TESTING IN BILINGUAL EDUCATION PROJECTS: LESSONS LEARNT FROM THE ADDITIVE BILINGUAL EDUCATION PROJECT

Elize Koch
University of the Western Cape

This paper discusses the topic of bilingual testing and the role played by the language of tests in home-language-based bilingual education projects such as the Additive BiLingual Education (ABLE) project. First, I highlight how national, macro-contextual contradictions in language in education affected the project. More specifically the impact of discrepancies between the additive bilingual language in education policy (LiEP) of the South African Department of Basic Education and practices around the language of tests such as the Annual National Assessments (ANA) are presented and analysed. The implications of these macro-contextual factors on bilingual education projects of a similar nature are discussed and suggestions for engagement are given. Secondly, the paper engages practically with the concept of bilingual educational testing and reveals how we dealt with it in the project. Research on the translation of a test used in the ABLE project is synthesised and the lessons learnt are discussed. It is argued that bilingual tests could contribute to a radical new way of approaching language in education in the context of South Africa.

PRE-AMBLE

As discussed in Koch, Landon, Jackson and Foli (2009), the primary aim of the Additive BiLingual Education (ABLE) project was to assist teachers at a rural Eastern Cape school with the translation of the 1997 South African Language in Education Policy into a workable model of implementation. The ABLE model advocated the maintenance of learners' mother tongue (MT), isiXhosa, as partial LoLT from grade R through grade 6. English, the first additional language, was introduced as a subject in the foundation phase, but learning the language and learning in the language occurred incrementally through the grades. In this manner, the previous abrupt switch to English as sole LoLT at grade 4 level was being circumvented. Learners in grade R in the experimental model spent 95% of curriculum time learning new concepts in their mother tongue. The remaining 5% represents their immersion in communicative English language activities. In grade 3, learners began to use English as a language of learning (a 20% time allocation) in preparation for the part-LoLT status of English in grade 4. Contrary to previous practices at the school, isiXhosa, under the bilingual model, served as primary-LoLT for an approximate 75% of the grade 4 curriculum. The model that was developed in consultation with the teachers increased English as LoLT to

71% of the time by the end of grade 6, in preparation for a complete switch to English as the only LOLT from grade 7 onwards.

This paper deals with the impact of the South African educational testing programme (the Annual National Assessment also called the ANAs) on the ABLE project and what was learnt during the process. The seemingly covert language-in-education policy promoted by national testing contradicts the South Africa's Department of Basic Education's Language in Education policy of 1997 and had a disabling and negative effect on the project. It is important to understand and analyse the contradictions and provide concrete suggestions about the way forward – more specifically about the role that well-conceptualised and researched bi- and multilingual tests and assessment can play. The lessons that were learnt about testing in bilingual (home language-based bilingual) education projects are therefore discussed at two levels. Firstly, that which was learnt more broadly about processes and forces around testing, multilingual testing specifically, and the extent to which these processes and forces affected the project are considered. Secondly, the specific lessons that were learnt about test translation and adaption flowing from the project team's need to have tests available in the two LoLTs of the schools, namely English and isiXhosa, are conveyed.

A social constructionist understanding of social reality and the discursive and constitutive interplay between power, knowledge, institutions and language (discourses) was useful as a broad framing of our understanding of what has happened and is happening. In addition, some theorising about the 'post-colonial' context was needed, as it is in this context that the above-mentioned interplay is taking place and the constitutive discourses exist. I also drew briefly on work discussing the role of testing in discourses around language in the contexts of the USA and Israel to explore how tensions and contradictions around tests play out in contexts such as South Africa. I will argue that the contradictions need to be resisted and managed, among other ways, via bilingual and multilingual testing.

BACKGROUND TO THE TESTING IN THE PROJECT

The two phases in the research (and testing) in the ABLE project

From the outset, the ABLE project was conceptualised with the focus on both research and implementation. Initially, however, the team focused more on research than implementation. The research took on a particular, more quantitative form. However, as a result of a number of contextual factors influencing the project, and intensive reflection about the project at the end of 2008, the project can be conceptualised as consisting of two phases, especially with regard to its approaches to research. These phases include the period from 2002–2009 and the period from 2008 to 2011, with an overlap of a year between the two periods.

During the first phase of the project, the need for tests was linked to the following quantitative, quasi-experimental research aim, namely to research the outcomes of the ABLE model of bilingual education on learners' language development, both in isiXhosa and English, their cognitive processing abilities and their academic achievements (see also Koch et al*.,* 2009). To do this, we compared the outcomes of learners at the research site with the language development, cognitive processing ability and academic achievement of two other 'control' groups of learners: those at a school implementing an early-exit bilingual

programme – the usual model in SA and many other post-colonial situations in Africa – and a straight-for-English English-medium school.

