SA Journal of Industrial Psychology, 2007, 33 (1), 90-101
SA Tydskrif vir Bedryfsielkunde, 2007, 33 (1), 90-101

DIE EVALUERING VAN 'N EENTALIGE TOELATINGSTOETS WAT VIR TOELATING TOT HOËR ONDERWYS IN 'N VEELTALIGE KONTEKS GEBRUIK WORD

ELIZE KOCH
elize.koch@nmmu.ac.za
Nelson Mandela Metropolitaanse Universiteit[1]

ABSTRACT

The evaluation of monolingual admissions test used for admission to higher education in a plurilingual context. This study aims to critically evaluate the practice to use monolingual admissions tests across diverse language groups. The specific aim of the study was, accordingly, to evaluate the bias, across language groups, of a reading comprehension test used for admission to higher education. The subsequent aim was to decide about the scalar equivalence of the test across three language groups, namely Afrikaans and English students and students who are first language speakers of an African language. Item bias and structural differences between the English first and English second language groups were found, while structural differences continued to be found after deleting the DIF items from the test. Implications for fair admissions testing in the South African context are discussed.

OPSOMMING

Hierdie stukkie het dit ten doel om die praktyk om eentalige toelatingstoetse oor taalgroepe heen te gebruik, krities te evalueer. Die oogmerk van die studie was gevolglik om 'n toets van leesbegrip wat vir toelating tot universiteit gebruik word en wat slegs in Engels beskikbaar is, te evalueer vir sydigheid. Die uiteindelike oogmerk was om oor die skaalekwivalensie van die toets ten opsigte van drie taalgroepe te besluit, naamlik Afrikaanse en Engelse studente en studente met 'n Afrika taal as eerste taal. Item sydigheid en strukturele verskille tussen Engels eerstetaal sprekers en Engels tweedetaal sprekers is gevind, terwyl strukturele verskille na die verwydering van die DIF items voorgeduur het. Implikasies vir billike toelatingstoetsing in Suid-Afrika word bespreek.

Key words
Mondingual tests, cross-lingual testing, admissions testing bias and equivalence

SYNOPSIS

In plurilingual and multicultural countries tests can be translated or adapted into more than one language for use across diverse language groups. It is commonly accepted that tests which are translated or adapted have to be evaluated for equivalence, that is to say, the extent to which test scores have the same meaning across groups. On the other hand, tests are often available in only one language, but are used across more than one language or cultural group. They could be called 'monolingual tests'. In South Africa, as is the case elsewhere in the world, it is a common practice to use monolingual tests to make decisions about admission into tertiary education. These tests are not always evaluated for their applicability across groups to the same extent as with adapted tests. The overall aim of the study is to evaluate this practice by presenting an overview of the literature and by conducting an empirical study. A comprehensive review of the literature led to the acceptance of a theoretical framework of test equivalence, as it was formulated by Van de Vijver, Poortinga and others, to conceptualise and focus the empirical section of the study. This framework conceptualises test bias in terms of test equivalence. Some of the most salient issues in the literature will be discussed as the different approaches to these issues have serious implications for fair admissions testing practices in South Africa. The aim of the empirical section of this study was accordingly, as a case study, to evaluate the bias and subsequently, the equivalence, across language groups, of a test of reading comprehension, available only in English, but which is used across language groups to decide about admission to university. The results of the study indicated that a large proportion of items displayed unacceptable levels of differential functioning, or DIF, across three language groups, namely English and Afrikaans students and students speaking African languages, mainly Xhosa. The structural equivalence of the test was also a problem. DIF accounted for some of these differences. However, structural non-equivalence between the English and non-English speakers continued to be found even after the removal of the DIF items from the test. The lack of structural equivalence continued to have practical implications. The implications of these results for the practice of admissions testing in SA are discussed.

Hierdie studie het as sy oorkoepelende oogmerk 'n kritiese bespreking van die praktyk om eentalige toetse vir toelating tot tersiêre onderwys in Suid Afrika (SA) te gebruik. Internasionaal raak dit 'n algemene praktyk om opvoedkundige toetse, selfs toelatingstoetse, wat in heterogene taal- en kultuurgroepsverband gebruik word, te vertaal en dan seker te maak dat die vertaling geskik is (sien bv Beller, 1994 vir toelatingspraktyke in Israel; Robin, Sireci & Hambleton, 2003). 'n Belangrike eienskap van al die toelatingstoetse in SA is egter dat hulle geadministreer word in die taal van onderrig aan tersiêre instellings, meestal Engels en soms Afrikaans. Dit is die geval ten spyte daarvan dat die toetse in weinig gevalle 'n direkte toets van taalvaardigheid is en meer dikwels toetse van akademiese geletterdheid, wiskundige vaardigheid of potensiaal (sien byvoorbeeld Yeld, 2001). Die toetse word aan beide eerste- en tweedetaal sprekers van die taal van onderrig geadministreer sonder om enige onderskeid te tref in die hantering/interpretasie van die toetstellings van die twee groepe. Die argument is dat die student haar vaardigheid in die taal van onderrig moet demonstreer en dat hierdie praktyk dus sonder meer geregverdig is (Yeld, 2001).

Dié argument klink korrek, maar sal nie sonder meer standhou indien dit in terme van die standaarde wat vir sielkundige en opvoedkundige toetsing gestel word, geëvalueer word nie. Internasionale riglyne oor evaluering en toetsing in heterogene verband is besig om sterker klem te begin lê op die rigiede evaluering van toetse vir gebruik in heterogene kultuur of

1 Dank word betuig aan die Centre for Access Assessment and Research, HEADS by die NMMU vir die gebruik van die data vir die studie, en aan Cheryl Foxcroft, NMMU, en Stephan Sireci, University of Massachucetts at Amherst, VSA, vir hulle bydrae as promotors tot die oorspronklike DPhil studie.

taalverband sowel as vereistes ten opsigte van die interpretasie van die toetstellings van heterogene groepe. Terwyl daar steeds kritiek uitgespreek kan word oor die gebrek aan duidelikheid in hierdie riglyne, veral ten opsigte van eentalige toetse (Koch, 2005a), kan die verskuiwing na 'n klem op billike toetsing oor groepe heen as 'n prysenswaardige ontwikkeling beskou word. So kan daar byvoorbeeld na die hersiende APA Standards for *Educational and Psychological Testing* (AERA et al, 1999) en die stel van 22 riglyne vir die vertaling en/of aanpassing van toetse in verskillende tale verwys word (Hambleton, 2001; International Test Commission, 2000; Van de Vijver & Hambleton, 1996). Die leser wat geïnteresseerd is, word verwys na die webblad vir die volledige riglyne oor die aanpassing van toetse: www.intestcom.org/adapt/adapt_test.htm.

In SA bestaan daar nie spesifieke riglyne nie en bogenoemde internasionale riglyne word meestal as relevant vir die SA konteks aanvaar (Huysamen, 2002). Die problematiek van toetsing in heterogene verband in SA word wel in wetgewing aangespreek. So word daar byvoorbeeld in die Employment Equity Act No 55 (1998) vereis dat daar bewyse van die geldigheid en betroubaarheid van toetse vir gebruik in heterogene groepe moet bestaan voordat hulle aangewend mag word vir keuring, terwyl duidelike vereistes ook daargestel word vir die aanwending van maatreëls om regstellende aksie te implimenteer.

Hierdie wet het verreikende implikasies vir sielkundige en opvoedkundige toetsing in SA, omdat instansies nou verantwoordelik gehou word vir bewyse oor die toepaslikheid van toetse wanneer hulle die toetse in heterogene verband gebruik. Dit is egter belangrik om te onthou dat die klem in hierdie wet nie net val op die toepaslikheid van die toetse vir gebruik in heterogene verband nie, maar ook op die impak van die toetsing op die seleksie van kandidate uit voorheen-benadeelde groepe. Dit mag wees dat beduidend minder kandidate uit hierdie groepe afsnypunte op die toetse maak as gevolg van, byvoorbeeld, onderwys opleiding wat steeds onderstandaard is vir baie individue uit hierdie groepe. Dit het dan wel ander implikasies soos die onderverteenwoordiging van sekere groepe, ook genoem differensiële impak, wat by wyse van kwotas en verdere opleiding aangespreek moet word. Die spesifieke fokus van hierdie artikel is egter nie op hierdie aspek nie, maar op die toepaslikheid van toetse vir gebruik in heterogene groepe. Dit impliseer egter nie dat die problem van differensiële impak nie belangrik is en aandag moet geniet nie.

In die konteks van Hoër Onderwys bestaan daar nie spesifieke wetgewing oor toetsing vir keuring nie. Bogenoemde wet word egter wel op hierdie tipe van toetsing ook van toepassing gemaak. Die vereiste van bewyse oor geldigheid en betroubaarheid vir gebruik in heterogene groepe is in die konteks van Hoër Onderwys dus net so streng as in die geval van toetsing vir indiensneming, terwyl die kwessie van differensiële impak ook baie relevant is.

By 'n ondersoek na die toepaslikheid van toetse vir gebruik in heterogene groepe is die konsep van 'sydigheid' baie belangrik. 'Toets-sydigheid' word gedefineer as stoornis-faktore (dus ongevraagde en onbedoelde faktore) wat die variansie in toetstellings van verskillende groepe sistematies differensieel affekteer (Van de Vijver & Poortinga, 2005). 'n Oorsig oor debatte ten opsigte van toelatingstoetse in SA toon aan dat die probleem van sydigheid in toetse nie voldoende aandag geniet het nie (Koch, 2005a). Huysamen and Raubenheimer (1999) het wel gefokus op die voorspellingsydigheid van die matrikulasie eksamen oor etniese groepe heen. Hulle het verskille in die snydingslyne en hellings sowel as voorspellingsfout in 'n meergroep regressie analise ondersoek. Hulle het geen aanduiding van voorspellingsydigheid gevind nie. Navorsing op ander toetse in die SA verband het egter wel bewyse van veral konstruksydigheid gevind, byvoorbeeld Claasen (1993) ten opsigte van die New South African Group Test, Owen (1989) ten opsigte van die Junior Aptitude Test, en Abrahams (1996, 2002) en Meiring, van de Vijver, Rothmann en Barrick (2005) ten opsigte van toetse soos die South African Personality Questionaire en die 16 PF. Hierdie sydigheid word dikwels toegeskryf aan gebrekkige taalvaardigheid in die taal van die toets, naamlik Engels (Meiring et al., 2005). Sydigheid in toelatingstoetsing is dus iets waaraan daar in konteks van 'n veeltalige SA baie meer aandag geskenk behoort te word.

