

Tobie van Dyk

Noordwes Universiteit (Potchefstroomkampus)

Geldigheid vanuit drie paradigmas beskou: 'n eenheid, of 'n veelheid van perspektiewe?

Opsomming

Taaltoetse is hedendaags die norm aan hoëronderwysinstellings, hetsy vóór of ná toelating. Die effek van sulke toetse, of die besluite wat op basis van die resultate daarvan geneem word, kan tot in- of uitsluiting bydra. In 'n post-1994 Suid-Afrika is dit vanselfsprekend van belang dat 'n konsep soos 'geldigheid' deeglik ondersoek word, omdat die impak van voorafgenoemde toetse verreikend kan wees. Hierdie artikel bespreek daarom vanuit 'n filosofiese perspektief op die grondslae van die toegepaste taalkunde die begrip 'geldigheid' en definieer dit vanuit drie paradigmas, te wete 'n tradisionele paradigma, die hedendaagse/konvensionele siening en 'n meer aanvaarbare, verruimde, standpunt daaroor.

Die kern van die argument is dat geldigheid terselfdertyd 'n attribuut van 'n toets kan wees en as konstitutiewe begrip ontsluit kan word deur verdere (regulatiewe) idees met betrekking tot taaltoetsing. Dit word aan die hand van 'n teoretiese raamwerk gedoen wat op die konseptuele werk van Albert Weideman gebaseer is. Hierdie raamwerk toon aan dat 'n veelheid van perspektiewe nodig is om toegepaste linguistiese of tegniese ontwerpe (soos taaltoetse) op verantwoordbare wyse te gebruik.

Sleutelwoorde: toegepaste taalkunde, taaltoetsing, geldigheid, validering, toetsattribuut, konstitutiewe begrippe, regulatiewe idees

Abstract

Language tests are the norm at institutions of higher education nowadays, either before or after admission. Such tests, or the decisions taken on the basis of the results, can contribute to inclusion or exclusion. In post-1994 South Africa it is obviously important that a concept such as 'validity' be investigated thoroughly, as the impact of the aforementioned tests may be far-reaching. This article therefore discusses the concept 'validity' from a philosophical perspective on the foundations of applied linguistics and defines it according to three paradigms, i.e. a traditional paradigm, the contemporary/conventional view and a more acceptable, extended

paradigm. The crux of the argument is that validity can be an attribute of a test and be explicated as constitutive concept by further (regulative) ideas about language testing. The conceptual work of Albert Weideman forms the theoretical framework for this article. This framework entails that multiple perspectives are necessary to ensure that applied linguistic or technical designs (such as language tests) are used in an accountable manner.

Keywords: applied linguistics, language testing, validity, validation, test attribute, constitutive concepts, regulative ideas

1. Inleiding en doel

Die gebruik van meetinstrumente aan universiteite is aan die orde van die dag op plaaslike en internasionale vlak en word onder andere aangewend om potensiaal te bepaal, vaardigheidsvlakke te meet, diagnoses te maak, vordering te monitor en ondersteuningsprogramme te rig. Taaltoetse vorm byna sonder uitsondering deel van toetsbatterye en die waarskynlike rede hiervoor is dat taal 'n integrale deel van onderwys is in die opsig dat dit 'n beduidende invloed op akademiese sukses het (Van Dyk, 2010:262-274; Yeld, 2001:249-250; McNamara, 1996:21). Yeld (2001:226) argumenteer dat taal eerstens gebruik word om te leer ('n heuristiese funksie) en tweedens om feite en kennis oor te dra/te produseer ('n produktiewe funksie) en as dit nie goed ontwikkel is nie, ly sukseskoerse daaronder. Cliff, Ramaboa en Pearce (2007:34) sluit hierby aan wanneer hulle aanvoer dat universiteite desnoods 'n verantwoordelikheid het om studente deeglik te toets alvorens hulle tot studies in hoër onderwys toegelaat word en uiteindelik ook daardie studente wat wel toegelaat word, toepaslik te ondersteun:

There is now more than anecdotal evidence that Higher Education institutions and admissions committees or panels are taking seriously the need for responsible, ethical and equitable approaches to admissions decisions, and a parallel need to make use of the multiple sources of information collected about applicants for the placement of those eventually registered into appropriate curricula.

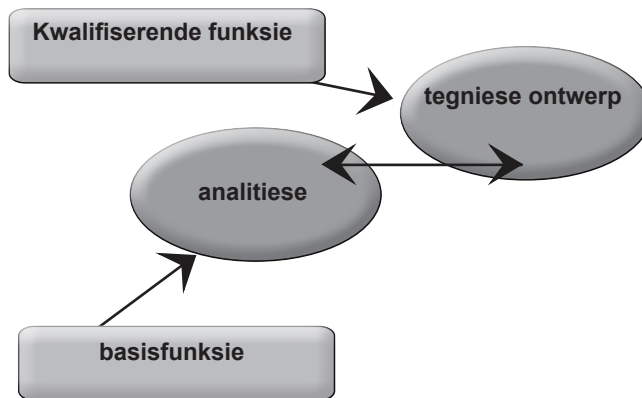
In hierdie aanhaling word begrippe soos ‘responsible’, ‘ethical’, ‘equitable’, ‘multiple sources’, en ‘appropriate’ gebruik, wat artikuleer met begrippe uit die konseptuele raamwerk wat hier onder as teoretiese vertrekpunt voorgelê sal word. Dit is gebaseer op die werk van Weideman (2006, 2007a, 2007b, 2009a, 2009b, 2009c) wat aanvoer dat toegepaste linguïste (die tipiese ontwerpers en ontwikkelaars van taaltoetse en taalkursusse) selde tyd maak om te reflekteer en teoretiese regverdiging te soek vir dit wat hul praktyk rig, aangesien daar altyd die een of ander dringendheid aan die werk van diesulkes is – oplossings vir taalprobleme (tegniese ontwerpe) moet vinnig bedink en meestal sonder uitsondering dringend geïmplementeer word (Weideman, 2006:71). In lyn met Cliff, Ramaboa en Pearce, en Weideman se argumente hier bo, is die doel van hierdie artikel daarom juis om te besin oor spesifieke konsepte wat verband hou met die geldigheid van taaltoetse.