In the second phase of the project (from the beginning of 2009), the research focus shifted to the development of literacy practices at the school; an understanding and support of the teaching of reading at the school and teachers' own understanding of reading; and support with bilingual teaching and curriculum implementation at the school. We also focused our energies on the development of an explicit, written language policy (LP) for the school and managed the processes for this to happen. In this phase, we planned to use participatory action learning and research (PARL) to generate a better understanding of the 'how' of bilingual teaching and the model we were implementing in a context such as this and to develop materials that could be used to support other similar initiatives in the South African (SA) context. Because we deliberately decided to focus our energies and time (and funding) on 'implementation' issues (it is more correct though to refer to 'development' rather than 'implementation'), we decided to discontinue the testing of the children. We thus shifted our focus to co-construct with the teachers and other partners, what it means to teach bilingually, using the mother tongue as a basis language, in contexts such as the current one and in subjects such as Science and Mathematics, for example. The change in our philosophy of research was reflected in our decisions around testing.

Considerations in the testing during the first phase

For the first phase of the project, we identified the need for tests to measure language development, a test of cognitive processing ability and tests of academic achievement. In addition, we needed these tests in the two languages of the project, namely English and isiXhosa. In the way the project was conceptualised in 2002, we decided that the academic achievement tests needed to be in the language of learning and teaching (LOLT) of the grade and the subject area, in accordance with the bilingual model that had been adopted at the school, or the LP of the school, in the case of the comparator schools. There were obvious challenges around these needs. We needed valid, standardised tests with proven track records in bilingual education projects, or testing across language and cultural groups, as well as valid South African developed academic tests that could be used across schools during a time when the curriculum was not prescriptive about content.

Various decisions therefore had to be taken to meet the challenges pertaining to testing in the project. As there were no acceptable South African standardised tests available, not even in English, we obtained three tests with the help of the Head of the Department of the Education Department of Edinburgh University, UK. These tests were the Early Years Easy Screen test (EYES) (Clerehugh et al., 1998) and the normed Woodcock Muňoz Language Survey (WMLS) (Woodcock & Muñoz-Sandoval, 1993) for the language testing and the normed Kaufman Assessment Battery for Learners (KABC) for the testing of cognitive processing (Kaufman and Kaufman, 1983). In all these tests, we used only those subscales that were deemed applicable; for instance, in the EYES test we selected only the one pertaining to oral language development (module 4) and not the "hand-eye coordination" subscale. The EYES test was administered in isiXhosa.

Because we needed to have the WMLS available in isiXhosa as well, we decided to translate this test. We also needed the KABC in the language the children at all the sites controlled best at the age of five and at the end of grade 3, which was isiXhosa. We followed the widely

used (Foxcroft and Roodt, 2001) but flawed approach of translating the instructions into isiXhosa and administering the test as is, without any adaptations to the content of the test, as the content was deemed acceptable for use in our context. We also decided to use the raw scores of both the WMLS and KABC tests, as we did not have any norms for a South African population. The available norms on the tests for the USA population would not have been appropriate for use with these children (see Foxcroft and Roodt, 2005).

In terms of the academic achievement tests, the main challenge was to find content area tests that could be used across schools against the background of a scenario in 2002 where the curriculum was not prescriptive about content. For the first round of academic achievement testing, we used tests available in English and isiXhosa, developed by and used in the Imbewu project (Imbewu, 2004) to test literacy and numeracy achievement at the end of grade 3 across selected schools in the Eastern Cape in their project evaluation (see Koch et al., 2009). For grades 4–6, we approached the JET programme at the beginning of 2008 and purchased science, mathematics and language tests for grade 6. These tests were available only in English but the plan was to translate them into isiXhosa for the relevant grades and content areas. However, before that took place, in the second phase of the project, all testing in the project itself was dropped as the team decided to continue only with qualitative research. Research on the validity of the translated WMLS continued.

THE IMPACT OF MACRO-CONTEXTUAL PROCESSES AND FORCES ON DEVELOPMENTS IN THE PROJECT

Processes

At about the same time that we were busy with the process of changing into a more PARL approach and entering a new phase in our approach to research, the South African Department of Basic Education (DoBE) was also busy with a process of decision-making around testing, which dramatically influenced the ABLE project in various ways. Their own processes were informed by a series of low achievement of South African children on testing conducted by international agencies, such as the comparative TIMMS test in 2003 (where SA children ranked last) and the Progress in International Reading Literacy Study (PIRLS) in 2006, as well as the department's own systemic evaluations of literacy and numeracy in 2001 (grade 3), 2004 (grade 6) and 2007 (grade 3) (in which children showed poor performance) (DoE, 2011a).

At the end of 2008, the ABLE project team was informed that the national DoBE had decided to conduct their own yearly assessments at a national level via testing, called the Annual National Assessment (ANA), of the progress South African schools were making in terms of literacy and numeracy teaching and learning, in addition to continued participation in international testing. At a provincial level, each provincial DoE embarked on testing of its own, in addition to the national annual assessments. In the Eastern Cape, this testing takes the form of common assessments every six months in various learning areas, such as mathematics, natural science and the languages (also called 'common tests').