Fokus van die studie: 'n toets van leesbegrip in Engels

In hierdie studie val die soeklig op 'n toets van leesbegrip of akademiese geletterdheid in Engels wat gebruik word vir toelating en/of plasing in oorbruggingsprogramme by die Nelson Mandela Metropolitaanse Universiteit (NMMU) in Port Elizabeth[2]. Die toets wat in die VSA ontwikkel is, maar aangepas is vir die SA konteks, meet akademiese taalverwante leesvaardighede wat geklassifseer word as, onder andere, deduksie, inferensie en toepassing. Die toets word dus nie beskou as 'n toets van Engelse taalvaardigheid in sigself nie en word aan eerste- en tweedetaal sprekers van Engels geadministreer. Eerstetaalsprekers van Engels kan arguments-onthalwe ook sleg vaar op die toets omdat hulle nie die vaardighede soos in die toets gemeet word, bemeester het nie.

Akademiese geletterdheid is 'n vaardigheid wat in enige taal kan ontwikkel. Literatuur oor tweetalige akademiese geletterdheid toon aan dat sodanige vermoë na 'n tweede taal oorgedra word sodra 'n sekere drumpel van taalvaardigheid in die tweede taal bereik word (Cummins, 1984; Koda, 1994). Dit is waar dat sekere taalkundige verskille tussen twee tale, sowel as gebrekkige bemeestering van 'n tweede taal, leesvermoë in 'n tweede taal kan affekteer (Koda, 1994). 'n Toets van leesvaardigheid moet egter nie die kanse van tweedetaal lesers om te demonstreer dat die vereiste vaardigheid wel aangeleer is, benadeel nie. Daar is heelwat navorsing wat aantoon dat akademies vaardige studente wat 'n hoë vlak van akademiese leesvaardigheid in hulle eerste taal het, dikwels akademies beter vaar in 'n tweede taal as wat uit hulle oënskynlike taalvaardigheid in die tweede taal aanvaar kan word (onder andere, Adamson, 1993). Navorsing het ook aangetoon dat tweedetaal- en eerstetaal lesers in hulle prosessering van inligting verskil en dat hulle waarskynlik van verskillende strategieë gebruik sal maak om te verstaan wat hulle lees (Koda, 1994; Valdés & Figueroa, 1994).

'n Goeie leser van watter taalagtergrond ookal moet dus die geleentheid gegee word om sy/haar vaardighede van deduksie, inferensie en toepassing te demonstreer selfs al sou strategieë aangewend word wat verskil van dié van 'n leser wie se eerste taal die taal van die toets is. Met ander woorde, so 'n leestoets moet lesers met goed ontwikkelende akademiese leesvaardighede onderskei van lesers met swak ontwikkelde vaardighede, ongeag taal of agtergrond en die toetstellings van die verskillende groepe moet dieselfde betekenis hê. Dieselfde konstruk(te) moet dus by eerste- en tweedetaal lesers gemeet word en bewys hiervan moet verskaf word.

Die taal van onderrig by die NMMU is Engels. Die universiteit kan as veeltalig en multikultureel beskou word deurdat ongeveer 36% van die studentepopulasie (uitsluitend die internasionale studente) Engelssprekend is, ingeveer 16% Afrikaanssprekend en die res een of ander Suid Afrikaanse Afrika taal as eerste taal het. Die oorgrote meerderheid van die laaste groep, ongeveer 70%, is Xhosasprekend (Focus, 2004). Daar is ook 'n klein groep internasionale studente met 'n verskeidenheid van tale as eerste taal. Terwyl uitsluiting van Hoër Onderwys in terme van taal in die SA konteks in sigself geproblematiseer kan word, is dit nie die spesifieke fokus van hierdie artikel nie en moet daar voorlopig aanvaar word dat Engelse taalvaardigheid wel 'n rol sal speel in die akademiese prestasie van Engelse tweedetaal sprekers aan dié universiteit. Sydigheid maak dus moontlik deel uit van assessering op universiteit en die argument mag wees dat sydigheid in die toets dus net hierdie 'realiteit' reflekteer.

2 Die oogmerk van die artikel is nie om praktyke aan een bepaalde universiteit te kritiseer nie, maar dit as 'n gevallestudie te gebruik om 'n baie algemene praktyk aan SA tersière inrigtings te problematiseer.

Navorsing het egter aangetoon dat tweedetaal sprekers se vermoë om inhoudsvakke in 'n tweede taal te begryp waarskynlik beter is as hulle vermoë om verbale toetse in die taal te prosesseer (Pennock- Roman, 1998). Verder is die konteks in die toets onder bespreking, soos algemeen die geval is in toelatingstoetse, beperk en die inhoud onbekend, iets wat nie so ooglopend is in assessering op universiteitsvlak nie. Dit moet ook aanvaar word dat terwyl daar vir die doeleindes van hierdie artikel voorlopig aanvaar word dat Engelse taalvaardigheid belangrik is vir funksionering in 'n Engelse akademiese opset en dit 'n belangrike rol in akademiese leesvaardigheid speel, die toetstellings van 'n sydige toets nogtans nie vir verskillende groepe ekwivalent is nie. 'n Sydige toets kan dus nie gebruik word om die tellings van eerste en tweedetaal sprekers te vergelyk of op dieselfde skaal te plaas nie.

Om die aard van die problematiek rondom sydigheid te verhelder, word daar eerstens 'n kort teoretiese oorsig van die belangrikste vraagstukke in toetsing binne heterogene kultuur en taalgroepverband soos dit in die internasionale literatuur bespreek word, gegee, waarna 'n empiriese gevallestudie van 'n eentalige leestoets wat vir toelating oor verskillende taalgroepe gebruik word, bespreek sal word. Om mee te begin, is dit nodig om 'n onderskeid te tref tussen billikheid en sydigheid. Daarna sal 'n 'n teoretiese raamwerk vir die evaluering van sydigheid en ekwivalensie in toetsing kortliks bespreek word.

TEORETIESE OORSIG

Die onderskeid tussen sydigheid en billikheid

In die literatuur word billikheid en sydigheid as belangrike vraagstukke in kruis-kulturele en -linguistiese toetsing beskou. Die sentrale probleem wat debatte in lande soos die Verenigde state van Amerika (VSA) onderlê, is die verskynsel dat minderheidsgroepe wat meestal tweedetaal sprekers van Engels of 'African-Americans' insluit, tot 1.8 van 'n standaard afwyking laer presteer op toetse soos die Standardised Achievement Test (SAT) en die American College Testing (ACT) as eertse taal sprekers van Engels. Dit is toetse wat algemeen gebruik word vir toelating tot prestige universiteite in die VSA. Hierdie feit veroorsaak dat 'n beduidend kleiner ratio van studente uit minderheidsgroepe tot hierdie universiteite toegelaat word. Dit kan beskou word as 'n bewys van 'differensiële impak' en 'n aanduiding van onbillikheid, dit is, tensy maatreëls soos regstellende aksie of kwotastelsels geld, en het op sy beurt gelei tot die aanwending van verskillende afsnypunte of norme vir minderheidsgroepe (Sireci & Geisinger, 1998). Die groot verskil in die gemiddelde prestasie van die verskillende groepe word ook dikwels, verkeerdelik, voorgehou as 'n bewys van die inherente sydigheid van gestandardiseerde toetse.

In die internasionale en SA literatuur is daar algemene verwarring oor die onderskeid tussen die twee terme, billikheid en sydigheid, en die implikasies wat die onderskeid inhou vir toetsing binne heterogene taal- en kultuurgroepsverband (Camilli, 1993; Cole & Moss, 1989; Yeld, 2001). Die twee terme word dikwels as sinomieme gebruik of word deur verskillende teoretici op teenstrydige maniere gedefinieer. Cole and Moss (1989) in hulle seminale werk oor sydigheid gebruik die woord 'metingssydigheid' ('bias' in Engels) in die sin van beide billikheid én sydigheid. Hulle stel, byvoorbeeld, dat toetse wat na 'n uitgebreide evaluering van sydigheid duidelik nie sydig is nie, nog steeds as sydig in die sin van onbillik beskou kan word indien minderheidsgroepe as 'n groep swakker vaar op so 'n toets as die meerderheidsgroep. Camilli (1993) tref 'n duidelike onderskeid en stel voor dat sydigheid 'n tegniese probleem is wat psigometries deur statistiese metodes geëvalueer moet word, terwyl billikheid 'n eties-morele probleem is wat deur middel van filosofiese debatering op wetlike vlak aangespreek moet word. Nadat daar besluit is op 'n bepaalde interpretasie van billikheid en nadat bewys is dat 'n toets nie sydig is nie, is die vraag dan of die uitkomste van 'n bepaalde keuringsprosedure billik is.

Beide hierdie groepe teoretici stel voor dat sydigheid verstaan en geïnterpreteer moet word binne die raamwerk van konstrukgeldigheid, soos gedefinieer deur Messick in sy seminale werk van 1989. Dit is, daar moet omvattende bewyse bestaan dat dieselfde konstruk of domein in die verskillende groepe gemeet word. Dit word gedoen, onder andere, deur middel van bewyse dat items nie moeiliker is vir sekere groepe as vir ander groepe van dieselfde vermoë nie, dat items nie sydigheid openbaar nie, dat dieselfde konstruk in verskillende groepe gemeet word en om te demonstreer dat toetse dieselfde voorspellingsgeldigheid het vir verskillende groepe, ook genoem kriteriumverwante geldigheid. Laasgenoemde twee vereistes kan byvoorbeeld gedemonstreer word deur statistiese tegnieke soos faktor analise en/of regressie analise, terwyl daar 'n verskeidenheid statistiese tegnieke en ontwerpmodelle bestaan om eersgenoemdes te evalueer.

Hierdie teoretici, en veral Cole en Moss (1989), stel egter voor dat die interpretasie van sydigheid afhang van die konteks van toetsing, 'n benadering wat tot heelwat teenstrydige interpretasies lei. So beskou Cole en Moss, byvoorbeeld, die differensiële effek van taalvaardigheid op die toetstellings van tweedetaal sprekers in rekenkundige toetse as 'n bewys van sydigheid, terwyl dieselfde effek in 'n toets van akademiese geletterdheid nie as sydigheid beskou word nie. Ander teoretici in dieselfde skool beskou die effek van taalvaardigheid in rekenkundige toetse egter as deel van die konteks van die toets en dus nie as sydigheid nie. Dit is, selfs al sou toetslinge beter vaar op soortgelyke toetse in hulle eie taal (Ebel & Frisbie, 1986).