2. Teoretiese raamwerk: die toegepaste taalkunde as dissipline van ontwerp

’n Verskeidenheid van perspektiewe is in ’n postmoderne aanpak nodig om oplossings vir probleme te bedink. Oplossings kan nie bloot ontwerp en geïmplementeer word sonder deeglike regverdiging daarvoor in terme van wetenskaplike teorie en analise nie en sonder om dit met die nodige verantwoordbaarheid en integriteit te doen nie. Weideman (2006:70) verwys in die verband veral na die werk van Schuurman (2005:26) wat hierdie punt bevestig wanneer laasgenoemde beweer dat “technology takes place in a historical, cultural, social and political context and that various groups within these contexts actively pursue various interests and goals.”

Hieruit blyk dit dat daar verskeie rolspelers is wat belange in tegniese ontwerpe (soos taaltoetse) het en dat hierdie rolspelers en die kontekste waarin hulle funksioneer dikwels in spanning met mekaar verkeer, veral waar hoërbelangaangeleenthede, soos toelating tot studie, ter sprake kom. Die gebruik van ’n konseptuele raamwerk wat as teoretiese- en eties-regverdigbare begroning vir die ontwerp kan dien, blyk dus ’n noodsaaklikheid te wees.

Weideman (2006:72) voer aan dat die toegepaste taalkunde oplossings vir taalprobleme bied in die vorm van tegniese ontwerpe, en dat hierdie oplossings twee funksies het: ’n leidende funksie (kwalifiserend) en ’n basisfunksie wat die teoretiese of empiriese fondasie vir die ontwerp bied. Hy oordeel verder dat die leidende of kwalifiserende funksie van ’n toegepaste linguïstiese oplossing te vinde is in die tegniese aspek daarvan, maar dat die ontwerp telkens ondersoek, verklaar of verander word vanuit die analitiese of teoretiese perspektief. Let daarop dat hierdie interpretasie of raamwerk nie veronderstel dat die teoretiese die ontwerp lei nie, maar dat dit eerder as rasionaal daarvoor aangebied word. Figuur 1 is ’n skematiese voorstelling van hierdie raamwerk en is oorgeneem uit Weideman (2007a:15).



Figuur 1: Leidende en basisfunksies van die toegepaste linguistiek

Weideman (2006:73) beweer verder dat die raamwerk in Figuur 1 hier bo as definisie vir die toegepaste linguistiek kan dien, aangesien dit al die kernaspekte van die vorming van die toegepaste taalkunde konseptueel sou kon byeenbring en veral ook ruimte laat vir 'n postmoderne en opvolgende beskouings daarvan. Hy waarsku egter dat daar nie, soos in die vroeëre tradisies en generasies van die toegepaste taalkunde, in die huidige tydgleuf bloot weer 'n verskuiwing plaasvind na die anderkant van die spektrum nie, maar dat 'n sobere inslag tot die ontwerp van oplossings gevolg moet word. Daar moet dus 'n balans tussen die konstitutiewe (wat die basisfunksie van die ontwerp insluit) en die regulatiewe (die ontslote, leidende funksie van die ontwerp) wees. In hierdie verband voer hy aan dat “[one] should carefully weigh a variety of potentially conflicting demands, and opt not only for the socially most appropriate, but also for a frugal solution” (Weideman, 2006:83). Dit is daarom byvoorbeeld soms nodig om die politieke oorwegings swaarder te laat weeg as die koste van die oplossing wat ontwerp word, en ander kere is dit weer nodig om ter wille van die bruikbaarheid of relevansie van die ontwerp tevrede te wees met 'n laer betroubaarheidswaarde.

In die verband is dit belangrik om te let op die respektiewelik terugskouende en antisipatoriese aard van die raamwerk wat Weideman (sien hier bo) as definisie vir die toegepaste taalkunde voorhou. Hy oordeel dat wanneer 'n oplossing ontwerp word, dit altyd vanuit die twee genoemde perspektiewe gedoen moet word: antisipatories in die opsig dat die ontwerper byvoorbeeld die moontlike gevolge van die gebruik daarvan in ag moet neem (as deel van die leidende of regulatiewe aard daarvan) en terugskouend in die opsig dat daar retrospektief 'n teoretiese begronding moet wees as deel van die basis of konstitutiewe aard daarvan. In 'n latere artikel van Weideman (2009b) bied hy 'n uitbreiding hiervan aan en gee 'n meer volledige beskrywing van welke elementêre toegepaste linguistiese grondbegrippe van belang mag wees. Die begrippe word as't ware gegeneer deur die verbande (analogieë, in die vorm van retrospektiewe of antisiperende momente) tussen die tegniese dimensie en 'n veelheid ander aspekte van die ontwerp. Die belang hiervan is daarin opgesluit dat elke grondbegrip uitdrukking gee

aan 'n stel tegniese beginsels vir die ontwerp van oplossings vir taalprobleme. Figuur 2 (uit Weideman, 2009b:66) is 'n illustrasie hiervan.

Toegepaste linguistiese ontwerp	Aspek / funksie / dimensie / ervaringsmodus	Tipe funksie	Terugskouende / antispatoriese moment
word gebaseer op	Numeriese	Konstitutiewe	eenheid in 'n veelheid van ontwerpbeginsele / bronne van bewyslewing
	Ruimtelike		Reikwydte
	Kinematiese		interne konsistensie (tegniese betroubaarheid)
	Fisiese		interne effek / krag (geldigheid)
	Organiese		differensiasie en aanpassing van ontwerp
	Psigiese		wilskrag en verbeelding
	Analitiese	Funderende	rasonaal vir ontwerp
word gekwalifiseer deur	Tegniese	kwalifiserende / leidende funksie (van die ontwerp)	
word ontsluit deur	Linguale	Regulatiewe	artikulasie van ontwerp met bloudruk / plan
	Sosiale		implementering / administrasie / interaktiwiteit
	Ekonomiese		tegniese nut / bruikbaarheid, matigheid / spaarsamigheid
	Estetiese		harmonisering, regstel van belyningsfoute / herstel van balans
	Juridiese		deursigtigheid, verdedigbaarheid, regverdigheid, legitimiteit / egtheid
	Etiese		verantwoordbaarheid, sorgsaamheid, diens
	Oortuigings		verbintenis, vertrou

Figuur 2: Konstitutiewe en regulatiewe momente in toegepaste linguistiese ontwerpe

Bostaande opsomming dui daarop dat toegepaste linguistiese ontwerpe, as oplossings vir probleme, nie slegs in diens staan van daardie persone of belanghebbendes waarvoor dit ontwerp word nie en nie altyd vanuit daardie perspektief benader moet word nie, maar dat die onderskeie ontwerpbeginsels altyd ook in wisselwerking met mekaar beskou moet word.