As a necessary caveat to contrast the language-in-education policy and language in testing practices, nationally and provincially, a brief overview follows of the Language in Education policy (LiEP) of the DoBE . Briefly, the LiEP of the national SA DoBE (1997) states that the aims of the policy are, among others, to support conceptual growth of learners and 'hence to

establish additive multilingualism as an approach to language in education' (1); improve participation in the economy by providing access to education; and develop programmes for the redress of the previously disadvantaged languages. The policy statement about the language of learning and teaching (LoLT) states simply that the LOLT of schools must be an official language. Given the aim of additive multilingualism, this policy statement, and the support in the South African constitution for the right to be educated in the language of one's choice (DoE, 1997), it is clear that any official language, including the African languages, can be used as LoLT in the classrooms – a policy principle the LP of the school in the ABLE project has adhered to.

Since the publication of the policy document very little has happened in real terms, either to unpack the concept of additive multilingualism for classrooms in South Africa or to implement additive multilingual education, especially with regard to the African languages, or to develop programmes to support the development of the African languages. As a result, and in contrast to the above clear statements about languages in education, including the African languages, the report on the ANAs that were conducted in all schools for the first time in 2011 (DoE, 2011b) stated that these tests were on 'languages and numeracy' in grades 1–3 and 'languages and mathematics' from grades 4–6. For grades 1–3, the language and numeracy tests were in all 11 official languages (DoE, 2011b: 12). However, from grades 4–6, both the language tests and the mathematics tests were only in English and Afrikaans (DoE, 2011b: 12). The approach to language in testing at the Eastern Cape provincial DoE level appears to follow the pattern of the national DoE.

While it is clear that the DoBE embarked on the ANAs and other provincial testing with the admirable motive of monitoring and improving education, one of the aims of the ANAs is also stated as promoting 'best assessment practices' and serving as an example to teachers in this regard (DoE, 2011b). One basic, internationally accepted principle of assessment practice in schools is that assessment and the LoLT must overlap (Figueroa and Hernandez, 2000). However, the practice of the DoBE in testing, at least nationally and in the Eastern Cape, continues to go against the grain of this 'best practice' with implications not only for schools like the school in the project but also in many other schools where the *de facto* LoLT is the MT (African languages), even when the LP of the schools states English as the LoLT (Probyn et al., 2002).

Despite a number of urgent contacts with both the provincial department and the national DoBE about the language of testing of the ANAs (the trials) in the school during 2009, as well as with the district officials on the same issue, the common tests during 2009, 2010 and 2011, the tests for grade 4 and onwards, continued to arrive at the school only in English. In 2009, the provincial common tests were translated by officials in the Eastern Cape department. However, in 2010 and 2011 and as a further demonstration of their commitment to the bilingual MT-based model of LOLT at the school, the teachers at the school translated the common assessments in the relevant learning areas into isiXhosa themselves, after several unsuccessful attempts to get the officials in the Department to commit to the translation of the tests.

Impact and forces

Regarding the practical impact on the school and the project, the most important effect was the poor results of learners on the English-only ANA of 2011. Internal testing was thus

dropped from the project, but it continued externally. The ANA results contradicted the results obtained on the common tests of 2010 (which were translated by the teachers) where the learners performed better than learners in schools that were similar to the ABLE project school (District Education Specialist, Curriculum, personal communication). These contradictions and the possible reasons for them were not explored further by either the Department or the Funder of the school and the English-only ANA results were accepted as the 'true' state of affairs. These results led to a number of 'interventions' and the decline of the project at the school (which is not explored further in this paper).

Of course, one might claim that the ultimate aim of MT-based bilingual education (MTbBE) is English language proficiency and if learners cannot demonstrate their learning in English and English literacy, MT education has failed. In the light of current developments around language in the Curriculum and Assessment Policy Statement (CAPS) documents and the role of English in exclusion from HE and the economy, this seems to be a realistic and valid claim. It is also clear that we will seriously disempower black and other African English second language (L2) learners via MT education if, in the current context, they do not end up with high levels of proficiency in English. This is an awareness also of the ABLE team. These statements bring us to the 'forces' and a theoretical framework to frame these understandings.

Social constructionism and post-colonialism in language use

Hook (2003: 111) views 'post-colonialism as a critical perspective that aims to understand the relationships of domination and/or resistance that manifest when one culture ... owns or controls another ... even after the era of formalised colonialism has ended.' Hook (2003) further draws on Fanon's description of the lived experience of black people during colonialism and apartheid, of racial alienation and estrangement in *Black Skin, White Masks* (1967), to explore, among others, the alienating identity effects of colonialism as cultural dispossession, more specifically also alienation through language via the adoption of the language of the coloniser. Bulhan's (1985) theory of 'cultural inbetweenity' proposed, as an expansion of Fanon's work, that colonised subjects (individually and as groups) psychologically deal with such alienation in three stages (and may or may not progress through all three). These stages are *capitulation* (when the dominant culture is completely adopted to the exclusion of the African culture), *revitalisation* (when there are attempts to go back and revitalise the dominated culture, as well as tribal cultures) and *radicalisation* (when the mutual influence of languages and cultures leads to new, synthesised forms of culture and hybridity in language use) (Bulhan, 1985; Hook, 2003).