Geisinger (1996), in teenstelling met die bogenoemde definisies en in navolging van Cleary (1968), definieer billikheid uitsluitlik in tegniese terme. Volgens hom beteken dit dat toetse wat nie die prestasie van verskillende groepe differensieël voorspel nie, as regverdig beskou kan word. Billikheid word op die volgende manier geëvalueer: '*Regression lines between the predictor test and the criterion are computed for each relevant group and their slopes, intercepts and the errors of prediction are compared. If any of these three components ...differs across groups, then the test is not thought to be fair*' (Geisinger, 1996, p.29). Hy beperk dus die evaluering van billikheid tot die tegniese evaluering van slegs een aspek van geldigheid, naamlik kriteriumverwante geldigheid. Differensiële impak maak, volgens Geisinger, nie deel uit van die problematiek van billikheid nie, selfs al sou dit as belangrik beskou word om dit te ondersoek.

Vir die doeleindes van hierdie artikel sal die bogenoemde debatte nie in meer besonderhede bespreek word nie. Dit is egter belangrik om kortliks kennis te dra van die volgende tendens in die VSA, omdat dit direkte implikasies vir SA inhou en alreeds ge-eggo word in debatte in SA asof dit sondermeer onproblematies is. Behalwe dat die benadering wat soms daar ten opsigte van die evaluering van sydigheid voorgehou word, problematies is, kan die kwessie van die onderverteenwoordiging van bepaalde groepe in die SA verband vanuit regs- en eties-morele oogpunte nie op dieselfde manier as in die VSA hanteer word nie.

In die VSA word daar tans hewiglik gedebateer oor billikheid en sydigheid in toetsing (Sireci & Geisinger, 1998). Die argument dat billikheid in terme van differensiële impak 'n belangrike aspek is om aan te spreek en die metodes wat voorgestel word om dit te hanteer, soos byvoorbeeld kwotastelsels of verskillende afsnypunte vir verskillende groepe, word tans vanuit neoliberale konserwatiewe kringe sterk teëgestaan as omgekeerde diskriminasie.

In geregtelike terme in die VSA is daar nou bewegings om billikheid te benader, nie meer in terme van die maatreels soos regstellende aksie of kwotastelsels nie, maar in terme van bewyse dat toetse nie voorspellingssydigheid openbaar nie (Sireci & Geisinger, 1998). Dit wil sê, in terme van die voorgestelde definisie van Geisinger (1996) en Cleary (1968), naamlik dat as 'n toets nie die prestasie van groepe differensieël voorspel nie, die toets as billik beskou kan word.

Verder word daar ook vereis dat items nie moeiliker mag wees vir lede van sekere groepe as vir lede van dieselfde vermoë van ander groepe nie (items moenie sydig wees nie). Daarna word verwys as die evaluering van DIF of "differential item functioning" (Sireci & Geisinger, 1998). Dié vereiste hang egter, volgens die voorstanders van hierdie benadering, van die konteks van die toetsing af. Met ander woorde, selfs al is sekere items moeiliker vir lede van sekere groepe as vir lede van dieselfde vermoë van ander groepe, kan op grond van sekere kriteria of die oordeel van 'n groep 'eksperts' dat die effek deel vorm van die konstruk van die toets, besluit word om dié items nietemin te behou (Sireci & Geisinger, 1998). Dit word die 'multidimensionale benadering tot DIF' genoem en lei ook tot heelwat teenstrydighede in die interpretasie van sydigheid, soos vroeër bespreek. Nietemin word die laer toetstellings van sekere groepe, na die evaluering van billikheid, dan as onproblematies of regverdig beskou, wat die toepassing van ander kriteria soos kwota stelsels or benaderings soos verskillende afsnypunte vir verskillende groepe in terme van die beginsel van differensiële impak, uitskakel.

Ander navorsers in die VSA verband is egter besig om bogenoemde benaderings te betwis. Pennock-Roman (1998,1999) het bevoorbeeld deur middel van empiriese bewyse aangetoon dat die SAT in die geval van akademies sterk studente wat meer vaardig in Spaans as Engels is, sterk statistiese verwantskappe met Engelse taalvaardigheid openbaar, byvoorbeeld, tot 36% van die variansie word in die geval van die verbale toets deur Engelse taalvaardigheid verklaar. Daar word dus iets anders as die konstruk van belang getoets. Pennock-Roman stel dit duidelik dat in die geval van studente wie se sterkste taal nie Engels is nie, hierdie toetse wel voorspellingsgeldig kan wees, maar dat sodanige resultate moontlik nie veel meer as net studente se taalvaardigheid in Engels aandui nie. Haar aanbeveling vir toelatingskomitees is dat die SAT tellings van eerste en tweedetaalsprekers van Engels as gevolg van konstruksydigheid nie op dieselfde skaal geplaas behoort te word vir die doeleindes van interpretasie nie, omdat daar statisties nie 'n basis vir 'n vergelyking bestaan nie.

Simplistiese benaderings tot die probleem van die sydigheid van items word ook betwis. Teoretici soos van de Vijver en Tanzer (1998) stel dit dat probleme met die konstrukgeldigheid van toetse DIF kan verskans, omdat die evaluering van DIF tradisioneel afhang van die aanname dat die totaaltellings op die toets nie ook sydigheid openbaar nie. Van de Vijver en Leung (1997) bevraagteken verder die praktyk om onder sommige omstandighede items wat DIF openbaar in 'n toets te behou. Volgens hulle is items wat differensiële patrone van moeilikheidsgraad (DIF) openbaar, 'n metingsvraag en sal hulle teenwoordigheid in 'n toets altyd die vergelyking van tellings oor groepe heen kompromiteer. Die konsepte sal in meer besonderhede in die volgende afdeling bespreek word, terwyl die leser wat geïntereseerd is, verwys word na die uitgebreide literatuur oor die onderwerp.

Die siening dat differensiële voorspelling en DIF as voldoende evaluering van sydigheid (hierdie navorsers noem dit billikheid) aanvaar behoort te word, kan dus empiries, maar ook teoreties, geproblematiseer word. In hierdie teoretiese benadering word sydigheid met metingsekwivalensie verbind en nie net met geldigheid, veral bloot kriteriumverwante geldigheid, nie. Die teoretiese raamwerk in terme waarvan dit gedoen word, word in die volgende afdeling bespreek.

'n Teoretiese raamwerk: sydigheid en ekwivalensie

Poortinga (1989) en Van de Vijver en sy medewerkers beskou die probleem van ekwivalensie en sydigheid as dié sleutelvraag in kruiskulturele en- linguistiese navorsing en toetsing. Volgens die navorsers verskaf die twee konsepte oorvleuelende, maar effens verskillende perspektiewe op die sentrale vraag of die toetstellings van verskillende groepe dieselfde betekenis het (Van de Vijver en Tanzer, 1998).

Ekwivalensie is 'n tegniese term wat verwys na die skaal van meting in terme waarvan vergelykings tussen indiwidue of groepe gemaak kan word. Om direkte vergelykings tussen die tellings van verskillende groepe te maak, hetsy vir navorsingsdoeleindes of keuring, moet daar afdoende bewyse bestaan dat die tellings van verskillende groepe op dieselfde skaal is en dat dieselfde konstruk in die verskillende groepe gemeet word.

Daar is drie vlakke in die hiërargie van ekwivalensie, naamlik strukturele ekwivalensie, ekwivalensie van metingseenheid, en skaalekwivalensie (Van de Vijver & Tanzer, 1998). Strukturele ekwivalensie tussen groepe bestaan wanneer dit bewys kan word dat dieselfde konstruk gemeet word in verskillende groepe, met ander woorde, as konstruksydigheid nie bestaan nie. Ekwivalensie van metingseenheid verwys daarna dat die metingseenheid dieselfde is, soos die geval is met temperatuur in Celsius en temperatuur in Kelvin. Skaalekwivalensie is die hoogste vlak van ekwivalensie en is 'n voorvereiste wanneer groepe vergelyk word in navorsingsverband of wanneer die tellings van toetslinge in verskillende groepe op een skaal geplaas word vir besluitneming soos in die geval van keuring. Skaalekwivalensie kan nie direk bewys word nie, en kan slegs indirek ondersoek word deur die evaluering van sydigheid. Skaalekwivalensie bestaan wanneer dieselfde konstrukte gemeet word, en wanneer die oorsprong van die skaal en die metingseenheid in die verskillende groepe dieselfde is.

Die evaluering van sydigheid geskied op 'n omvattende en geïntegreerde manier en dit word as belangrik beskou om al die vorms van sydigheid te evalueer alvorens 'n toets in heterogene verband aangewend kan word. Sydigheid word geklassifiseer in konstruk- en itemsydigheid en sydigheid wat verband hou met metodologiese faktore, soos differensiële bekendheid met itemformaat. Konstruksydigheid toon aan dat verskillende konstrukte in die verskillende groepe gemeet word, terwyl itemsydigheid verwys na die nou reeds bekende DIF. DIF kan geklassifiseer word in uniforme en nie-uniforme DIF. Uniforme DIF beteken dat 'n item 'n bepaalde groep, ongeag vermoë, bevoordeel, terwyl in die geval van nie-uniforme DIF die helling van die regressie van die item op die latente konstruk oor groepe heen verskil (Jodoin & Gierl, 2001). In al die gevalle bestaan daar 'n verskeidenheid van statistiese en ontwerpmetodes wat gevolg kan word om sydigheid te evalueer (Van de Vijver & Tanzer, 1998).

Navorsers soos Van de Vijver en Tanzer, (1998), Van de Vijver en Lueng (1997), Helms-Lorenz en Van de Vijver (1995) en Poortinga (1989) gaan van die veronderstelling uit dat toetse wat sydigheid openbaar se toetstellings noodwendig nie dieselfde betekenis in verskillende groepe het nie, of op dieselfde skaal is nie (dus nie ekwivalent nie) en dat dié groepe se toetstellings dus nie vergelyk kan word of op dieselfde skaal geplaas kan word vir interpretasie nie. Toetse nie inherent sydig of nie sydig nie; dit moet in elke nuwe konteks van gebruik, gedemonstreer word (Helms-Lorenz & Van de Vijver, 1995).

Konstruksydigheid affekteer strukturele ekwivalensie. DIF en sydigheid wat verband hou met metodologiese faktore affekteer soms, soos in die geval van nie-uniforme item sydigheid, maar nie altyd nie, strukturele ekwivalensie. Hulle affekteer wel altyd

skaalekwivalensie. DIF of sydigheid wat verband hou met metodologiese faktore verander die oorsprong van die skaal en mag die metingseenheid ook affekteer (van de Vijver & Leung, 1997). Bewyse van enige vorm van sydigheid word dus as 'n aanduiding van 'n gebrek aan een of ander vlak van ekwivalensie beskou.