In die volgende deel van die artikel word die begrip geldigheid binne konteks van bostaande raamwerk bespreek. Geldigheid het in terme hiervan in die eerste plek te doen met die tegniese krag of effek wat van die meetinstrument vereis word 'n instrument beskik slegs oor geldigheid as dit die nodige effek tot gevolg kan hê. Let reeds nou al daarop dat die bewysbare bydra tot krag, maar dat die besluite wat op grond daarvan geneem word, bepaalde gevolge het.

3. Geldigheid gedefinieer

Geldigheid word seker as dié belangrikste eienskap van enige toets / toetsresultaat geag, of soos Van der Walt en Steyn (2007:138) dit stel: “validity is the central concept in language assessment.” Geldigheid blyk egter 'n komplekse begrip te wees (Brown, 2004:22), wat met verloop van tyd verskeie interpretasies ondergaan het. Dit word veral gekompliseer deur die invloedrykheid van Samuel Messick (1989; 1994) en sy interpretasie daarvan, naamlik dat alle aspekte van toetse onder die begrip geldigheid ingetrek kan word en dat geldigheid nie 'n attribuut van 'n toets kan wees nie, maar dat dit alles omvat en insluit. Ten einde die kompleksiteit van die begrip geldigheid meer verstaanbaar te maak, word definisies van geldigheid vervolgens sistematies benader en in drie paradigmas opgedeel, te wete die tradisionele paradigma, die hedendaagse paradigma, en 'n alternatief tot die huidige konvensie.

3.1 Die eerste paradigma: 'n tradisionele benadering tot geldigheid

Geldigheid word reeds dekades gelede, binne konteks van die psigometrika, deur Magnussen (1966:123) gedefinieer as “... the accuracy with which meaningful and relevant measurements can be made ... in the sense that it actually measures the traits it was intended to measure.” In dieselfde jaar verklaar French en Michael (1968:165) ook dat verskillende toetse vir verskillende doeleindes ontwikkel is en dat die geldigheid van 'n toets afhanklik is van die mate waartoe 'n spesifieke toets gemeet het wat dit ten doel gehad het om te meet: “[t]ests are used for several types of judgements and for each type of judgement, a different type of investigation is required to establish validity.” Beide definisies bou voort op die werk van Kelley (1927), Cattell (1946) en Cronbach en Meehl (1955), maar die definisies van Magnussen en French en Michael word verkies, omdat dit eerstens 'n goeie opsomming van bostaande werk is, en tweedens reeds 'n suggestie is van wat in die volgende paradigma sou volg. Let intussen daarop dat die klem in hierdie definisies implisiet, dog sterk, val op die empiriese. Daar is egter ook suggesties dat relevante interpretasies van toetspunte gemaak moet kan word ten einde oordele oor toetslinge te kan vel. Lado (1961:30) wys ook daarop dat geldigheid

nie algemeen nie, maar wel spesifiek is. Dit vind steeds weerklank in hedendaagse beskouings (die tweede én derde paradigmas) van geldigheid, naamlik dat “validity is contextual, local and specific” (Van der Walt & Steyn, 2007:138).

Die tradisionele beskouing van geldigheid kom dus daarop neer dat 'n toets nie net konsekwent moet wees in die resultate wat dit oplewer nie (betroubaar dus), maar dat dit veral moet meet wat dit veronderstel is om te meet. 'n Toets is daarom net geldig vir 'n spesifieke doel in 'n spesifieke konteks 'n toets kan opsigself nie oor geldigheid beskik nie, maar dit kan slegs geldigheid vir 'n spesifieke intensie hê. Geldigheid word derhalwe geag as 'n inherente deel van 'n toets, 'n eienskap waaroor 'n toets sal beskik, mits dit meet wat dit veronderstel is om te meet. Sluijter (1998:69) argumenteer egter dat die beantwoording van die vraag of 'n toets meet wat dit veronderstel is om te meet, nooit die einddoel van 'n geldigheidsondersoek kan wees nie. Hy stel dit dat

[e]en positief antwoord op deze vraag is een noodzakelijke, maar geen voldoende, voorwaarde voor een test om een rol te kunnen spelen bij het nemen van een bepaalde beslissing. Indien een valideringsonderzoek zich zou beperken tot de vraag of een test meet wat deze beoogt te meten, is het niet mogelijk een uitspraak te doen over de geschiktheid van deze test voor het nemen van een bepaalde beslissing.

Interessant genoeg word daar in die tradisionele benadering tot geldigheid, soos reeds hier bo gesien kan word, gesinspeel op wat logieserwys uit hierdie paradigma na die volgende, wat 'n meer hedendaagse benadering tot geldigheid is, sou voortvloei. Dit is naamlik dat geldigheid daardie eienskap van 'n toets is wat toetsresultate funksioneel of bruikbaar maak – 'n toets op sigself is nie geldig nie; dit is slegs geldig vir spesifieke doeleindes (Sundberg, 1977:43; Lado, 1961:30). Dit word bevestig in die volgende woorde van Van der Walt en Steyn (2007:138): “[validity pertains] to a specific use of a test, i.e. one asks whether the test is valid for this situation.”

Binne konteks van die teoretiese raamwerk wat vir hierdie artikel gebruik word, is dit ook duidelik dat 'n konstitutiewe begrip soos geldigheid inderdaad een van die grondbegrippe is waarop tegniese ontwerpe berus, maar daar is reeds aanduidings daarvan dat hierdie grondbegrip slegs ontsluit kan word deur die leidende funksies van die ontwerp in hierdie geval spesifiek die juridiese en etiese dimensies wat opvattinge oor regverdigheid en billikheid antisipeer, binne spesifieke kontekste en met spesifieke doeleindes voor oë.