With regard to language in the postcolonial context, Pennycook (2001) theorises about the politics of language in post-colonial contexts mainly in terms of exciting new possibilities of resistance and appropriation of, for example, English, and the creation of 'third spaces', also in terms of language use, 'language crossing' and non-essentialist 'hybridity' (Dua, 1994; Canagarajah, 1999). In terms of the concepts of third spaces and hybridity, the use of 'another's' language then is not about the dominance of one culture over another and acceptance of this dominance any more (*capitulation* or the 'colonialisation of the mind'), but it is then about appropriation and the creation of new possibilities – also in terms of language and culture. Adopting (and adapting) the language of the dominant culture then becomes subversion and resistance (thus *radicalisation* in Bulhan's terms). However, Pennycook (2001) also refers to Caragarajah (1999) who warns that conceptual linguistic hybridity might

lead to apathy and is potentially apolitical and supposedly neutral about language and language loss. He therefore urges one to deal contextually with language politics in postcolonial contexts when engaging in critical work.

As a possible alternative reading of the politics of language in the South African postcolonial context, Fanon in *The Wretched of the Earth* (1961) wrote almost prophetically about negotiated settlements between colonisers and the elites in Africa, with the continuation and even deepening of the oppression and exploitation associated with colonialism in the period following colonialism. While he did not specifically deal with language but more with power and politics in general, the adoption of the language of the coloniser as the language of the postcolonial society (*capitulation* in Bulhan's terms), also in education, may be regarded as part of the continued systemic and structural exclusion of the poor people of the country who often have low levels of proficiency in this language (see also Neville Alexander's 2002 discussion of language in the context of post-apartheid South Africa).

From a social constructionist perspective, institutions and the practices of institutions interact in complex manners with power, while discourses reflect and play a constitutive role in the construction and maintenance of power and in the acceptance by people of systems of power and knowledge (Burr, 2003). Particular ways of doing and acting then become part of internalised 'common sense' knowledge and talking. The uncritical acceptance of large-scale testing in English in education (e.g. for admission to higher education), where the majority of people are second or even third language (L2) speakers of these languages and the official LP in education supports additive multilingualism, may be highlighted as an example of a colonial institution that was constituted and is kept in place, by a plethora of discursive statements about empowerment, English as a global language, and so on. These statements have become part of the internal realities (in terms of identity and common sense) of many English L2 speakers in South Africa. While I acknowledge that we could interpret these statements as the constructions of 'third spaces', it is also possible to regard them as typical postcolonial discourse that discursively maintains and continues to construct the dominance of the elite. The contradictions between policy and practice, especially with regard to language in testing, appear to point us in the last direction.

Languages of tests to promote implicit language policies

Shohamy (2001, 2006), in both *The Power of Tests: a critical perspective of the uses of language tests* (2001) and *Language Policy: Hidden agendas and new approaches* (2006) has explored comprehensively how in many contexts, language tests (and language in tests) are used by people in power implicitly to promote agendas and ideologies that are often in direct contrast to explicit policies. Often these tests are used to promote a monolingual (one language) ideology or habit regarding society, in contrast to the multilingual pretence of policy (2006: 98). She has provided the examples of the USA, Latvia and Israel. We may want to add South Africa.

In the USA context, the use of English for the testing of bilingual children is highly contested (the debates are mainly about Spanish-English bilingual children); in fact, testing of bilingual children is contested, period (Figueroa and Hernandez, 2000). Extensive research points out the undesirable consequences of English-only tests and produces evidence of the bias of such tests, while it is clear that the interpretation of the scores of bilingual children, even on tests in their home language, is problematic. Several documents in this regard have served before

senate, such as Figueroa and Hernandez' 'Testing of Hispanic students in the USA, technical and policy issues' (2000). While they seemingly have had little effect in many of the USA's states, these documents continue to raise awareness of these issues and debates continue. Evidence of the negative consequences of English-only tests, as well as bias in these tests, has also been produced in South Africa (Koch, 2007; Koch and Dornbrack, 2008). However, the use of English-only tests in education in South Africa remains largely uncontested and this institution of (employing tests in a language that is not even the language of LOLT) continues to be supported by discourses, such as the ones mentioned above.

