely representative of the cultural groups that the test will be applied to. Other than focusing on the content validity of the items, the panel should indicate whether the items are free from stereotyping and potential bias. In this way, cultural sensitivity is woven into all aspects of the development of a CRT. Very few of the psychological and educational tests developed in South Africa to date are criterion-referenced. Test developers should be encouraged to change this situation as CRTs can serve useful purposes in educational and industrial settings and they have the potential to be cross-culturally appropriate based on the way in which they are designed.

Documenting test specifications. Whether a theory-based or a criterion-referenced approach is used to guide the content development of a test, test specifications should be prepared that document the content domains, behaviours, or constructs to be tapped by the test, the specific dimensions (or objectives) of each content domain, behaviour or construct that will be tapped, and an estimate of the number of items that the final test should ideally have for each content domain/ behaviour/construct and for each of the specific dimensions. This is often presented in the form of a two-way table in which the content domains/behaviours/constructs to be assessed are listed as row headings and the specific dimensions as column headings. The descriptions of specific items that fall in each of the cells of the table are also specified in the table together with the ideal number of items per cell. With a clear picture in mind of the test specifications, the format of the test, items, and responses need to be addressed next in the test plan.

Test, item and response modes and formats

In simple terms, a test consists of a stimulus (item) to which the test-taker responds using a specified response mode. There are various modes in which a test is presented (e.g., paper-based, computer-based); various item formats (e.g., multiple choice, performance tasks), and various response modes (e.g., verbal, written, typing on a computer keyboard). In multicultural assessment, Hambleton (1994, p. 232) expresses the view that "instrument developers/publishers should provide evidence that the choice of testing techniques, item formats, test conventions, and procedures are familiar to all intended populations". It should not merely be assumed that the different presentation and response modes or item formats are equally familiar to and appropriate for all cultural groups. Some examples in this regard will be provided below.

Presentation mode. As regards the mode in which a test is presented, there is some research evidence to suggest that the level of computer familiarity will influence computer-based test performance in South Africa, especially for test-takers who have low levels of technological sophistication (Foxcroft, Watson, Greyling & Streicher, 2001). Furthermore, research suggests that sociocultural factors influence the test-takers' ability to interact with the computer, which could impact on computer-based test performance (e.g., Evers & Day, 1997). Greyling (2000) notes that some of the interface design issues that are affected by culture are, for example, format conventions for numbers, dates, time and currency, as well as the use of icons, symbols and colours.

The implication of the above for the planning and design of tests is that, should the test developer decide to develop a computer-based test for multicultural use, consideration will have to be given to first researching the sociocultural factors that could impact on test performance in the various cultural groups so that these factors can be minimised in the design of the test. Consideration will also have to be given regarding how to deal with the impact of differing levels of computer familiarity and technological sophistication among various cultural and socio-economic groups in South Africa. One possibility is to develop a preparatory tutorial where test-takers can gain familiarity with the computer based format, can develop the necessary computer skills to respond to the test items, and can be introduced to differences in test-taking skills between computer-based and paper-based tests. An example of the latter is that the test-taking behaviour of moving quickly through items and then going back to review those missed, is not always possible in a computer-based test, especially with adaptive testing. Research has shown that by employing preparatory tutorials, the variations in test performance between test-takers with high and low levels of computer familiarity become insignificant (Powers & O'Neill, 1993).

Another way of catering for differing levels of computer familiarity among cultural and socio-economic groups is to develop a parallel paper-based version of the computer-based test. If test-takers have never used a computer before and are technologically fairly unsophisticated, the paper-based version of the test could be administered to them. However, as is the case when different language versions of a test are to be developed, it is essential that a test plan specify the methods that will be used to empirically demonstrate that the computer-based and paper-based versions of the test are equivalent (Bartram & Coyne, 2003).