Alhoewel die vereiste van ekwivalensie op hierdie stadium in die internasionale riglyne hoofsaaklik met vertaalde toetse verbind word, is daar tog wel sterk teoretiese en empiriese aanduidings dat toetse wat slegs in in een taal beskikbaar is (ook genoem "eentalige toetse") maar in heterogene taalgroepsverband gebruik word, aan dieselfde vereistes moet voldoen. So kan daar verwys word na die werk deur Pennock-Roman (1998, 1999), soos hierbo bespreek. Daar is ook 'n toename in navorsing van hierdie aard op eentalige toetse in Suid-Afrikaanse verband, byvoorbeeld, 'n studie op die 15 FQ+ ('n alternatief op die 16 PF) wat vir keuring tot die SA Polisiediens gebruik word (Meiring, et. al., 2005. In hierdie studie is daar gevind dat strukturele sydigheid tussen swart, bruin en wit etniese groepe die ekwivalensie van die toets (in Engels) tot so 'n mate kompromiteer dat aanbeveel is dat die voortgesette gebruik van die toets heroorweeg moet word. Ander studies waar die effek van Engelse taalvaardigheid op tellings as problematies uitgewys is, is Claasen (1993), Owen (1989), Abrahams (1996, 2002), Abrahams & Mauer (1999a, 1999b) en Meiring et al., (2005).

In die volgende afdeling word 'n gevallestudie van die empiriese evaluering van die sydigheid van 'n eentalige leestoets wat in SA verband gebruik word ter illustrasie van bogenoemde teoretiese raamwerk aangebied. Die toepaslikheid van die toets is ten opsigte van drie taalgroepe ondersoek, naamlik 'n Afrikaanssprekende groep, 'n Engelssprekende groep en 'n groep wat 'n verskeidenheid van inheemse Suid Afrikaanse Afrika tale praat. Die Afrikaanssprekende groep en die Afrika taalgroep is nie in een nie-engelssprekende groep geplaas nie, omdat die groepe aansienlik verskil in terme van die opsies oor medium van onderrig op skoolvlak – die meerderheid van Afrikaanssprekendes word in Afrikaans onderrig, terwyl die groep wat Afrika tale praat, hoofsaaklik in Engels onderrig word. Taalverskille tussen Engels en Afrikaans is ook minder opvallend as taalverskille tussen Engels en die Afrika tale, byvoorbeeld ten opsigte van sin-struktuur en kognate (Kotze, persoonlike kommunikasie, September 2006). Die Engelse taalgroep was die kontrole- of verwysingsgroep en die ander twee groepe die fokus groepe. Met ander woorde, die prestasie van die Engelse groep is as verwysing gebruik waarteen die prestasie van die twee ander groepe geëvalueer is.

NAVORSINGSONTWERP

Loodsstudie

'n Loodsstudie op 'n beskikbaarheidssteekproef van 989 eerstejaarstudente (2003 en 2004 toelating gekombineerd) het aangetoon dat daar groot verskille tussen die taalgroepe bestaan ten opsigte van hulle prestasie op die toets (Koch, 2005a). Die Afrikaanse groep (n = 190, gemiddeld = 22,7, standaard afwyking = 6,02) het 'n gemiddelde van 0,5 standaard afwyking laer as die Engelse groep (n = 260; gemiddeld = 25,6, standaard afwyking = 5,06) op die toets behaal, en die Afrika taalgroep, 'n gemiddelde van (n = 539, gemiddeld = 17,09, standaard afwyking = 6,05) van 1,5 standaard afwyking laer as die Engelse groep. Die standard 1,5 afwyking van die Engelse groep is as verwysing gebruik. Die totaaltelling van die toets is 35 (die telling uit 35 word na 'n telling uit 120 getransformeer vir interpretasie). Die Cronbach Alpha's vir die drie groepe het gewissel van 0,80 vir die Engelse groep tot 0,85 vir die Afrikaanse groep, met 0,81 vir die Afrika taalgroep.

Ongeveer 44% van Afrika taal sprekers en sowat 16% van Afrikaanssprekendes behaal 'n punt wat laer as die afsnypunt van 16 (uit 35) wat op die toets gestel is (Koch & Foxcroft, 2003) in teenstelling met slegs 3% van Engelssprekendes (Koch, 2005b), 'n verskynsel met ernstige implikasies vir billike toelatingspraktyke. In ander navorsing by die NMMU is gevind dat die Afrikaanse groep nie akademies swakker vaar as die Engelse groep op universiteit of in matriek nie, terwyl korrelasies van die toets met akademiese prestasie vir die Afrika taalgroep swakker is as vir die ander twee groepe (Koch, 2003; Koch, 2005b).Verder is ook gevind dat die item-totaal korrelasies van die Afrika taalgroep en die Afrikaanse groep swak tot matig met dié van die Engelse groep korrreleer (r = 0,14 en 0,38 onderskeidelik) in teenstelling met 'n sterk korrelasie van r = 0,77 met mekaar (Koch, 2005a). Laasgenoemde is 'n aanduiding van verskille in die relatiewe orde van item-totaal korrelasies in die verskillende groepe en die moontlikheid dat verskille in die konstruk oor groepe heen bestaan (Van de Vijver & Lueng, 1997).

Hierdie resultate saam met die teoretiese vereistes ten opsigte van toetse wat in heterogene taal- of kultuurverband gebruik word, dien as 'n motivering vir die evaluering van die skaalekwivalensie van die toets. Skaalekwivalensie is 'n voorvereiste van toetse wat oor taal en kultuurgroepe heen gebruik word vir toelating of plasing vir tersiëre onderwys. Hierdie tipe van navorsing word veral as belangrik beskou as groot groepsverskille gevind word (AERA et al., 1999).

Navorsingsdoelwitte

Die oorkoepelende navorsingsoogmerk van die studie is om die skaalekwivalensie van die toets oor drie taalgroepe heen te evalueer.

Die spesifieke oogmerke is:

1. Om groepsverskille ten opsigte van betroubaarheid, standaardmetingsfout en item eienskappe, naamlik item-totaal korrelasie en moeilikheidsgraad te ondersoek;
2. Om die differensiële moeilikheidsgraad van die items oor groepe heen te ondersoek; en
3. Om die stukturele ekwivalensie van die toets vir die verskillende taalgroepe te ondersoek.

Die post hoc oogmerke is:

1. Om die stukturele ekwivalensie van die toets vir die verskillende taalgroepe na die verwydering van problematiese items te ondersoek.
2. Om groepsverskille na die verwydering van problematiese items te ondersoek.

Steekproef

Studente wat in 2003 en 2004 die leestoets as deel van 'n battery van toelatingstoetse van die NMMU afgelê het, het die aanvanklike beskikbaarheidssteekproef van 989 uitgemaak. Alle eerstejaarstudente skryf die toetse, hetsy vir toelating of vir navorsingsdoeleindes. 'n Toestemmingsbrief is deur al die studente geteken om toestemming tot navorsing te verleen. Geen internasionale studente is ingesluit in die steekproef nie, terwyl swart studente wat hulleself as eerstetaal Engelssprekend beskou, ook nie opgeneem is in die steekproef nie. Die loodsstudie is op hierdie steekproef uitgevoer.

Vorige navorsing op die gebied van ekwivalensie het aangetoon dat groot groepsverskille op die totaalpunte van toetse sowel as ongelyke getalle in die groepe tot sydigheid in die resultate kan lei (Sireci & Khaliq, 2002). 'n Frekwensie distribusie passing is dus op die oorspronklike steekproef gedoen en 'n steekproef van 563 studente is bekom. Die Engelse en Afrikaanse taalgroepe het etniese groeperinge soos Kleurling, Indiër en Wit ingesluit in proporsies wat die algemene studentepopulasie weerspieël. Tabel 1 gee 'n aanduiding van hoe suksesvol die steekproeftrekking was om distribusies en gemiddeldes op die totaaltellings van groepe min of meer dieselfde te hou, terwyl tabel 2 die proporsionele verteenwoordiging van die etniese groepe in die taalgroepe aandui.

Tabel 1
Beskrywende statistiek van steekproef per taalgroep: gemiddeldes, standaardafwyking en interval van totaaltellings

Taalgroepe	Gemiddeld*	n	Standaard Afwyking	Interval
Afrikaans	23,59	181	5,52	11-35
Afrika taalgroep	23,24	195	5,00	11-34
Engels	24,96	187	4,99	7-35
Totale groep	23,93	563	5,21	7-35

*Totaaltelling uit 35

Tabel 2
Beskrywing van die taalgroepe per etniese groep

Taalgroepe	Kleurling		Wit		Swart		Indiër	
	n	%	n	%	n	%	n	%
Afrikaans	80	44	100	56	0	0	0	0
Afrika taal	0	0	0	0	195	100	0	0
Engels	48	26	108	58	0	0	31	16

Om die groepsverskille tussen die taalgroepe na die verwydering van die DIF items te ondersoek is 'n totaal van 695 eerstejaar studente wat in die 2005 die toets afgelê het, en wat geregistreer en die Junie-eksamen afgelê het, geselekteer. Daar was 191 Afrikaanse studente, 194 sprekers van Afrika tale, meestal Xhosa, en 310 Engelse studente. Om te kontroleer vir verskille tussen die moeilikheidsgraad van programme oor die taalgroepe heen, het die navorser 'n analise gemaak van die proporsie studente in die verskillende taalgroepe wat ingeskryf is by verskillende fakulteite. Verskille het nie tussen die drie taalgroepe ten opsigte van hierdie faktor bestaan nie

Meetinstrument

Die papier weergawe van die 'ACCUPLACERTM Reading Comprehension' toets is gebruik vir die studie. Die toets is ontwikkel deur die Educational Testing Services (ETS) in die VSA en word tans besit en versprei deur die College Board. Die toets is met toestemming, aangepas vir die SA konteks deur sekere van die itembewoordigings te verander. Die toets bestaan uit twee hooftipe vrae en 35 items. Die eerste tipe vrae bestaan uit leesstukke gevolg deur vrae, terwyl die tweede tipe uit vrae bestaan wat verbande tussen sinne ondersoek. Inhoud uit 'n verskeidenheid algemene en akademiese areas is gebruik. Die prosessering van inligting bestaan uit, onder andere, eksplisiete stellings wat verband hou met die hoofgedagte, implisiete stellings wat verband hou met die hoofgedagte, afleiding en toepassing. Ten spyte hiervan word die konstruk as eendimensioneel beskou en word 'n Cronbach Alpha vir die totale toets gerapporteer en nie vir verskillende subskale nie (College Board, 1993). Die toets is uitvoerig vir geldigheid, betroubaarheid en kruiskulturele geldigheid in die VSA verband geëvalueer (College Board, 1993). Die toets het 'n bevredigende Cronbach Alpha van 0,86 vir 'n totale NMMU populasie (Davies, 2003). Korrelasies met akademiese prestasie wissel van 0,34 – 0,52 in verskillende fakulteite en van laag negatief tot hoog positief vir verskillende taal- en etniese groepe. Die hoogste korrelasies is vir wit en Engelssprekende studente gevind (Koch, 2002).