Bi- and multilingual tests as resistance in practice

Contradictions between practice and policy have dire consequences for projects such as the ABLE project and need to be resisted and/or managed. It appears that in South Africa we have not yet reached the stage of *radicalisation* where we would be working on thinking in new ways about language in education, including the languages of tests. Instead, we seem to be trapped in either *capitulation* or *revitalisation*, to use Bulhan's terms. Multilingualism in testing and assessment is one way forward.

If in principle one accepts that learners in bilingual education programmes need to be assessed and tested in both languages of instruction, one also needs to engage with the concept of bilingual testing and what it means in practice. I will deal with this very briefly based on our experiences in the project regarding the translation of a standardised language test from English into isiXhosa, namely the Woodcock Muñoz Language Survey (WMLS).

TEST TRANSLATION AND ADAPTATION

Test adaptation of the Woodcock Muñoz Language Survey

Adaptation was the process that was deemed appropriate in the case of translating the WMLS. Adaptation refers to the process of translating a set of items literally and to change the words or content of other items so as to enhance their appropriateness in the new cultural context (Van de Vijver and Poortinga, 2005), and then to evaluate the different versions of the test for equivalence. The term 'equivalence' in test construction theory deals with measurement issues and, more specifically, whether scores can be compared across groups. Equivalence is a prerequisite for valid comparisons and needs to be demonstrated by preliminary analyses in any cross-cultural (cross-linguistic) comparisons or uses (Van de Vijver and Poortinga, 2005). This is done by investigating bias. If bias exists, test scores cannot be used for comparison or across groups. 'Bias' is a technical psychometric term referring to the occurrence of score differences between groups when there are no differences in the actual underlying trait or ability (Van de Vijver and Poortinga, 2005). Equivalence is always challenged when bias occurs.

The processes of adapting the test into isiXhosa and the subsequent research done on the test were conducted in line with a series of 22 guidelines by the International Test Commission (ITC) for the development or adaptation of tests into more than one language (ITC, 2000).

The Woodcock Muñoz Language Survey (WMLS) is a typical prompt-response test with several items making up four individual subscales (table 1). Some of the subtests contained Multiple Choice (MC) items. As a test of word recognition and phonetic awareness, as well

as expressive vocabulary and verbal reasoning skills, this test has proved to be a useful predictor of reading in the USA. The WMLS was adapted into isiXhosa during two workshops: one two-day workshop at the end of 2004 and a one-day workshop at the beginning of 2006. During the first workshop, the main adaptation work took place. The first round of data was collected during 2005 to conduct the first exploratory statistical analyses of the equivalence of the two language versions of the test. The second workshop then took place to make changes to the test based on the first round of results. During this workshop, the team also changed some of the wording and correct options of the English version of the test.

Table 1: Summary of each subtest in terms of content, prompts and responses

Subtest	Linguistic and Curriculum areas	Stimuli	Test Requirement	Response
Picture Vocabulary (PV)	Oral expression; Language; development Expressive; vocabulary	Visual (pictures)	Identifying objects	Oral (word) Total=57
Verbal Analogies (VA)	Receptive expression; Vocabulary.	Auditory (phrases)	Stating a word to complete an Analogy.	Oral (word) Total=35
Letter-word Recognition (LWR)	Reading; Reading decoding.	Visual (text)	Identifying printed letters and words.	Oral (letter name, word) Total=57
Dictation (Dict)	Spelling, writing; Language development; English usage.	Auditory (words)	Spelling orally presented; Pre-writing and writing skills.	Motor (writing) Total=56

The adaptation and translation of the test were conducted by a multilingual and multi-disciplinary team consisting of bilingual English and isiXhosa-speaking linguists, language educators and accredited translators, mainly monolingual English-speaking language educators, and a bilingual (Afrikaans and English) psychometric expert with a background in research psychology. The multilingual and multidisciplinary nature of the teams enabled them to focus on identifying the underlying psychological (specifically cognitive) and linguistic processes as measured by the test and to take those into account in the adaptation process.

Two strategies that were used throughout were: (a) extensive relexification, such as the translation of roots and use of totally different phrases because of the lack of overlap in metaphors in the two languages or the lack of available words in the target language, isiXhosa; and (b) attention to the grading of difficulty in the items. All the subtests of the English version of the test consist of graded items in terms of difficulty level – easy to very difficult – as developmental norms for the USA population are set in terms of mental age – from 3 to 70 – and grade levels. This often necessitated the selection of other words or concepts instead of translating them from English into isiXhosa. The challenges and the

strategies for Verbal Analogies subtests are discussed briefly (see Koch, 2009, for a more comprehensive description of the process of adaptation).