In die volgende gedeelte word daar na die hedendaagse beskouing tot geldigheid gekyk wat as't ware 'n teenreaksie op die oordrewe klem van die bewysbare was. Daar sal duidelik gesien word hoe die verskillende dimensies van toegepaste linguïstiese artefakte in wisselwerking met mekaar behoort te tree, maar dat daar in die tweede paradigma 'n oordrewe verskuiwing na die regulatiewe kant van die teoretiese raamwerk plaasgevind het, wat aanleiding gegee het tot die ontstaan van die derde paradigma wat aan die hand doen dat die konstitutiewe en die regulatiewe momente in tegniese ontwerpe in harmonie met mekaar moet wees.

3.2 Die tweede paradigma: die huidige konvensie

Die tradisionele benadering tot geldigheid is geag as beperkend van aard, omdat dit volgens teoretici nie genuanseerd genoeg nie, en selfs simplisties, na geldigheid gekyk het. So is daar n t drie sogenaamde tipes geldigheid vir die eerste paradigma geidentifiseer: kriterium-gerigte (wat beide konkurrente en voorspellingsgeldigheid insluit), inhouds-, en konstruktgeldigheid – die klem is duidelik op dit wat gemeet en bewys kan word. Geldigheid is egter meer kompleks as die definisies hier bo en daarom word ’n fokus op slegs die drie genoemdes tipes geldigheid nie meer as voldoende geag nie. Die hedendaagse benadering tot geldigheid tref dus nie meer ’n onderskeid tussen hierdie drie nie, maar verenig dit in ’n raamwerk waar konstruktgeldigheid as die kern geag word en die ander twee as aspekte daarvan beskou word. Dit kry reeds in 1955 beslag in die werk van Cronbach en Meehl (1955:283) waar hulle konstruktgeldigheid as volg definieer:

Construct validity is ordinarily studied when the tester has no definite criterion measure of the quality with which he is concerned and must use indirect measures. Here the trait or quality underlying the test is of central importance, rather than either the test behaviour or the scores on the criteria.

Hoewel die eerste paradigma se definisie, naamlik dat ’n toets moet meet wat dit veronderstel is om te meet, en dat dit bewys moet kan word, steeds as die basiese vetrekpunt vir geldigheid geag word (cf. Alderson, Clapham & Wall, 1995:170; Davies, 1990:21; Hughes, 2003:50; Jordan, 1997:88; Weir, 1993:19), en die tweede paradigma in lyn daarmee is, word dit deesdae tog meer genuanseerd en selfs nog versigtiger omskryf as wat Cronbach en Meel hier bo gedoen het, en daar vind ’n klemverskuiwing na die interpretatiewe kant van toetsresultate plaas.

McNamara en Roever (2006:11) wys daarop dat geldigheid in essensie te make het met die maak van afleidings oor die vermoens van toetslinge, vir ’n spesifieke domein, en dat dit gedoen word op grond van toetsresultate wat geïnterpreteer moet word. Dieselfde outeurs voer aan dat hoewel Cronbach aangedui het dat die nodige bewyse vir die geldigheid van ’n toets ingesamel moet word ten einde toetsresultate geregverdig te kan interpreteer, hy nie in sy aanvanklike werk die kwessie van sosiale verantwoordbaarheid in ag geneem het nie en eers veel later op sake soos konsekwensi le geldigheid gefokus het (McNamara & Roever, 2006:10).

Etlke jare na Cronbach se interpretasie van die begrip geldigheid wys Bachman (1990:289) steeds daarop dat geldigheid, naas die feit dat ’n toets moet meet wat dit veronderstel is om te meet, alleen bepaal kan word as dit ’n proses is wat aan deurlopende analise en empiriese ondersoek van toetsresultate onderwerp word, wat nogmaals ’n aanduiding daarvan is dat die interpretasie van toetsresultate ’n belangrike kwessie begin raak. Miller, Linn en Gronlund (2009:70) voer dieselfde argument as Bachman hier bo, maar selfs in meer besonderhede, naamlik dat geldigheid slegs bereik kan word as toetsresultate toepaslik, betekenisvol en bruikbaar geïnterpreteer kan word, en as die gevolge van besluitneming gebaseer op hierdie interpretasies billik en regverdig sal wees.

Bachman en Palmer (1996:9) wys daarop dat daar twee basiese beginsels is waaraan enige taaltoets moet voldoen. Die eerste hiervan dikteer die nodigheid vir 'n noue verband tussen 'n toetsling se prestasie op 'n toets en sy/haar toekomstige taalgebruik in 'n spesifieke situasie (soos in die eerste, psigometriese, paradigma). Dit gaan dus oor "a correspondence between language test performance and language use [in the target language use situation]". Ten einde hierdie 'correspondence' te bewerkstellig, behoort daar 'n agtergrond of 'n konteks te wees waarteen taaltoetsresultate in terme van werklike taalgebruiksituasies geposisioneer en beoordeel kan word. Dit is dus duidelik dat hierdie soort siening, waar konteks en die effek van toetsresultate op toetslinge aandag begin geniet het, tot die tweede paradigma behoort, hoewel daar reeds blyke daarvan in die eerste paradigma was.

Die tweede basiese beginsel waaraan taaltoetse moet voldoen (Bachman en Palmer, 1996:9), het te make met die bruikbaarheid van 'n toets, en bruikbare toetse is by implikasie met 'n spesifieke doel voor oë ontwerp vir 'n spesifieke doelgroep en resultate moet dus interpreteerbaar wees. Hierdie beskouing van geldigheid kom daarop neer dat die konsep altyd te make het met die benutting van toetsresultate vir die een of ander doel (dit artikuleer dus met die tradisionele beskouing), maar dat dit aan die voorwaardes van bewysbaarheid en billikheid moet voldoen, wat weereens in lyn is met die teoretiese vertrekpunt van hierdie artikel, naamlik dat konstitutiewe en regulatiewe momente van tegniese instrumente in wisselwerking met mekaar is, en moet wees.

'n Meer hedendaagse benadering tot geldigheid kom dus daarop neer dat dit, in aansluiting by die eerste paradigma, daaroor gaan of 'n toets meet wat dit ten doel het om te meet, sowel as die adekwaatheid en korrektheid/toepaslikheid waarmee toetsresultate geïnterpreteer word. Geldigheid gaan verder in die kontemporêre siening dikwels oor graadverskille – dit is nie 'n alles-of-niks-situasie nie, en 'n toets, of die interpretasie van die toetsresultaat, kan dus binne die tweede paradigma nie bloot as geldig of ongeldig beskryf word nie. Messick (1989:13) stel dit byvoorbeeld dat

[v]alidity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores ...