Dataontleding

Dataontleding is gedoen in Excel, Office 2000, SPSS weergawes 11,5 en 12 en Statistica 6. Die volgende analises is per navorsingsdoelwit gedoen:

Om groepsverskille ten opsigte van betroubaarheid en item eienskappe, naamlik item-totaal korrelasie en moeilikheidsgraad te ondersoek, is die Cronbach Alpha's van die drie taalgroepe bereken. Die item-totaal korrelasies en moeilikheidsgraad (p-waardes) per item is bereken, waarna die gemiddelde item-totaal korrelasie en p-waardes per taalgroep bereken is. Die gemiddeldes is nie statisties vergelyk nie, slegs beskrywend. Daarna is die item-totaal korrelasies en p-waardes van die verskillende taalgroepe gekorreleer as 'n aanduiding van die relatiewe orde van item eienskappe in die verskillende groepe. Die benadering om item-totaal korrelasies te vergelyk, word gebruik as 'n ondersoek van moontlike verskille in die konstruk oor groepe heen (Van de Vijver & Lueng, 1997). Die Pearsonkorrelasie is gebruik.

Om die differensiële funksionering van die moeilikheidsgraad van die items oor groepe heen te ondersoek, is twee tegnieke gebruik, naamlik die Delta-DIF indeks metode, en logistiese regressie analise. Twee tegnieke is gebruik omdat die analise van DIF welbekend is vir onstabiele resultate oor metodes en steekproewe heen (Robin et al, 2003). Twee tegnieke verleen dus meer interne geldigheid aan die studie indien 'n hoë mate van ooreenstemming tussen die twee metodes gevind word. Die groep wie se eerste taal die taal van die toets is, in die geval van hierdie studie die Engelse groep, word as die verwysingsgroep beskou, en die ander groepe (wie se eerste taal nie die taal van die toets is nie), die fokus groepe.

Die eerste metode, naamlik die Delta-DIF indeks metode, is verwant aan die delta-plot metode waar die p-waarde (moeilikheidswaardes) van die items per groep gekarteer word op 'n grafiek en dan vergelyk word. Om dit te doen word die p-waardes van elke taalgroep apart linieêr getransformeer na z- waardes op 'n skaal van M = 13 en SD = 4, ook genoem ETS delta waardes (Missisipi State University, 2004). Met hierdie prosedure dui laer delta waardes moeilike items aan en hoër waardes, makliker items. 'n Delta-DIF indeks word dan per item as 'n aanduiding van verskille tussen die taalgroepe bereken (sien Robin et al., 2003 vir die formule). 'n Waarde van 1,5 delta eenhede is as 'n aanduiding van DIF gebruik (Robin et al., 2003). 'n Negatiewe waarde het beteken dat die item die fokus groep bevoordeel en 'n positiewe waarde, dat die item die verwysingsgroep bevoordeel. 'n Aparte Delta-DIF indeks is vir die Engels-Afrikaans en die Engels–Afrikataal groepe bereken. Die navorsingshipoteses was dat daar verskille sou wees in die Delta-DIF indekse van die Engels-Afrikaans en die Engels–Afrikataal groepe vir al die items i, en die nul-hipoteses dat daar geen verskille sal wees nie.

Die tweede metode, logistiese regressie (LR), bereken die waarskynlikheid van 'n korrekte respons per item. Zumbo (1999) se prosedure en sintaks vir binêre items is gebruik om die analise met behulp van SPSS weergawe 11,5 te doen. Die afhanklike veranderlike was die itemtelling, 0 vir verkeerd en 1 vir korrek. Die onafhanklike veranderlikes was die totaal telling as die kondineringsveranderlike (*conditioning variable*), en groep lidmaatskap en die interaksie tussen groep lidmaatskap en leesvermoë as die DIF veranderlikes, en is stapsgewys in die analise ingevoer. Die nul-hipoteses was dat die waarskynlikheid om 1 te behaal op 'n item 'n funksie van die snydingslyn en die totaal telling sou wees (dus, slegs leesvermoë speel 'n rol of 'n item reg of verkeerd is) , terwyl die navorsingshipoteses was dat die waarskynlikheid 'n funksie sou wees (1) van die snydingslyn en groep lidmaatskap en (2) van die snydingslyn en die interaksie tussen groep lidmaatskap en leesvermoë (iets anders as leesvermoë speel 'n rol in die reg of verkeerd beantwoording van 'n item). Aparte vergelykings is vir die Engels-Afrikaans en die Engels–Afrikataal groepe gedoen.

'n Beduidende verskil – Chi-kwadraat (DIFF Chi-kwadraat), kritiese waarde 9,55 ($p < 0,01$), sowel as die R^2 verskil ($R^2 \Delta$) tussen die eerste en derde stappe van die analise is gebruik om

moontlike DIF te identifiseer. Om te kontroleer vir tipe-1 fout is slegs effekgrootes van 0,035 < R^2 Δ < = 0,060 as 'medium DIF' en R^2 Δ > 0.060 as 'groot DIF' gebruik om die nul hipetese te verwerp (Jodoin & Gierl, 2001), ongeag die DIFF Chi-kwadraat se beduidendheid. Uniforme DIF is gevind as τ^2 <> 0 en τ^3 = 0 en nie-uniforme DIF as τ^3 <> 0 ongeag die waarde van τ^2. As τ^2 < 0 was, het die item die verwysingsgroep bevoordeel, en as τ^2 > 0 was, het die item die fokus groep bevoordeel. In die geval van nie-uniforme DIF het die item lede van die verwysingsgroep met 'n hoë vermoë en lede van die fokus groep met 'n lae vermoë bevoordeel as die τ^3 < 0, en andersom as τ^3 >0.

Om die strukturele ekwivalensie van die toets vir die verskillende taalgroepe te ondersoek, is gebruik gemaak van geweegde meerdimensionele verskaling. In simulasie studies is gevind dat die metode meer sensitief is vir die effek van kleiner DIF as byvoorbeeld, faktoranalise en strukturelevergelykings modellering (Sireci & Khaliq, 2002). Die metode is ook met welslae gebruik om die effek van DIF op strukturele ekwivalensie te bestudeer (Robin et al., 2003). Die metode vereis nie die spesifisering van toets-strukture a priori nie, maak geen aannames oor die verwantskap tussen items nie en hou die voordeel in dat 'n gemeenskaplike struktuur te gelykertyd op meer as een groep (of toetse) gepas kan word (Sireci, Patsula, Hambleton, 2005).

Meerdimensionele verskaling gebruik nabyhede tussen objekte (items in die geval van hierdie studie) as invoerdata (dit wil sê, 'n getal wat aandui hoe ver of hoe naby objekte aan mekaar is), en 'n ruimtelike voorstelling van 'n geometriese konfigurasie van punte, as uitvoer. Nabyhede kan vir data wat nie in hulle oorspronklike vorm nabyhede is nie bereken word deur 'n profiel van ongelyksoortighede of gelyksoortighede af te lei. Dit word gedoen deur die korrelasies tussen data te bereken of die afstande tussen die stimuli te kwadreer (Wish & Carroll, 1974). In kleiner datastelle word die Euclideaanse formule as meer geskik beskou as die inter-item korrelations (sien Robin et al., 2003). In hierdie studie is inter-item korrelations gebruik en die drie matrikse van die drie taalgroepe is met die data-hanteerder in SPSS in een datastel bymekaar gevoeg.

Meerdimensionele verskaling kan vir 'n enkele groep of veelvuldige groepe gedoen word. In die geval van 'n enkele groep word 'n matriks van geobserveerde ongelyksoortighede tussen item j en j', Δ = $\{\delta_{jj'}\}$ in 1, 2 ... or R dimensionele ruimte by wyse van die Euclideaanse afstand gemodelleer (sien Robin et al., 2003 vir die formule). Hierdie model voorsien 'n voorstelling van die geobserveerde data in enige R-dimensionele ruimte, byvoorbeeld moontlike oplossings van een to vyf dimensies, en stel die struktuur van die toets voor. Die items word dan óf grafies óf by wyse van hulle koordinate voorgestel, x_{jr} vir item j on dimensie r (r = 1...R). Die dimensies van die oplossing wat aanvaar word, word beskou as 'n voorstelling van die struktuur van die data.

In 'n veelgroep analise word meer as een matriks van ongelyksoortighede gemodelleer. Elke matriks stem met die groepe van belang, k = 1,2 ...K ooreen, in hierdie studie die drie taalgroepe. Die matrikse word dan gelyktydig in 'n gemeenskaplike 2, 3,...R-dimensionele ruimte met die volgende geweegde Euclideaanse afstandsformule gemodelleer (om 'n gemeenskaplike struktuur te pas):

$$d_{jj}^{k} = \sqrt{\sum_{r=1}^{R} w_r^k \, (x_{jr} - x_{jr})^2} \qquad (1)$$

waar w_r^k ooreenkom met die gewig op dimensie r vir groep k (Robin et al., 2003). Die gewigte word gebruik om die strukturele verskille tussen groepe te evalueer. WMDS verskaf dus 'n metode om die dimensionaliteit wat response onderlê, te evalueer en vas te stel of die dimensionaliteit dieselfde is oor groepe heen. 'n Gemeenskaplike struktuur word gelyktydig op al die groepe gepas, en groepsgewigte word dan gebruik om verskille te evalueer (Sireci, et al., 2005). Die gewigte is 'n aanduiding van die mate waartoe die gemeenskaplike struktuur gewysig moet word om die data van die bepaalde groep die beste te pas. Indien die patroon van gewigte tussen groepe verskil en een of meer van die groepe 'n gewig van naby zero op 'n dimensie het waarop 'n ander groep 'n redelike groot gewig het, word verskille aanvaar. Dit is egter belangrik dat die verskille interpreteerbaar moet wees voordat die verskille aanvaar word (Sireci & Khaliq, 2002).

Vir hierdie studie is die INDSCAL WMDS model in SPSS 11.5 hiervoor gebruik.