Self-report tests. Cultural groups differ in terms of the relative importance attached to individual needs as opposed to group needs. One of the implications of this is that in cultures where individualism is valued, individuals learn how to introspect (reflect), to be aware of their personal needs, and to strive towards personal growth and development. In contrast, in a collectivist society, individuals learn to place their needs below that of their community and involvement in activities for the greater good of their society is valued more than pursuing personal happiness. Consequently, a self-report test, which requires the use of self-reflection and introspection when responding to the items, will pose different challenges for test-takers from individually and collectively orientated cultures, with the latter group being at a distinct disadvantage. Test developers need to ascertain whether or not a self-report test will introduce unwanted bias and thus whether it is wise to develop the test using this format.

Setting time limits. Sometimes test-takers are required to complete test items as quickly as possible within a time limit. However, cultures tend to differ in terms of their perception of time and the extent to which they are driven by it. Consequently, when a test is to be developed for multicultural use it might be wise to plan to rather develop a test where no time limits will be imposed. The Wechsler Intelligence Scale for Children–Fourth Edition (WISC IV) is in the process of being developed for a more multicultural target population than previous versions. Among the various innovations in the WISC IV design, the use of time limits for certain subtests has either been dropped or where time limits will still be imposed, scoring methods and norms will be provided so that the individual's performance can be determined and compared with and without time limits (Weiss, 2003).

Item format and content. There are a variety of different item formats and types (e.g., open-ended, forced-choice, sentence completion, essay format, performance type items where objects need to be manipulated). Furthermore, new item types have emerged in recent times that enhance the assessment of higher-order cognitive skills (e.g., dynamic problem-solving items where the online situation adapts and evolves as the test-taker enters response actions into the computer) (Hambleton & Zenisky, 2003). Test developers need to choose the most appropriate item type based on the purpose that the test intends to serve. However, when a multicultural test is developed, Hambleton (1994) points out that the item format as well as the item content and stimulus materials should be familiar to test-takers from all cultural groups in the intended target population. The implication of this is that the test developer cannot only base the choice of item format and content on the purpose of the test, but variations among cultural groups in terms of familiarity with item types or response modes also needs to be taken into account.

Owen (1989) extensively explored the characteristics of differentially functioning items for black and white test-takers in South Africa on an aptitude test battery. Among other things, he found that items in which a statement was made functioned worse than other verbal item types, that test-takers performed worse on story-type items, and that figure analogies functioned better than verbal analogies. Furthermore, stimulus materials such as graphs, diagrams, tables and pictures may not be equally familiar to test-takers from various cultural groups (Hambleton, 1994). Consequently, the test developer needs to review the research literature regarding item types and content that has been found to introduce bias into tests for various cultural groups. Thereafter a decision could be made to either omit problematic item types, or to include a balance of different item formats in the test, or to include practice items that would allow test-takers the opportunity to familiarise themselves with unfamiliar item types or content (Kanjee, 2001).

The development of item content can also be enhanced if the test development team immerses itself in the world of target populations and also consults with cultural experts, anthropologists and experts to develop a sense of what type of item content and test tasks the various subgroups in the intended target population are likely to be familiar with and could relate to.

Throughout this section it should be clear that test developers should be sensitive to the fact that their choice of presentation mode, item format, and response mode, represent potential sources of construct-irrelevant variance, which should be avoided or minimised so as to not give certain subgroups of test-takers an unfair advantage or disadvantage. The use of practice examples or even practice tests is often suggested as a solution for minimising the impact of unfamiliar presentation and response modes and item formats.

Specifying administration and scoring methods

Having chosen a test format, presentation and response modes, and item formats, the test developer needs to specify how the test will be administered and scored in the test plan. This information is an integral part of the test plan and can affect the way in which some of the test items are written. As regards administration, the test plan needs to specify whether the test will be administered in writing (paper-based), orally, or by computer; whether the test will be administered to groups or individuals; and how long it will take the test-taker to complete the test. As regards the scoring of the test, the test plan should specify who will score the test (e.g., the assessment practitioner or the computer); how the test will be scored (e.g., by awarding one mark for each correct response and then totalling up the marks per sub-scale); and what type of data the test is expected to yield (e.g., a raw score that must be converted to a norm score, or a descriptive category with respect to a performance standard related to the content domain sampled).