In die tweede paradigma is daar, soos reeds genoem, derhalwe wegbeweeg van die redenasie dat daar drie tipes (bewysbare) geldigheid is en dit is vervang met 'n verenigde benadering tot geldigheid waarin konstruktgeldigheid sentraal staan, en kriterium-gerigte geldigheid en inhoudsgeldigheid as deel van konstruktgeldigheid geag word, omdat dit deel is van die verdedigbaarheid ('n konstitutiewe begrip) van die konstruk (Messick, 1989:20; 1996:248). In aansluiting hierby strek konstruktgeldigheid volgens Bachman (1990:290) veel verder as kriterium-gerigte geldigheid of inhoudsgeldigheid, aangesien dit "empirically verifies (or falsifies) hypotheses derived from a theory of factors that affect performance on tests – constructs, or abilities, and characteristics of the test method."

Messick (1980:1015) het dieselfde punt egter al 'n dekade vroeër geopper toe hy verklaar het dat kriterium-gerigte geldigheid en inhoudsgeldigheid nie geskik geag kan word "[to] bear the name validity and to wear the mantle of all that name implies." Sy regverdiging hiervoor in sy seminale werk oor geldigheid (Messick, 1989:17) is dat inhoudsgeldigheid bewyse oor die domein en die verteenwoordigende aard van die toetsinhoud van die spesifieke domein verskaf, terwyl dit behoort te gaan oor 'n regverdiging vir die interpretasie van toetsresultate, wat 'n regulatiewe moment is. Let alreeds hier op die paradoks in sy redenasie, naamlik dat hy die tradisionele beskouings van kriterium-gerigte geldigheid en inhoudsgeldigheid afmaak as onvoldoende en ontoepaslik, maar dat hy nie krities staan teenoor die feit dat dit bewyse lewer wat bydraend is tot die regverdiging van die konstruk nie.

Benewens die kwessie van teoretiese verdedigbaarheid gaan hy in dieselfde hoofstuk selfs nog verder en voeg twee komponente by sy verenigde benadering tot geldigheid, naamlik die sosiale aspek van toetsing, oftewel die impak of effek daarvan op toetslinge, en die bruikbaarheid daarvan (Messick, 1989:13). Hy voer aan dat geldigheid 'n samevoeging behoort te wees van bestaande bewyse vir, en die moontlike gevolge van, die interpretasie van toetsresultate en die bruikbaarheid daarvan, terwyl hy 'n dekade vroeër aangedui het dat 'n fokus op verskillende tipes geldigheid bydra tot "confusion and, in the face of confusion, oversimplification" (Messick, 1980:1014).

Hierdie benadering van Messick (1989:10) tot geldigheid word in Figuur 3 aangehaal en kom opsommenderwys op die volgende neer: geldigheid is eerstens 'n verenigende begrip wat verskillende aspekte byeenbring en konstrukgeldigheid lê ten grondslag van alles. Die tradisioneel-onderskeibare konsepte van geldigheid, wat hier bo bespreek is, vervaag dus en word van gelyke waarde geag – die een is nie van 'n hoër prioriteit as 'n ander nie, behalwe wat konstrukgeldigheid betref. Tweedens is daar blyke van ooreenstemming met die tradisionele benadering, naamlik dat geldigheid gaan oor die argument wat gevoer word (gegrond op die bewyse), maar tog verklaar hy dat die verskillende bewyse behoort saam te vloei in 'n verenigde argument wat aandui of die toets sy doel gedien het – dit gaan dus nie meer oor verskillende tipes geldigheid nie, maar oor verskillende tipes bewyse.

Messick argumenteer ten derde dat geldigheid nie 'n kenmerk van 'n toets kan wees nie, maar slegs van die afleidings wat oor die toetsresultaat gemaak word. Die rede hiervoor is dat die aanvanklike doel van toetse soms vervaag en derhalwe in situasies gebruik, of liever misbruik, word en dat die resultate dan geïnterpreteer word asof dit vir daardie spesifieke doel ontwerp is. Ten slotte dra Messick se argument by tot die ondersoek van die gevolge van toetsresultate op individue, gemeenskappe, ensovoorts. In Messick (1989:10) se eie woorde is konstrukgeldigheid die "integrating force that unifies validity issues into a unitary concept... [that] ...binds the validity of test use to the validity of test interpretation... [and that] ...binds social consequences of testing to the evidential basis of test interpretation and use."

	Test interpretation	Test use
Evidential basis [empirical]	Construct validity (CV)	CV + Relevance / Utility (R/U)
Consequential basis [impact]	CV + Value implications (VI)	CV + R/U + VI + Social consequences

Figuur 3: Messick se benadering tot geldigheid

Borsboom, Mellenbergh en Van Heerden (2004:1061) stel dit egter dat “most of the validity literature either fails to articulate the validity problem clearly or misses the point entirely”, wat weerklank vind in Van der Walt en Steyn (2007:139) se opmerking dat hoewel Messick se werk oor geldigheid seminaal van aard was, en steeds is, dit ook betreklik vaag is, en selfs moeilik interpreteerbaar. Bostaande artikuleer ook met McNamara en Roever (2006:33) se opinie dat toetsontwikkelaars nie noodwendig kundiges op die gebied van toetsvalidering is nie, maar dat daar tog ’n dringendheid is vir diesulkes om hulle daarmee te bemoei. Hulle aanbieding van Messick se raamwerk is daarom moontlik ’n beter vertolking daarvan en dit word in Figuur 4 hier onder aangehaal (McNamara en Roever, 2006:14):

	What test scores are assumed to mean	When tests are actually used
Using evidence in support of claims: test fairness	What reasoning and empirical evidence support the claims we wish to make about candidates based on their test performance?	Are these interpretations meaningful useful and fair in particular contexts?
The overt social context of testing	What social and cultural values and assumptions underlie test constructs and hence the sense we make of scores?	What happens in our education systems and the larger social context when we use tests?