 The main challenge of this subtest was the way in which the prompts were framed. The English version's prompts were in the form of, for example, "*mother* is to *father* as *sister* is to …" The analogies increase in difficulty with regard to the underlying logic and the pattern that needs to be identified for the analogy to be completed. However, the form remains the same throughout the subtest. While this phrasing is also an unnatural way of speaking in English, it still makes grammatical sense in this language. However, in isiXhosa, this way of phrasing a prompt would have made no grammatical sense. The response of the team in the second workshop was thus to choose a totally different form for the phrasing of the prompt, for example: "*Gqibezela oku ndikutshoyo – umama uhambelana notata, udade ahambelane no …*" [*Complete what I say – mother can be matched with father, and sister can be matched with...*]. Some of the items were completely changed, while keeping the underlying process/analogy type and the difficulty level in mind. The question was whether this made these items easier or more difficult in isiXhosa, specifically because of the grammatical clues with regard to the correct responses that were an inherent part of the form in which the prompt was phrased.

Lessons learnt

A number of studies on the item and construct bias of the subtests were conducted and can be accessed from the reference list. They consisted of studies on the equivalence of the English version of the test across English first language speaking and isiXhosa first language speaking learners (Haupt and Koch, 2012; Ismail and Koch, 2012). The equivalence of the isiXhosa version across two different isiXhosa dialect groups, namely urban and rural standard isiXhosa-speaking groups (Silo, 2010; Brown, 2012), and on the equivalence of the two language versions of the test, i.e. the English and isiXhosa versions (Arendse, 2010; Koch, 2009; Koch 2010; Roomaney and Koch, 2013). Interested readers can consult the articles dealing with the results in more depth. Table 2 summarises the main findings. All research procedures and data collection were done strictly in accordance with the ethical regulations of the then University of Port Elizabeth (now the NMMU). The researchers received permission from the main researcher to use and re-analyse the data collected for the main study.

Table 2: Bias results on the English and isiXhosa tests used across language and dialect groups and the two language versions of the test

Subscales	English test used across English and isiXhosa 1st language speakers.		isiXhosa test used across rural and urban isiXhosa 1st language speakers.		English and isiXhosa versions of the test.	
	N of biased items	Construct bias present	N of biased items	Construct bias present	N of biased items	Construct bias present
Picture Vocabulary	7	NA*	6	No	9	Yes
Verbal Analogies	4	Yes	3	NA	2	No
Letter-word Identification	21	NA	15	NA	10	Need further research
Dictation	4	NA	20	NA	18	Yes

*NA: not applicable because no research have been conducted yet

The main lessons from this research were that equivalence might well end up always being an issue whether the test is in one language and used across language groups, or available in two languages, and that we might never be able to prove complete comparability of scores across different language groups. As a result of the findings, we recommend that when we engage in multilingual testing (that is testing across multiple language groups), we need to follow a criterion-referenced (competence-based) approach in the interpretation of the scores. It is vital to keep in mind first, that tests need to be developed with criterion-referenced use in mind and, second, that one needs to ensure the correct interpretation of the scores for the different groups. This might mean developing different cut scores for different groups (see Cizek, 2001).

While debates around the uses of the African languages in education in South Africa often centre on the issue of dialect and whose dialect should be used for what purpose, our research on the vocabulary scale seemed to indicate that dialect differences do not have such great impact on the comparability of test scores across dialect groups in tests of this nature (Brown, 2012). Slight bias found on this scale might mean, though, that approaches to the assessment of vocabulary that are more holistic (e.g. embedded in reading passages) would be more valid than the atomistic approach followed in this test. This is a recommendation one can take forward to other constructs and tests as well.

The adaptation approach that we followed, especially with the verbal analogies scale, is one strong lesson that can be taken forward (Koch, 2009; Koch, 2010). Direct translation methods do not work on constructs such as the ones tested in this subscale in the case of languages that differ to the same extent as English and isiXhosa. This is in line with international approaches and findings, and is supported by the International Testing Commission's guidelines for adapted tests (2000). After first trying a more direct translation method that did not work (as revealed by the results of the first round of analysis), we adopted an approach of operationalising the underlying construct (verbal reasoning) and difficulty level of items in a manner that makes sense in isiXhosa and captures expression and analogy in this language. This last approach is consistently providing us with impressive results. The multilingual multidisciplinary nature of the team who enabled this approach is non-negotiable and possible in the South African context, and can be replicated in other projects and in testing in general in this country.

LESSONS AND RECOMMENDATIONS

In addition to the specific lessons that were learnt during the adaptation of the WMLS into isiXhosa, there were also a number of more general lessons that were learnt during the process as a whole.

First, national testing continued and affected the project negatively at the exact time that we decided to move away from testing into a more dynamic PARL approach in our research. We were not able to counteract the negative effects of an emphasis on only English use in tests. To an extent, we were caught by surprise by these developments. This demonstrated the need for continued engagement with the Department of Education. A major lesson was, thus, that at this stage, keeping international and South African developments in mind. It appears that we are going to have to accept the inevitability of large-scale educational testing. There is, of

course, nothing wrong in resisting large-scale testing. In fact, in the USA context, many teachers are mobilising to do exactly that and are putting forward convincing arguments why large-scale testing should be scrapped from the educational agenda (Torres, 2005). An alternative may be to acknowledge the usefulness of these tests in the evaluation of large-scale projects and interventions, but to engage with the validity of the tests and to resist an over-reliance on the results of these tests.