Meerdimensionele verskaling verskaf nie statistiese toetse om die beste oplossing te selekteer of om te besluit oor verskille tussen groepe nie en hipoteses word nie vir hierdie tegniek geformuleer nie. Daar is wel 'n aantal praktiese reëls wat ten opsigte van passingsindekse, naamlik die STRESS en R2 indekse, en die interpretasie van dimensies geld om sodoende besluite oor die mees geskikte oplossing en groepsverskille te neem (Sireci et al., 1998). Die reëls word in die volgende stappe wat vir hierdie studie gevolg is, gereflekteer:

a) Die STRESS passingsindeks en R^2 indeks is gebruik om die finale model te selekteer, dit wil sê, of twee of drie of meer dimensies die data die beste voorstel. 'n Skerp verbetering in passing by 'n spesifieke punt van passing, ook genoem 'die knakpunttoets', is gebruik as die kriterium om 'n oplossing te aanvaar. Passing verbeter wanneer die STRESS indeks skerp daal en die R^2 indeks styg en beide daarna afplat (Sireci et al., 1998).
b) Om oor die strukturele verskille tussen taalgroepe te besluit, is 'n aantal stappe gevolg om die dimensies en die verskille tussen die groepe te interpreteer (Robin et al., 2003; Sireci & Khaliq, 2002). Indien die patroon van gewigte tussen die groepe verskil en een of meer van die groepe 'n gewig van naby zero op 'n dimensie gehad het waarop 'n ander groep 'n redelike groot gewig het, is verskille aanvaar. Verder is die item koördinate van elke dimensie in Excel georden en koördinate van groter as ¦1¦ is gebruik vir die interpretasie van die dimensie. Die p-waardes, die diskriminasie waardes van die items en die DIF terme van die Delta-DIF indeks en die LR is ook met die item koördinate gekorreleer. Hierdie korrelasies is gebruik om redes vir verskille tussen die groepe te vind en hierdie verskille te interpreteer. Slegs korrelasies van 0,6 en hoër is as van praktiese waarde vir interpretasie aanvaar in navolging van Robin et al., (2003). Verskille tussen groepe word aanvaar wanneer die patroon van gewigte tussen groepe aansienlik verskil en die verskille interpreteerbaar is (Sireci & Khaliq, 2002).

Vir die post hoc analises is geweegde meerdimensionele verskaling na die verwydering van die DIF items herhaal. Groepsverskille is met behulp van beskrywende statistiek en standaard afwyking ondersoek. Die verskille is op 'n nuwe steekproef wat die toets in 2005 afgelê het, ondersoek.

RESULTATE

Groepsverskille ten opsigte van betroubaarheid en item eienskappe

Geen groepsverskille is ten opsigte van die Cronbach Alpha's gevind nie. Die betroubaarheid het gewissel van 0,78 vir die Engelse en Afrika taal groepe tot 0,83 vir die Afrikaanse groep en was dus aanvaarbaar in al drie groepe (Koch, 2005). Die gemiddelde item-totaal korrelasies (van 0,26 tot 0,31) en p-waardes (van 0,67 tot 0,71) van die verskillende groepe was ook dieselfde. Terwyl daar groot ooreenstemming was tussen die

groepe ten opsigte van die p-waardes van die items (r = 0,91 tot 0,95), het die item- totaal korrelasies van die Engelse groep egter min ooreenstemming getoon met dié van die Afrika taal en die Afrikaanse groepe (r = 0,01 en 0,22, p > 0,05). Dié van die Afrikaanse en Afrika taal taalgroepe het egter sterk ooreenkomste getoon (r = 0,70, p < 0,05). Die patroon was dus dieselfde vir die oorspronklike steekproef soos gerapporteer onder die loodsstudie. Die laaste bevinding dui op die moontlikheid van strukturele inekwivalensie en dien as 'n verdere motivering vir die studie.

Differensiële moeilikheidsgraad funksionering van die items oor groepe heen

In die Delta-DIF Indeks metode is die nul hipotese van geen DIF vir tien items verwerp. Daar was 'n gelyke aantal items wat die Engelse (5) en die ander twee groepe (5 in totaal) bevoordeel het (Tabel 3).

Tabel 3
Dif items: delta-dif indeks metode

Analise groepe	Rigting van voordeel	Aantal items	Items
Afrikaans-Engels	Engels	2	18,27
	Afrikaans	2	21,35
Afrika taal-Engels	Engels	4	16,18,19,20
	Afrika taal	4	2,4,24,35

In die Logistiese regressie ontleding is altesaam 18 items met DIF gevind. Tien items het matige DIF openbaar, en 8 items klein DIF. Die nul hipotese van geen DIF is slegs ten opsigte van items met matige DIF verwerp. Terwyl drie items die nie-engelse groepe bevoordeel het, het die ander óf die hele Engelse groep bevoordeel óf hulle was nie- uniforme DIF items wat die Engelse studente met hoë totaalpunte op die toets en nie-engelse studente met lae totaalpunte bevoordeel het. Dieselfde patroon is ten opsigte van die agt items met lae DIF gevind (Tabel 4).

Tabel 4
Matige dif items: logistiese regressie metode

Vergelykingsgroepe	Tipe van DIF	Rigting van voordeel	Getal items	Items
Afrikaans – Engels	Uniform	Engels	2	18,27
		Afrikaans	1	21
	Nie-uniform	HV* Engels LV** Afrikaans	3	20, 33, 35
		HV Afrikaans LV Engels	0	
Afrika taal – Engels	Uniform	Engels Afrika taal	3 2	16, 18, 19 2,14
	Nie-uniform	HV Engels LV Afrika taal	1	20
		HV Afrika taal LV Engels	0	

* Hoër Vermoë
** Laer Vermoë

Nege van die tien items wat deur die Delta-DIF Indeks metode geindentifiseer is, is dus ook met LR as beduidende DIF geidentifiseer. Die ander item was wel geidentifiseer maar was nie statisties beduidend nie. Verder het die mate van ooreenstemming tussen die twee metodes ondersteuning verleen aan die interne geldigheid van die bevinding dat 'n groot proporsie (10) van die 35 items DIF openbaar (Robin et al., 2003). Terwyl die rigting van die DIF in die items wat deur die Delta-DIF Indeks metode geidentifiseer is, aandui dat 'n gelyke aantal items die verwysing en fokus groepe bevoordeel, het die items wat deur LR geidentifiseer is meerendeels Engelse studente met hoë totaalpunte op die toets en nie-engelse studente met lae totaalpunte bevoordeel. In totaal is die nul hipotese van geen DIF vir 10 items verwerp. Hierdie items vorm 'n groot proporsie van die totale getal items en dui reeds op ernstige probleme met die skaalekwivalensie van die toets.

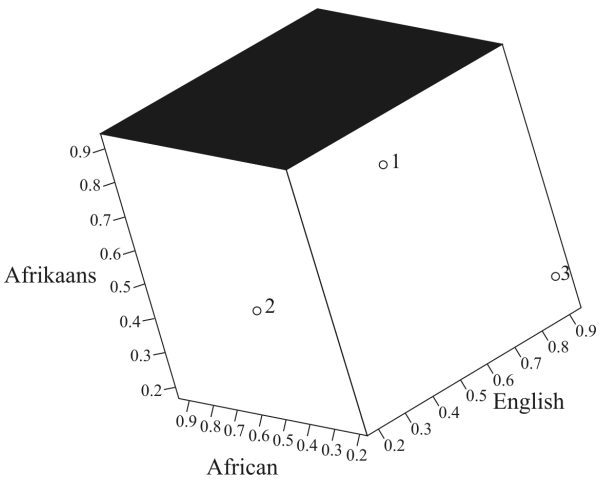
Strukturele ekwivalensie van die toets vir die verskillende taalgroepe

Die verskillende stappe wat onder "Dataontleding" bespreek is, is gevolg in die selektering van die beste oplossing. Op grond van die kriteria is drie dimensies geselekteer. Die STRESS statistiek het van 0,34 tot 0,16 gedaal van die tweede tot die derde dimensie, terwyl die R^2 gestyg het van 0,82 tot 0,93. Verskille het daarna afgeplat. Groot strukturele verskille tussen die drie taalgroepe is aangedui deur die verskille in die gewigte per taalgroep in die drie dimensies. Tabel 5 illustreer die verskille in terme van die verskil in gewigte oor groepe heen.

Tabel 5
Die gewigte per taalgroep en % variansie deur die dimensies verklaar

	Dimensies	Groepe			Variansie verklaar
		Afrikaans	Afrika taal	Engels	
Gewigte	1	0,49	0,39	0,91	0,41
	2	0,80	0,33	0,27	0,27
	3	0,20	0,81	0,19	0,24

Die eerste dimensie kan beskou word as relevant vir die Engelse groep, die tweede dimensie is relevant vir die Afrikaanse groep en die derde dimensie, vir die Afrika taalgroep. Met ander woorde: die betrokke dimensies lê prominent onderliggend aan die berekende ongelyksoortighede van die betrokke groepe. Die betrokke dimensies verklaar ook die grootste persentasie variansie in die getransformeerde data in die betrokke groepe. Die afstand tussen die drie groepe word geïllusteer deur die voorstelling van groep sentroïdes in figuur 1.



Figuur 1: Grafiese voorstelling van die afstand tussen die drie taalgroepe (1 = Afrikaans, 2 = Afrika taal, 3 = Engels)

Die stappe wat gevolg is om die dimensies te interpreteer, het aangedui dat verskille tussen die groepe ten opsigte van die

dimensies as 'differensiële moeilikheidsgraad van items oor dimensies vir die verskillende groepe' gekategoriseer kan word. Die item köordinate van die dimensies wat relevant was vir die verskillende groepe het sterker met die item p-waardes van die betrokke groepe gekorreleer as met dié van die ander groepe (Tabel 6).

Tabel 6
Korrelasies tussen die item koördinate en die p-waardes en item-totaal korrelasies per taalgroep

	Dimensies	Groepe		
		Afrikaans	Afrika taal	Engels
p-waardes	1	0,84	0,83	0,95
	2	0,86	0,72	0,73
	3	0,77	0,94	0,79
Item totaal korrelasies	1	0,62	0,41	-0,28
	2	0,06	0,11	-0,55
	3	0,43	0,34	-0,45

Die item köordinate van geen van die dimensies het meer as r = 0,60 met die item-totaal korrelasies van die Afrika taalgroep gekorreleer nie. Die item koordinate van die eerste dimensie (wat relevant was vir die Engelse groep) het sterker as 0,6 met die DIF terme van die items in Engels-Afrikaanse groepe gekorreleer, maar met nie met die DIF terme in die Engels-Afrika taalgroepe nie. Die item koordinate van die derde dimensie (wat relevant was vir Afrika taalgroep) het ook sterker as 0,6 met DIF terme van die LR analise in Engels-Afrikaanse groepe gekorreleer, maar met nie met die DIF terme in die Engels-Afrika taalgroepe nie (tabel 7). Die item koordinate van die tweede dimensie (wat relevant was vir die Afrikaanse groep) het nie met enige van die DIF terme van die items gekorreleer nie.