Figuur 4: McNamara en Roever se herinterpretasie van Messick se benadering tot geldigheid

Hierdie herinterpretasie het primêr ten doel gehad om op ’n eenvoudiger manier aan te toon hoe Messick se raamwerk op grond van die nodige bewyse ook kennis neem van die sosiale dimensie van toetsing en in sekere kontekste geregtig is om te gebruik. Dit lê egter ook ’n swak plek in Messick se matriks bloot, naamlik dat die sosiale dimensie van toetsing slegs in die onderste deel van die matriks na vore tree en

die regverdigingsdimensie ('fairness') in die boonste deel, wat 'n vraag oor die werklike geïntegreerdheid van die Messick-raamwerk laat opkom. Hierdie vaaghede van die Messick-raamwerk het inderdaad die potensiaal om verwarring te skep en uiteindelik by te dra tot 'n komplekse netwerk van argumente. Dit word bevestig deur Borsboom et al. (2004:1062) wat verklaar dat

... in the past century, the question of validity has evolved from the question of whether one measures what one intends to measure ... to the question of whether interpretations and actions based on test scores are justified – not only in the light of scientific evidence but with respect to social and ethical consequences of test use... Thus, validity theory has gradually come to treat every important test-related issue as relevant to the validity concept and aims to integrate all these issues under a single header. In doing so, however, the theory fails to serve either the theoretically oriented psychologist or the practically inclined tester: The theoretically oriented are likely to get lost in the intricate subtleties of validity theory, whereas the practically oriented are unlikely to derive a workable conceptual scheme with practical implications from it. A theory of validity that leaves one with the feeling that every single concern about psychological testing is relevant, important, and should be addressed in psychological testing cannot offer a sense of direction to the working researcher. ... Validity is not complex, faceted, or dependent on nomological networks and social consequences of testing. It is a very basic concept and was correctly formulated, for instance, by Kelley (1927, p. 14) when he stated that a test is valid if it measures what it purports to measure.

Hoewel die saak moontlik nie heeltemal so eenvoudig is as wat Borsboom et al. dit aanbied nie, is dit tog interessant dat hierdie kwessie nog nie meermale deur teoretici ondersoek is nie en dat daar bloot net indirekte en geïmpliseerde uitsprake hieroor gemaak word. 'n Voorbeeld is Bachman en Palmer (1996:17-36) wat, sonder om direkte kritiek op Messick uit te spreek, subtiele klemverskuiwings van 'n vereniging van onderskeibare konsepte onder die begrip geldigheid na 'n vereniging van konsepte onder die begrip nuttigheid aanbring. Hierdie konsepte sluit betroubaarheid, konstrugeldigheid, outentisiteit, interaktiwiteit, impak en bruikbaarheid in. Wat hul klemverskuiwing wel prysenswaardig maak, is dat dit die genoemde ses konsepte oënskynlik as gelykwaardig ag en nie een as 'n superordinaat daarstel nie. Die probleem bly egter steeds dat konseptuele skerpheid tot 'n mate verlore gaan in 'n verenigde benadering en dat enige verenigende begrip derhalwe tog as superieur tot die ander geag kan word; in hierdie geval die begrip nut. 'n Moontlike rede vir die gebrek aan deeglike ondersoek is dat eers Messick en later ander noemenswaardige teoretici se invloedrykheid daartoe bygedra het dat hierdie benadering tot geldigheid, waar geldigheid as 'n kenmerk of attribuut van 'n toets verskuif het na die interpretasie van toetsresultate, hedendaags steeds as die norm aanvaar word.

Dit is dus die vereniging van toetskenmerke of –attribute onder 'n enkele sambreelterm, geldigheid, wat kritici genoodsaak het om daarop te reageer en wat aanleiding tot 'n alternatiewe paradigma gegee het, wat hier onder bespreek sal word. Vir nou is dit egter

voldoende om hierdie afdeling af te sluit deur wel erkenning te gee aan Messick dat hy die kwessie van impak ('consequential validity') op die voorgrond geplaas het. Die belang hiervan is dat daar verskeie faktore is wat tot sogenaamde konstruk-irrelevante variansie kan bydra en dus 'n impak op toetsresultate kan hê. Konstruk-irrelevante variansie verwys na daardie aspekte in 'n toets wat veroorsaak dat die interpretasie van die resultaat ongeldig is. Dit gaan dus daaroor dat die toets te wyd meet of aspekte meet wat nie deel van die domein vorm wat gemeet moet word nie.

Die teoretiese raamwerk van die studie waarop hierdie artikel gebaseer is, voer aan dat tegnies-linguistiese artefakte ontwerp behoort te word met 'n antisipatoriese en 'n terugskouende ingesteldheid en daar is dus nie fout te vinde met Messick en ander se soektog na bewyse of teoretiese begroning nadat 'n toets afgelê is nie. Wat hulle egter uit die oog verloor, is dat die beste moontlike interpretasie van 'n toetsresultaat nie die probleem van konstruk-irrelevante variansie oplos of selfs aanspreek nie. Die resultate sal verdag bly, omdat die instrument (nie die interpretasie nie) ongeldige elemente kan bevat. Interpreteerbaarheid kan nie probleme uit 'n objektiewe instrument verwyder nie en dit is juis hier waar teoretici wat tot hierdie paradigma behoort se argumente blyk om op losse skroewe te staan.

3.3 Die derde paradigma: 'n alternatief tot die hedendaagse siening

Soos reeds in die vorige afdeling genoem, het die vereniging van dit wat tradisioneel as attribute van toetse gesien is, byvoorbeeld betroubaarheid, geldigheid, outentisiteit, interaktiwiteit, impak en die praktiese sy daarvan (implementeerbaarheid en bruikbaarheid), tot die derde paradigma wat 'n alternatief tot die huidige ortodoksie bied, aanleiding gegee. Die kern hiervan lê in die woorde van Van der Walt en Steyn (2007:140), naamlik dat

... the psychometric tradition of language testing has obscured the role and effect of language testing in society, especially its sorting and gate-keeping roles, which ultimately depend on the policies and values that underlie any test. The practice of decision-based interpretations has now become part of validity, although as yet there is no coherent theory of the social context in current validity theory.

Binne konteks van die teoretiese raamwerk wat vir hierdie artikel gebruik is, kan dit voorts beredeneer word dat die grondliggende funksies waarop toegepaste linguisties-tegniese ontwerpe berus, ontsluit word deur regulatiewe aspekte en dat die twee kante van die raamwerk ideaal-gesproke in balans met mekaar moet wees, eerder as wat daar te sterk op die een of die ander gefokus word. Die konsep geldigheid sou dus, binne konteks hiervan, eintlik beperk word deur dit slegs vanuit die Messick-perspektief te benader.