In multilingual settings, the language of tests is one crucial aspect that affects validity. Bias, and the effect of bias on the equivalence of test scores across language groups (and the interpretation of test scores), is one way in which "language" affects validity, as seen from the above discussion. Another is the discursive nature (or discourse "consequence") and other negative consequences of English-only tests in a multilingual context. The negative consequences of the English-only tests on this project was clearly demonstrated. This can be seen as "consequential validity" (Messick, 1989). In his seminal article, Messick (1989) argued convincingly that exploring the negative consequences of tests on groups (mainly minority and oppressed groups) has to be part of investigations into the validity of tests.

One way to resist these consequences is to engage with the concept of bi- or multilingual tests. The first level of engagement is the purpose of such tests. In an LP context of additive bilingual education, as is the case in South Africa, developmental maintenance bilingualism and biliteracy form the underlying philosophy and guiding framework that should not be compromised. This underlying philosophy and guiding framework also apply to language tests and languages in tests. In practice, this means that multilingual tests first create a space and are part of the space for learners to develop cognitively in languages they control best. Secondly, multilingual tests also create the space for learners to demonstrate their competence and cognition in languages they control best. However, these tests will only assist improved performance if children are educated in those languages, as they also need to develop conceptually and academically in their home languages (Figueroa and Hernandez, 2000)

The second level of engagement regarding the languages of tests would be issues of bias and equivalence, both to raise awareness and provide insight into these issues, and to provide practical support for research into, and handling of, bias in tests. We also need to engage with the content validity of tests. Not only does this mean that there needs to be coherence and alignment between the curriculum and the blueprints and content of tests (Messick, 1989), but it also means that these tests need to capture and test the underlying cognitive processes that go into the completion of actual tasks and display a range of difficulty levels that will allow for discrimination between strong (competent) and weak (not yet competent) learners.

Large-scale educational tests with a lack of validity at a content level and accordingly a lack of construct validity, which are used for large-scale evaluations and monitoring of progress in educational systems, have dire implications for classroom practices. It has been demonstrated that teachers tend to start teaching to the test under conditions of teacher and school evaluations (Madaus, Russell and Higgens, 2009). The last level of engagement with the validity of these tests is therefore the connection between tests and classroom instruction. Teachers must be able to work with the results of tests to help improve instruction. Beyond the need to engage with the content validity of tests and its implications for teaching, we also therefore need to assist teachers to develop a deep understanding of the results on (valid)

tests. Where the tests are not valid, teachers need to develop the critical skills to recognise and argue about this.

In engaging with projects dealing with the implementation of home-language-based bilingual education, it is therefore important to take macro-level developments into account in making decisions at a micro level. Where these developments contradict what happens at a project level, strategic engagement with other role players needs to be considered and motivated strongly also from international and theoretically defendable perspectives.

REFERENCES

ALEXANDER, N. 2002. *An ordinary country.* Pietermaritzburg: University of Natal Press.

ARENDSE, D. 2010. *Evaluating the Structural Equivalence of the English and isiXhosa Versions of the Woodcock-Muñoz Language Survey on Matched Sample groups.* Unpublished Master's dissertation, University of the Western Cape.

BROWN, Q. 2012. *Exploring the scalar equivalence of the picture vocabulary scale of the Woodcock-Muñoz language survey across rural and urban isiXhosa speaking learners.* Unpublished Masters dissertation, University of the Western Cape.

BULHAN, HA. 1985. *Frantz Fanon and the Psychology of Oppression*. New York: Plenum Press.

BURR, V. 2003. *Social Constructionism* (2nd ed). London: Routledge.

CIZEK, GJ. 2001. Conjectures of the rise and call of standard setting: an introduction to context and practice. In Cizek, GJ (Ed), *Setting performance standards*. New Jersey: Lawrence Erlbaum Ass Publishers.

CLEREHUGH, J, K HART, R PITHER, K RIDER, & K TURNER. 1998. *Early Years Easy Screen Short Form.* Slough: NFER.

DEPARTMENT OF EDUCATION (DOE). 1997. *Language in Education Policy*. Government Gazette 17997(383). Pretoria: Government Printer.

DEPARTMENT OF EDUCATION (DOE). 2011a. *Background to the Annual National Assessments, 2011.* Available from http://www.education.gov.za/Curriculum/AnnualNationalAssessment/tabid/424/Default.aspx [Accessed: 20 October 2011]

DEPARTMENT OF EDUCATION (DOE). 2011b. *Report on the Annual National Assessments, 2011.* Available from http://www.education.gov.za/Curriculum/AnnualNationalAssessment/tabid/424/Default.aspx [Accessed: 20 October 2011]

FIGUEROA, RA & S HERNANDEZ. 2000. *Testing Hispanic students in the United States: technical and policy issues.* President's advisory commission on educational excellence for Hispanic Americans. Available from http://www.clintonlibrary.gov/assets/storage/Research%20-%20Digital%20Library/brookssub/Box%204/647992-education-hispanic-2b.pdf [Accessed: 30/8/2012]

FOXCROFT, C. & G ROODT (Eds). 2005. *An introduction to psychological assessment in the South African context.* Cape Town: Oxford.