Tabel 7
Korrelasies van die koördinate van die 3-dimesionele oplossing met dif terme

Vergelykingsgroepe	Dimensies	Korrelasies van die koördinate met DIF terme		
		DIF indeks	LR groep helling	LR interaksie helling
Afrikaans-Engels	1	0,58	-0,70	0,64
	2	-0,05	-0,42	0,44
	3	0,35	-0,68	0,65
Afrikataal-Engels	1	0,22	-0,35	0,30
	2	-0,03	-0,32	0,31
	3	-0,35	-0,26	0,35

By nadere ondersoek by wyse van die ordening van die item köordinate van die dimensies – die items met köordinate groter as ¦1¦was meestal DIF items- het dit geblyk dat die DIF items wat dimensies 1 en 2 getipeer het, ooreengestem het met die items wat hoër item-totaal korrelasies gehad in die Engelse groep, maar met dié wat laer item-totaal korrelasies gehad het in die Afrikaanse groep. Met ander woorde, die DIF items het daarin geslaag om te onderskei tussen goeie en swak lesers in die Engelse groep (in terme van die totaalpunt op die toets), maar nie in die Afrikaanse groep nie. Dieselfde patroon is ten opsigte van die Afrika taalgroep gevind, selfs al het die DIF terme swakker as 0,6 met die item koordinate van die dimensies gekorreleer.

Die strukturele verskille tussen die Afrikaanse en Engelse groepe kon dus gedeeltelik in terme van DIF verklaar word, maar DIF het nie die strukturele verskille tussen die Afrika taalgroep en die Engelse groep verklaar nie. Daar is dus verwag dat die verwydering van die DIF items tot meer strukturele ekwivalensie sou lei, maar dat alle strukturele verskille nie sou verdwyn nie. Hierdie vermoede is in die post hoc ontledings ondersoek.

Post hoc analises
Strukturele ekwivalensie van die toets vir die verskillende taalgroepe na verwydering van DIF items

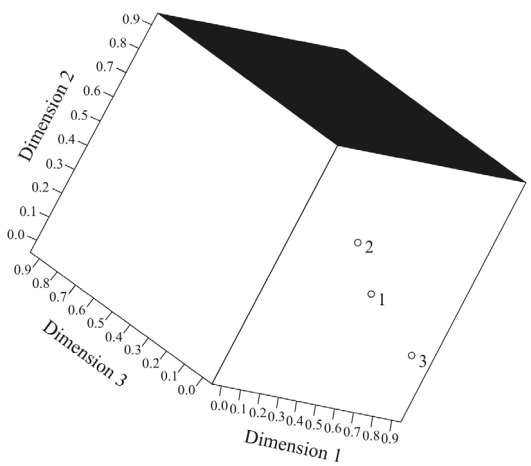
Die 10 DIF items met medium DIF waardes is uit die toets verwyder. Dieselfde stappe as voorheen is gevolg deurdat korrelasie matrikse vir die drie groepe sonder die DIF items geskep is, die matrikse bymekaar gevoeg is en 'n analise van veelvuldige meerdimensionele verskaling gedoen is.

Op grond van die afname in die STRESS waardes (van 0,17 na 0,13) en die toename in die R^2 waardes (van 0,93 tot 0,95), is drie dimensies ook in hierdie analise as die mees bevredigende oplossing aanvaar. Daarna was daar 'n afplatting in die waardes. Die eerste dimensie het 77% van die totale variansie in die getransformeerde data verklaar en was dus 'n sterk dimensie. Die ander twee dimensies het onderskeidelik 10% en 7% verklaar. Alhoewel die laaste twee dimensies slegs 'n klein persentasie van die totale variansie verklaar het, het die passingsindekse tog wel aangedui dat hierdie oplossing die mees bevredigende is. Al drie dimensies kon ook geinterpreteer word en het bygedra tot die dissektering van groepsverskille. Dit het verdere ondersteuning verleen aan die aanvaarding van hierdie oplossing (sien Sireci & Khalig, 2002 in hierdie verband).

Tabel 8 gee 'n aanduiding van die verskille tussen die drie taalgroepe in terme van die verskille in gewigte oor groepe heen, terwyl figuur 2 dit visueel demonstreer.

Tabel 8
Die gewigte per taalgroep en % variansie deur die dimensies verklaar met dif items verwyder

	Dimensies	Groepe			Variansie verklaar
		Afrikaans	Afrika taal	Engels	
Gewigte	1	0,89	0,79	0,96	0,77
	2	0,29	0,45	0,17	0,10
	3	0,28	0,36	0,05	0,07



Figuur 2: Grafiese voorstelling van die afstand tussen die drie taalgroepe na verwydering van DIF items (1 = Afrikaans, 2 = Afrika taal, 3 = Engels)

Daar was duidelik 'n verbetering in die strukturele ekwivalensie na die verwydering van die DIF items. Geringe strukturele verskille tussen die drie groepe het egter voortgeduur. Die eerste dimensie was relevant vir al drie groepe, maar veral vir die Engelse groep. Alhoewel die gewigte van die ander twee groepe nie naby aan zero was nie, kon verskille met die Engelse groep op hierdie dimensie interpreteer word (sien Sireci & Khaliq, 2002). Dit kan dus aanvaar word dat die groepe tot 'n geringe mate selfs op hierdie sterk en belangrike dimensie verskil het. Dimensies 2 en 3 was relevant vir die Afrika en Afrikaanse taalgroepe. Dimensie 2 het minder relevansie vir die Engelse groep gehad, terwyl dimensie 3 geen relevansie vir die Engelse groep gehad het nie.

Die stappe wat geneem is om die dimensies te interpreteer, het aangedui dat dimensie 1 beskou kan word as "algemene item moeilikheidsgraad" in al drie groepe, maar "item diskriminasie" slegs in die Engelse groep (Tabel 9).

Tabel 9
Korrelasies tussen die item koördinate van die toets met dif items verwyder, en die p-waardes en item-totaal korrelasies per taalgroep

	Dimensies	Groepe		
		Afrikaans	Afrika taal	Engels
p-waardes	1	0,93	0,90	0,97
	2	0,64	0,69	0,57
	3	-0,56	-0,55	-0,49
Item totaal korrelasies	1	0,40	0,35	-0,80
	2	0,08	0,07	-0,39
	3	-0,70	-0,55	0,06

Die item koördinate van die dimensie het ook gekorreleer met DIF terme. Weereens wil dit voorkom of die moeiliker items met geringe DIF wat hierdie dimensie tipeer wel diskrimeer tussen goeie en swak lesers in die Engelse groep, maar dat daardie items nie tot dieselfde mate diskriminerend is in die nie-Engelse groepe nie. Die geringe verskille tussen die drie groepe op hierdie dimensie kan dus toegeskryf word aan differensiele item diskrimasie oor die drie groepe heen. Dimensie 2 was "algemene item moeilikheidsgraad" in al drie groepe, effens meer so in die geval van die Afrika taalgroep. Dimensie 3 was "item moeilikheidsgraad" en "item diskriminasie" in die Afrika taal en Afrikaanse groepe. Die dimensie het geen relevansie vir die Engelse groep gehad nie. Die grootste verskille tussen die Engelse en nie-Engelse groepe was dus op hierdie 'swak' derde dimensie.

Daar kan dus tot die gevolgtrekking gekom word dat die verwydering van die matige DIF items gelei het tot 'n verbetering in strukturele ekwivalensie tussen die drie groepe, maar dat geringe strukturele verskille voortgeduur het. Dis was egter belangrik om vas te stel of hierdie geringe verskille praktiese implikasies het vir die gebruik van die toets oor die drie groepe heen.

Groepsverskille tussen die taalgroepe na die verwydering van die DIF items

Tabel 10 gee 'n aanduiding van die verskille tussen die taalgroepe ten opsigte van hulle gemiddeldes op die 25 items wat nie as DIF geïdentifiseer is in die vorige afdelings nie, sowel as ten opsigte van hulle akademiese resultate in die eerste semester van 2005.

Alhoewel die verskille tussen die nie-Engelse en Afrika taalgroepe kleiner is as met die toets wat die DIF items insluit, is die verskille tussen die twee groepe steeds aansienlik met die Afrika taalgroep wat heelwat laer presteer as die Engelse groep. Die standaard afwyking verskil tussen die Afrikaanse en Engelse groepe (die standaard afwyking van die Engelse groep is as verwysing gebruik) is dieselfde as die toets met die DIF items met die Afrikaanse groep wat laer presteer as die Engelse groep. Hierdie verskille word egter nie gereflekteer in die verskille op akademiese prestasie nie, waar die Afrikaanse en Engelse groep dieselfde vaar, en die verskille tussen die Engelse groep en Afrika taalgroep kleiner is as op die leestoets. Dit wil dus voorkom asof die voorgesette strukturele verskille op die toets sonder die DIF items wel praktiese implikasies inhou vir die gebruik van die toets oor die drie taalgroepe heen. Met ander woorde, indien sterk klem geplaas word op hierdie toets in besluitneming oor toelating sonder om strukturele verskille in ag te neem, bestaan die moontlikheid dat nie-engelstalige studente aansienlik benadeel sal word. Dit is ook belangrik dat die voorspellingsgeldigheid van die toets oor taalgroepe heen geëvalueer moet word. Die gevolgtrekkings word in die volgende afdeling bespreek.

Tabel 10
Groepsverskille op die leestoets na verwydering van dif items en eerste semester 2005 akademiese prestasie

Taal-groepe	Gemid-delde totaal op leestoets	Std afwyking	Standaard afwyking verskil van Engelse groep	Akade-miese prestasie	Std afwyking	Standaard afwyking verskil van Engelse groep
Afrikaans	13,30	2,68	0,59	61,32	13,95	0,01
Afrika taalgroep	12,16	8,58	1,09	50,78	10,80	0,76
Engels	14,63	2,86		61,42	14,04	

BESPREKING EN GEVOLGTREKKING

Die resultate van die empiriese afdeling van die studie het gedemonstreer dat die skaalekwivalensie van die toets oor taalgroepe heen gekompromiteer is. Daar was ook bewyse dat die toets nie 'n invariante konstruk in die Engelse en nie-Engelse taalgroepe meet nie selfs na die verwydering van die DIF items. Die gevolgtrekking kan dus gemaak word dat die tellings van die Engels eerstetaal en tweedetaalsprekers nie gebruik kan word om dieselfde afleidings oor groepe heen te maak ten opsigte die kontruk/domein wat in die toets gemeet word nie en dat die tellings van die twee groepe nie op dieselfde skaal geplaas kan word word vir vergelying nie. Verder is dit duidelik dat die geringe strukturele verskille tussen die groepe na die verwydering van die DIF items steeds op 'n praktiese vlak 'n impak het. Die voortgesette gebruik van die toets moet dus heroorweeg word, tensy die resultate van die groepe in 'n differensiële manier hanteer word soos, byvoorbeeld deur verskillende afsnypunte vir verskillende groepe te stel, of die toets aangepas word vir gebruik oor diverse groepe heen.

Die differensiële gebruik van die toetstellings van die verskillende groepe is op hierdie stadium die benadering wat aan die NMMU gebruik word, terwyl daar ook verdere aanpassings aan die toets aangebring is. Die toets word ook altyd gebruik as deel van 'n profiel van toetstellings wat studente se matriekpunte insluit, en besluitneming word nooit op grond van hierdie toetstellings alleen gemaak nie (Koch & Foxcroft, 2003).