Hoewel die Messick- en Bachman en Palmer-benaderings hier bo bespreek oënskynlik ruimte laat vir 'n verskeidenheid van perspektiewe en bewyse om geldigheid te bepaal (wat opsigself alreeds problematies is), gaan geldigheid volgens Van der Walt en Steyn

(2007:140) in die eerste plek altyd oor die waarheid. Die vraag ontstaan egter of die waarheid absoluut of relatief is wanneer hulle aanvoer dat “[t]ruth remains a relative concept, a question of judgement, a matter of degree, subject to new or more relevant evidence. There is no such thing as an absolute answer to the validity question...” (Van der Walt en Steyn, 2007:140). Die kwessie van waarheid word dus vervaag in bogenoemde benaderings deur ’n aantal veranderlikes te verenig onder ’n oorkoepelende begrip, hetsy dit geldigheid of nut genoem word. Lank voor Van der Walt en Steyn se opmerking hier bo word dit reeds in die werk van bekende taaltoetskundiges soos Bachman en Palmer (1996:38), Davies (1990:6) en Alderson et al. (1995:180) geïmpliseer dat geldigheid liewer gesien moet word net vir wat dit is, naamlik ’n soeke na juistheid deur ’n verskeidenheid van toetsattribute te ondersoek. Sluitjer (1998:71-72) verwoord die problematiek hiermee as volg:

Het aantal eisen ten aanzien van validiteit, zoals met name door Messick ... verwoord zijn, is uitgebreid. De relevantie van de verschillende eisen varieert met het specifieke doel van een test. Het werk van Messick biedt echter nauwelijks criteria met behulp waarvan de mate van relevantie van al deze eisen, gegeven een specifiek gebruiksdoel, valt te bepalen ... De afwezigheid van concrete richtlijnen voor het opzetten van een valideringsonderzoek maakt het moeilijk om onderzoeksvragen te prioriteren. Een bijkomend probleem is dat er veel bronnen van informatie zijn waar onderzoekers uit kunnen putten om na te gaan of een bepaalde interpretatie van een score op een test of toets juist is, of om een bepaalde vorm van gebruik te rechtvaardigen. In de meest recente versie van de ‘Standards for Educational and Psychological Testing’ staat dat ‘Resources should be invested in obtaining the combination of evidence that optimally reflects the value of a test for an intended purpose.’ ... De twee voornoemde probleme tezamen zorgen er echter voor dat nie expliciet vast te stellen is of in een valideringsonderzoek sprake is van een optimale of zelfs goede ‘combination of evidence’.

Van der Walt & Steyn (2007:140) voer verder aan dat die bepaling van geldigheid deurlopend behoort te geskied. Empiriese data, versamel oor tyd heen, is inderdaad nodig om uitsprake oor die geldigheid van ’n toets te kan maak. Uiteindelik is elke toets en die omstandighede daaromheen ook uniek en dit moet daarvolgens beoordeel word.

Dit bring die kwessie van validering (’n subjektiewe proses) wat staan naas geldigheid (’n objektiewe, meetbare eienskap) weer ter sprake, wat nie verder in hierdie artikel bespreek sal word nie. Daar word volstaan by die feit dat die probleem met hierdie twee begrippe is dat dit ineenvloei en daarom vervaag. Weideman (2009a:243) verduidelik dit as volg:

... the distinction between the subjective process of validation and the objective validity of a test is an essential one, which seems to be forgotten in some ... discussions ... Viewed subjectively, validity is the achievement of validation. Viewed objectively, it is a function of test scores. If the latter were not the case,

we would not have been able to ascribe or impute an adequate interpretation to such scores – for those scores would have lacked not only validity, but also interpretability.

Met verloop van tyd het dit egter duidelik geword dat in stede daarvan dat Messick en ander soos Bachman en Palmer dit kon bewerkstellig om aan te toon hoe die attribute van 'n toets in wisselwerking tot mekaar staan, dit eerder veroorsaak het dat daar konseptuele verwarring ontstaan, wat denke na die tradisioneel-onderskeibare tipes geldigheid terugneem. Binne die konteks van die tweede paradigma is daar dan weinig ander keuse as dat onderskeibare tipes geldigheid bloot weer bewys moet word aan die hand van sogenaamde hipoteses, sodat die resultate van die toets minstens sosiaal verantwoordbaar sal wees. Davies en Elder (2005:798) beskryf hierdie situasie as volg:

... in spite of the unitary view now taken of validity, it has to be operationalized through the usual suspects of content and construct validity, concurrent and predictive validity ... To the usual suspects we need to add reliability; some will want to include face validity. However, the joker is validity ... called consequential validity.

Weideman (2009c:241) sluit hierby aan en argumenteer dat “[there is a] need to be critical of notions that subsume others. Conceptual clarity is essential – indeed, a necessary condition for test design ...” Ten einde dit te bewerkstellig bied dieselfde outeur (Weideman, 2009a:240) gevolglik nog 'n interpretasie van Messick se matriks oor geldigheid aan wat toon dat Messick se stellings nie meer oor geldigheid per se handel nie, maar eerder oor die interaksie wat verskillende aspekte van toetse met mekaar het en fundamenteel is tot die interpretasie van toetsresultate. Dit word in Figuur 5 en die deel wat daarop volg aangehaal.

	<i>Adequacy of ...</i>	<i>Appropriateness of ...</i>
inferences made from test scores	depends on multiple sources of evidence	relates to impact considerations / consequences of tests
the design decisions derived from the interpretation of empirical evidence	is reflected in the usefulness / utility or (domain) relevance of the test	will enhance and anticipate the social justification and political defensibility of using the test

Figuur 5: Weideman se herinterpretasie van Messick se benadering tot geldigheid met 'n fokus op die verband tussen fundamentele oorwegings in taaltoetsing

The matrix ... can be read as a number of claims about or requirements for language testing, as follows (left to right, top to bottom):

- ... The technical adequacy of inferences made from test scores depends on multiple sources of empirical evidence.
- ... The appropriateness of inferences made on from test scores relates to the detrimental or beneficial impact or consequences that the use of a test will have.
- ... The adequacy of the design decisions derived from the interpretation of empirical evidence about the test is reflected in the usefulness, utility or relevance to actual language use in the domain being tested.
- ... The appropriateness of the design decisions derived from the interpretation of empirical evidence about the test will either undermine or enhance the social justification for using the test, and its public or political defensibility.