HAUPT, G. & E. KOCH. 2012. The argument for evaluating monolingual language tests for equivalence across language groups: empirical evidence of item bias. *South African Linguistics and Applied Linguistics Studies,* 30(1):65–76.

HOOK, D. 2003. Frantz Fanon and racial identity in postcolonial contexts. In K Ratele & N Duncan (Eds), *Social Psychology: Identities and relationships*. Lansdowne: UCT Press.

IMBEWU. 2004. *Imbewu II longitudinal evaluation: grade 3 learner assessment. Baseline study.* An unpublished report for the Department of Education, Eastern Cape.

ISMAIL, G. & E KOCH. 2012. Investigating item and construct bias in an English Verbal Analogies scale. *South African Linguistics and Applied Linguistics Studies*, 30(2):325–338.

ITC (INTERNATIONAL TEST COMMISSION). 2000. *Guidelines for adapting educational and psychological tests*. Available from http://www.intestcom.org/adapt_test.htm [Accessed: January 2008]

KAUFMAN, AS & NL KAUFMAN. 1983. *Kaufman Assessment Battery for Learners (KABC)*. Upple Saddle River NJ: Pearson Educational.

KOCH, E. 2007. The monolingual testing of competence: acceptable practice or unfair exclusion. In Cuvelier, P, T Du Plessis, M Meeuwis & Teck, L. (Eds). *Multilingualism and exclusion. Policy, practice and prospects*. Pretoria: Van Schaik Publishers. 79–103.

KOCH, E. & DORNBRACK, J. 2008. The use of language criteria for admission to higher education in South Africa: issues of bias and fairness investigated. Special edition on multilingual universities. SALALS 26 (3):333–350.

KOCH, E, J LANDON, MJ JACKSON & C FOLI. 2009. First brushstrokes: initial comparative results on the additive bilingual education project (ABLE). SALALS (27(1):109– 27.

KOCH, E. 2009. The case for bilingual language tests: a study of test adaptation and analysis. *Southern African Linguistics and Applied Language Studies.* 27(3):301–317.

KOCH, SE. 2010. *Adapting language tests from analytical languages such as English into an agglutinative tonal language of the Bantu family: Rasch modelling results.* Brief oral presentation: 27th International Congress of Applied Psychology, National Convention Centre Melbourne, Melbourne, 11–17 July 2010.

MADAUS, G, M RUSSELL & J HIGGINS. 2009. *The paradoxes of high stakes testing. How they affect students, their parents, teachers, principals, schools and society.* New York: Information Age Publishing.

MESSICK, S. 1989. Validity. In RL Linn (Ed.), *Educational Measurement* (3rd ed).London: Collier Macmillan publishers.

PENNYCOOK, A. 2001. *Critical Applied Linguistics.* London: Erlbaum.

PROBYN, M, S MURRAY, L BOTHA, P BOTYA, M BROOKS & V WESTPHAL. 2002. Minding the gaps – an investigation into language policy and practice in four Eastern Cape districts. *Perspectives in Education,* 20(1):29–46.

SHOHAMY, E. 2001. *The power of tests. A critical perspective on the uses of language tests*. Singapore: Pearson Education Asia Pty Ltd.

SHOHAMY, E. 2006. *Language policy: hidden agendas and new approaches.* London: Routledge.

SILO, U. 2010. *An evaluation of group differences and item bias, across urban isiXhosa learners and rural isiXhosa learners, of the translated version of the Woodcock-Muñoz Language Survey.* Unpublished Master's dissertation, University of the Western Cape.

TORRES, CA. (2005). *No child left behind: A brainchild of neo-liberalism and American politics.* Available from http://newpol.org/node/283 [Accessed: 22/8/2012]

VAN DE VIJVER, F. & POORTINGA, YH. 2005. Conceptual and methodological issues in adapting tests. In Hambleton, RK, PF Merenda & CD Spielburger (Eds), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment.* New Jersey: Lawrence Erlbaum.

WOODCOCK, RW & AF MUÑOZ-SANDOVAL. 1993. *Woodcock-Muñoz language survey.* Chicago: Riverside Publishing Company.

Biographical note

I am currently an extraordinary associate professor at the Psychology department, UWC. I specialise in measurement and assessment in education and psychology, and have a background and interest in bilingual developmental education and literacy development. My publications were mainly in the areas of bias and equivalence in measurement, and bilingual testing.