Op 'n meer algemene vlak kan daar geargumenteer word dat die strukturele verskille wat op hierdie toets oor die Engelse en nie-engelse groepe gevind is, bloot 'n weerspieëling is van die realiteit van alle leesaktiwiteite wat op universiteitsvlak

plaasvind omdat meeste van die leesmateriaal in elk geval in Engels is. Dit mag dus net prakties wees om hierdie realiteit te te aanvaar en voort te gaan om die toets, en andere soos die, as geldige en realistiese metings van leesvaardigheid in die SA konteks te gebruik. Die benadering om die teenwoordigheid van sydigdigheid in te bou in die konteks van toetsing word egter gekontesteer deur navorsers soos Valdés and Figueroa (1994) and Pennock-Roman (1998; 1999), wat argumenteer vir geldiger toetsing van tweetalige studente en kinders in die VSA, 'n argument wat ernstig opgeneem moet word in die konteks van SA. Die teoretiese raamwerk wat in hierdie studie aanvaar is, sluit ook so 'n benadering vanuit 'n psigometriese, en uiteindelik 'n regsoogpunt, uit.

Dit moet duidelik gestel word dat die resultate van die empiriese deel van die studie nie veralgemeen kan word na ander toetse of selfs die SA populasie in die algemeen nie. Die metode van steekproeftrekking wat gebruik is, sluit die moontlikheid van veralgemening uit, terwyl dit in die literatuur duidelik gestel word dat toetse vir ekwivalensie geëvalueer moet word in alle kontekste van gebruik en dat 'n toets wat in een konteks sydigheid openbaar nie noodwendig in 'n ander konteks sydigheid sal openbaar nie (Van de Vijver & Leung, 1997; Van de Vijver & Tanzer, 1998). Hierdie toets was ook vir die VSA konteks ontwikkel wat dit veral gevoelig maak vir evaluasies soos in hierdie studie, iets wat moontlik nie die geval mag wees by toetse wat vir en in hierdie land ontwikkel is nie. Die resultate van hierdie studie moet egter as 'n ernstige motivering dien vir soortgelyke studies op ander toetse wat in die veeltalige SA konteks gebruik word.

As 'n slotopmerking kan daar gestel word dat hierdie studie daarin geslaag het om aan te dui dat die praktyk om eentalige toelatingstoetse te gebruik vir toelating tot Hoër Onderwys in SA inderdaad problematies mag wees, en dat dit belangrik is om navorsing soos in hierdie studie op alle soortgelyke toetse uit te voer. Alternatiewe tot die gebruik van eentalige toelatingstoetse moet ook dringend ondersoek word. In dié verband kan die praktyk in Israel (Beller, 1994), waar toelatingstoetse in tot agt tale vertaal word, ten spyte van die feit dat die taal van onderrig Hebreeus is, met vrug ondersoek word. Sulke navorsing moet verder aangevul word met evaluerings van die voorspellings-sydigheid van toetse sowel as ondersoeke na die differensiële impak van toelatingsprosedures op verskillende taalgroepe.

VERWYSINGSLYS

Abrahams, F. (1996). *The cross-cultural comparability of the Sixteen Personality Factor Inventory* (16PF). Unpublished doctoral thesis, University of Pretoria, Pretoria, South Africa.

Abrahams, F. (2002). Fair usage of the 16PF (SA 92) in South Africa: A response to C. H. Prinsloo & I. Ebersohn. *South African Journal of Psychology*, *32*, 58-61.

Abrahams, F., & Mauer, K. F. (1999a). The comparability of the constructs of the 16PF in the South African context. *Journal of Industrial Psychology*, *25*, 53-59.

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 South African context. *South African Journal of Psychology*, *29*, 76-86.

Adamson, H.D. (1993). *Academic competence: Theory and classroom practice: Preparing ESL students for content courses*. New York: Longman Publishing Group.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association (also published 1985, 1974, 1969).

Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. *Educational Measurement: Issues and Practices*, *13*, 12-21.

Camilli, G. (1993). The case against item bias techniques based on internal criteria: do item bias procedures obscure test fairness issues? In P.W. Holland and H. Wainer (Eds.), *Differential item functioning*. (pp 397-413). Hillsdale: Lawrence Erlbaum Associates, Publishers.

Claassen, N. C. W. (1993). *Verslag oor die funksionering van die NSAG intermedier G in verskillende bevolkingsgroepe*. Pretoria: Raad vir Geesteswetenskaplike Navorsing.

Cleary, T.A. (1968). Test bias: prediction of grades of Negro and white students in integrated colleges. *Journal of Educational measurement*, *5* (2), summer 1968.

Cole, N.S.& Moss, P.A. (1989). Bias in test use. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed.). (pp. 201-220). London: Collier Macmillan publishers.

College Board (1993). *ACCUPLACER™: Computerized placement tests technical data supplement*. New York: College Entrance Examinations Board.

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. San Diego, CA: College Hill Press

Davies, C.L. (2003). *A psychometric evaluation of the equivalence of the paper-based companion tests and the ACCUPLACER Computerised Placement tests*. Unpublished MA dissertation, University of Port Elizabeth.

Ebel, R.L. & Frisbie, D.A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Employment Equity Act No 55 of 1998 (1998). *Government Gazette Vol. 400*, No. 19370. Cape Town, 19 October 1998.

Focus (UPE). *The Commemorative Edition*. (2004). Port Elizabeth: Universiteit van Port Elizabeth

Geisinger, K.F. (1996, September). *The testing of Hispanics in civil service settings*. Paper presented at the Personnel Testing Council of Metropolitan Washington, Washington, DC.

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

Helms-Lorenz, M. & Van de Vijver, F. (1995). Cognitive assessment in education in a multicultural society. *European Journal of Psychological Assessment*, *11*, 158-169.

Huysamen, G.K. (2002). The relevance of the new APA standards for educational and psychological testing for employment testing in South Africa. *South African Journal of Psychology*, *32* (2), 26-33

International Test Commission (2000). *Guidelines for adapting educational and psychological tests*. [Online] Available: http://www.intestcom.org/adapt_test.htm

Jodoin, M.G. & Gierl, M.J. (2001). Evaluating type 1 error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14* (4), 329-349.

Koch, E. (2002, September). *Commerce, science and Pharmacy entry criteria: 2002 development and tracking*. Unpublished report for APAP. University of Port Elizabeth, Suid Afrika.

Koch, E. (2003, March). *Results on the APAP language tests*. Paper presented at Language Testing Colloquium, University of the Vrystaat, Bloemfontein, South Africa.

Koch, E., & Foxcroft, C.D. (2003). A developmental approach to admissions testing: Admissions and placement standards development. *South African Journal of Higher Education*, *17* (3), 192-208.

Koch, E. 2005a. *Evaluating the equivalence, across language groups, of a reading comprehension test used for admissions purposes*. Unpublished D.Phil thesis, Nelson Mandela Metropolitan University, South Africa.

Koch, E. (2005b). *Group differences on a reading comprehension test: What are the practical implications of inequivalence across language groups*. Unpublished report for CAAR, HEADS, Nelson Mandela Metropolitan University, South Africa.

Koda, K. (1994). Second language reading research: Problems and possibilities. *Applied Psycholinguistics*, *15*, 1-28.

Meiring, D., Van de Vijver, F., Rothmann, S. & Barrick, M.R. (2005). Construct, item and method bias of cognitive and personality tests in South Africa. *South African Journal of Industrial Psychology*, *31* (1), 1 – 8.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed). (pp. 13-104). London: Collier Macmillan publishers.

Missisipi State University (2004). *Item bias/differential item functioning*. Chapter 16. Notes and supplement. [Online] Available: http://www2.mstate.edu/=dmorse/8993chap16.pdf)

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.

Pennock-Roman, M. (1998, August). *Measuring developed academic abilities using Spanish-language and English-language tests; PAEG/GRE relationships for Puerto Ricans who are more proficient in Spanish than in English*. (GRE Board Professional Report No. 89-01cP; ETS Research Report No. 98-40). Princeton, NJ: Educational Testing Service.

Pennock-Roman, M. (1999, June). *English proficiency and differences among racial and ethnic groups in mean SAT and GRE scores: A longitudinal analysis*. (GRE Board Professional Report No. 86-09cP; ETS Research Report No. 99-17). Princeton, NJ: Educational Testing Service.

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

Robin, F., Sireci, S.G., & Hambleton, R.K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, *3* (1), 1-20.

Sireci, S.G., Bastari, B., Xing, D., Allalouf, A.,& Fitzgerald, C. (1998). *Evaluating construct equivalence across tests adapted for use across multiple languages*. Paper presented at the Annual Meeting of the American Psychological Association (Division Research), San Francisco, CA.

Sireci, S.G., & Geisinger, K.F. (1998). Equity issues in employment testing. In J. Sandoval, C.L. Frisby, K.F. Geisinger, Scheuneman, and J. R. Grenier (Eds.), *Test interpretation and diversity. Achieving equity in assessment* (pp 105-140). Washington: American Psychological Association.

Sireci, S.G., & Khaliq, S.N. (2002). *Comparing the psychometric properties of monolingual and dual language test forms*. (Center for Educational Assessment Research No. 458). Amherst, MA: School of Education, University of Massachucetts Amherst.

Sireci, S.G., Patsula, L., & Hambleton, R.K. (2005). Statistical methods for identifying flaws in the adaptation process. In R.K. Hambleton, P.F. Merenda & C.D Spielburger (Eds), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93 – 116). New JerseY; Lawrence Erlbaum Associates, Inc.

Valdés, G., & Figueroa, R.A. (1994). *Bilingualism and testing. A special case of bias*. Norwood, NJ: Ablex Publishing corporation.

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

Van de Vijver, F., & Lueng, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks: Sage.

Van de Vijver, F., & Poortinga, Y.H. (2005). Conceptual and methodological issues in adapting tests. In R.K. Hambleton, P.F. Merenda & C.D Spielburger (Eds), *Adapting educational and psychological tests for cross-cultural assessment*. (pp. 39 – 64). New Jersey; Lawrence Erlbaum Associates, Inc.

Van de Vijver, F., & Tanzer, N.K. (1998). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47*, 263-279.

Wish, M., & Carroll, J.D. (1974). Applications of individual differences scaling to studies of human perception and judgment. In E.C. Carterette and M.P. Friedman (Eds.) *Handbook of Perception*, vol. 2. New York: Academic Press.

Yeld, N. (2001). *Assessment, equity and language of learning: Key issues for higher education selection in South Africa*. Unpublished PHD thesis, University of Cape Town, South Africa.

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