The implications of the administration mode for multicultural test development were dealt with in the previous section and there are not any substantial implications of the choice of scoring method for multicultural test development.

CONCLUSIONS

This article has attempted to outline important factors to consider when developing a test plan for a multicultural test. As such, it has only touched on the first step of the test development process. At appropriate points in the article, readers were referred to additional sources that can be consulted regarding the remaining steps in the test development process.

In closing, mention needs to be made of who should develop the test plan. Given the sensitivity to cultural factors that needs to be shown throughout the development of the test plan, it is unlikely that one test developer representing one cultural and language group could take on the task of developing the test plan or indeed the remainder of the process of developing the multicultural test. It is thus recommended that a multicultural test development team be assembled that demonstrates a rich mix of cultural and language groups and test development expertise. In addition, it is recommended that a reference panel of cultural experts, anthropologists, psychologists, linguists, and so forth be assembled to assist the test development team to develop the test plan.

REFERENCES

Abrahams, F. & Mauer, K.F. (1999b). Qualitative and statistical impact of home language on responses to the items of the Sixteen Personality Factor Questionnaire (16PF) in the South African context. *South African Journal of Psychology*, *29*, 76-86.

Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall International Inc.

Bartram, D. & Coyne, I. (2003). *International guidelines on computer-based and Internet delivered testing* (Draft version 0.3, March 2003). International Test Commission Web site: http://www.intestcom.org.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxomony of educational objectives. Handbook I: Cognitive domain*. New York: McKay.

Bracken, B.A. & Barona, A. (1991). State of the art procedures for translating, validating and using psycho-educational tests in cross-cultural assessment. *School Psychology International*, *12*, 119-132.

Brislin, R. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*, 185-216.

Claassen, N.C.W. (1997). Cultural differences, politics and test bias in South Africa. *European Review of Applied Psychology*, *47*, 297-307.

Das, J.P., Bisanz, G.L., & Mancini, G. (1984). Performance of good and poor readers on cognitive tasks: Changes with development and reading competence. *Journal of Learning Disabilities*, *17*, 549-555.

Das J.P. & Naglieri, J.A. (1994). *Das-Naglieri Cognitive Assessment System* (CAS). Chicago: The Riverside Publishing Company.

Els, C. (2004). *Occupational aspirations and gender stereotyping of Xhosa-speaking senior primary learners*. Unpublished masters' treatise, University of Port Elizabeth, Port Elizabeth, South Africa.

Evers, V. & Day, D. (1997, July). *The role of culture in interface*. Proceedings of INTERACT '97, pages 260-267, Sydney, Australia.

Foxcroft, C.D. (1997). Psychological testing in South Africa: Perspectives regarding ethical and fair practices. *European Journal of Psychological Assessment*, *13*, 229 -235.

Foxcroft, C.D. & Roodt, G. (2001). *An introduction to psychological assessment in South Africa*. Johannesburg: Oxford University Press.

Foxcroft, C.D., Watson. A.S.R., Greyling, J.H., & Streicher, M. (2001, July). *CBT challenges relating to technological unsophisticated test-takers in multilingual contexts*. Paper presented at the 7th European Congress of Psychology.

Freeman, M.C. (1984). *The effect of cultural variables on the Goodenough-Harris Drawing Test and the Standard Progressive Matrices*. Unpublished master's dissertation, University of the Witwatersrand, Johannesburg, South Africa.

Greyling, J.H. (2000). *The compilation and validation of a computerised selection battery for computer science and information systems students*. Unpublished doctoral thesis, University of Port Elizabeth.

Grieve, K.W. (2001). Factors affecting assessment results. In C.D. Foxcroft & G. Roodt (Eds.), *An introduction to psychological assessment in South Africa* (pp 315-343). Johannesburg: Oxford University Press.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests. *Bulletin of the International Test Commission*, *10*, 229-244.

Hambleton, R.K. & Zenisky, A. (2003). Advances in criterion-referenced testing methods and practices. In C.R. Reynolds & R.W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (2nd ed.) (pp. 377-404). New York: The Guilford Press.

International Test Commission, (ITC, 2001). International Guidelines for Test Use. *International Journal of Testing*, *1*, 93-114.

Kaplan, R.M. & Saccuzzo, D.P. (1997). *Psychological testing: Principles, applications, and issues* (4th ed.). Pacific Grove, CA: Brooks/Cole Publishing Company

Kanjee, A. (2001). Cross-cultural test adaptation and translation. In C.D. Foxcroft & G. Roodt (Eds.), *An introduction to psychological assessment in South Africa* (pp. 86-102). Johannesburg: Oxford University Press.

Krathwohl, D.R., Bloom, B.S. & Masia, B.B. (1956). *Taxomony of educational objectives*. Handbook II: Affective domain. New York: McKay.

Luiz, D.M. & Jansen, J.M. (2001). Assessment of young children, physically disabled and mentally handicapped individuals. In C.D. Foxcroft & G. Roodt (Eds.), *An introduction to psychological assessment in South Africa* (pp. 153-174). Johannesburg: Oxford University Press.

McIntire, S.A. & Miller, L.A. (2000). *Foundations of psychological testing*. Boston: McGraw-Hill Companies, Inc.

Meiring, D., Van de Vijver, F.J.R., Rothmann, S., & Barrick, M.R. (2003). Construct, item, and method bias of cognitive and personality tests in South Africa. Unpublished manuscript submitted for publication to *The International Journal for Selection and Assessment*.

Mochela, H. & Seymour, B.B. (2003, October). *Exploring learner's reported language preference in an assessment context*. Poster presentation at the 7th annual Prestigious Student Research Conference, University of Port Elizabeth.

Murphy, K.R. & Davidshofer, C.O. (1998). *Psychological testing: Principles and applications* (4th ed). Upper Saddle River, NJ: Prentice-Hall International.

Naglieri, J.A. & Das, J.P. (1990). Planning, attention, simultaneous and successive cognitive processes as a model for intelligence. *Journal of Psychoeducational Assessment*, *8*, 303-337.

Nell, V. (1994). Interpretation and misinterpretation of the South African Wechsler-Bellevue Adult Intelligence Scale: a history and prospectus. *South African Journal for Psychology*, *27*, 43-49.

Owen, K. (1989). *Test and item bias: The suitability of the Junior Aptitude Test as a common test battery of White, Indian and Black pupils in standard seven*. Pretoria: Human Sciences Research Council.

Powers, D.E. & O'Neill, K. (1993). Inexperienced and anxious computer users: Coping with a computer administered test of academic skills. *Educational Assessment, 1* (2), 153-173.

Retief, A. (1992). The cross-cultural utility of the SAPQ – bias or fruitful differences. *South African Journal of Psychology*, *22*, 202-207.

Robertson, G.J. (1990). A practical model for test development. In C.R. Reynolds & R.W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 62-85). New York: The Guilford Press.

Schepers, J.M. (1974). Critical issues which have to be resolved in the construction of tests for developing groups. *Humanitas RSA*, *2*, 395-406.

Taylor, T.R. & Boeyens, J. (1991). The comparability of the scores of blacks and whites on the South African Personality Questionnaire: An exploratory study. *South African Journal of Psychology*, *21*, 1-11.

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

Van Ede, D.M. (1996). How to adapt a measuring instrument for use with various cultural groups: a practical step-by-step introduction. *South African Journal of Higher Education*, *10*, 153-160.

Weiss, L. (2003, July). WISC-IV: *Clinical reality and the future of the four-factor structure*. Paper presented at the 8th European Congress of Psychology, Vienna, Austria.