Bostaande matriks het ten doel om aan te toon dat daar koherensie en passing is tussen 'n aantal onderskeibare konsepte wat fundamenteel tot enige toets behoort te wees. Hy beweeg met hierdie herinterpretasie van Messick dus weg van die oordrewe klem op konstrugeldigheid en verskuif die aandag na 'n konseptueel meer gefokusde benadering wanneer hy aanvoer dat "... if one does not deliberately distinguish what is conceptually distinct, the distinction so avoided subsequently obtrudes itself upon the conceptual analysis" (Weideman, 2009a:241).

Hierdie meer gefokusde benadering sluit onder andere aspekte van toetse soos betroubaarheid, geldigheid, bruikbaarheid, impak, verantwoordbaarheid, en regverdiging in. Weideman (2009a:240-241) som dit die beste op met die woorde "[s]urely concepts like technical adequacy, appropriateness, the technical meaningfulness (interpretation) of measurements (test scores), utility, relevance, public defensibility and the like must be conceptually distinguishable to make sense..." Waarna hier verwys word, is niks anders as die elementêre begrippe wat die ruggraat vorm van die teoretiese raamwerk van hierdie artikel nie. Hoewel Messick ook (weergawes van) hierdie begrippe gebruik, beskou hy dit alles, jammerlik genoeg, as onderdele van een (geldigheid).

Ander soos Fulcher en Davidson (2007:279), en Borsboom et al. (2004) gaan selfs verder en stel dit eksplisiet, naamlik dat die Messick-perspektief foutief en selfs onsinnig is. Fulcher en Davidson (2007:279) vra daarom tereg

[h]as this validity-as-interpretation mantra perhaps become over-used? If a test is typically used for the same inferential decisions, over and over again, and if there is no evidence that it is being used for the wrong decisions, could we not speak to the validity of that particular test – as a characteristic of it?

4. Samevatting en gevolgtrekking

Uit die bostaande bespreking behoort dit duidelik te wees dat geldigheid vanuit verskillende perspektiewe benader kan word. Dit blyk wel die beste te wees om die benadering van die derde paradigma te volg, omdat dit 'n meer gebalanseerde en gefokusde kyk op geldigheid bied as enige van die ander paradigmas. Dit kom daarop neer dat geldigheid inderdaad 'n attribuut van 'n toets kan wees, in teenstelling met Messick se opinie daaroor. Hoewel die tweede paradigma, naamlik dat geldigheid in wese oor die interpretasie van toetsresultate gaan, wyd aanvaar word (ook deur taalonderrysers en –dosente), moet daar in ag geneem word dat dit afhanklik is van die doel waarvoor die toets ontwerp is.

In aansluiting by die titel van hierdie artikel, moet dit voorts duidelik gestel word dat 'n swak meetinstrument nie reggestel kan word deur die interpretasie van die resultate daarvan nie. Interpretasie en geldigheid kan dus onmoontlik dieselfde wees en moet as onafhanklik geag word. Taalonderrysers en –dosente behoort veral hierop te let en versigtig daarvoor te wees om nie in die interpretasie-slaggat van die hedendaagse siening tot geldigheid te trap nie. Daar is anekdotiese getuigenis van onderrysers en dosente wat punte gewoon opwaarts of afwaarts aanpas wanneer toetsresultate nie dit wat verwag is reflekteer nie. Sonder om te reflekteer oor die adekwaatheid of die toepaslikheid van 'n toets vir 'n spesifieke konteks, word daar dus aanpassings gemaak. Dit dra ook nie tot 'n verhoging in die betroubaarheid van die instrument by nie.

Verder is daar verskeie faktore (intern en ekstern) wat toetsresultate kan beïnvloed wat weer die eerste paradigma se denke sterk op die voorgrond laat tree. Myns insiens behoort, soos wat Borsboom et al. (2004) voorstel, 'n volledige terugbeweeg na die tradisionele benadering tot geldigheid egter nie nodig te wees nie; dit is trouens nie geregverdig nie, omdat sosiale aspekte wel in ag geneem moet word.

Daar blyk dus wel 'n saak uit te maak te wees vir onderskeibare konsepte soos nut, regverdiging, betroubaarheid en geldigheid. Al hierdie konsepte sluit egter ook by mekaar aan in die opsig dat dit die belang van konstrugeldigheid beklemtoon, wat Messick ter ere nagegee moet word. Teen die konteks van die teoretiese raamwerk wat vir hierdie artikel gebruik is, is dit egter nodig om versigtig te wees om konstrugeldigheid as die begin en einde van enige toets te ag, omdat, soos Weideman (2009a:249) dit stel,

... the constitutive concept of technical validity can be enriched by articulation of the theoretical idea or rationale for a test. It can also be further enriched and opened up when one considers the notions of its social results or impact. But such enrichment does not constitute either a basis for privileging the concept, or for subsuming everything under it. It merely points to the unfolding of or opening up of the design to the regulative conditions for language testing.

Dit is dus belangrik dat konsepte wat onderskeibaar is liefs apart van mekaar gehou behoort te word ten einde konseptuele skerpheid na te streef wat sal bydra tot 'n beter begrip van die funderende en leidende vereistes vir die ontwerp, ontwikkeling en hantering van toetse. Slegs wanneer die konstitutiewe en die regulatiewe momente in wisselwerking met mekaar kom en daar gepoog word om 'n balans te bewerkstellig, met bepaalde toegewings na die een of die ander kant, kan tegniese ontwerpe soos taaltoetse verantwoordelik gebruik word.

Bronnelys

- Alderson, C., Clapham, C., Wall, D. 1995. *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Bachman, L.F. & Palmer, A.S. 1996. *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
- Bachman, L.F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Borsboom, D., Mellenbergh, G.J. & Van Heerden, J. 2004. The concept of validity. *Psychological Review* 111(4): 1061-1071.
- Brown, J.D. 2004. Research methods for applied linguistics: Scope, characteristics, and standards. In: Davies, A. & Elder, C. (Reds.) 2004. *The handbook of applied linguistics*. Victoria, Australië: Blackwell. pp. 476-500.
- Cattell, R.B. 1946. *Description and measurement in personality*. New York: World Book Company.
- Cliff, A.F., Ramaboa, K. & Pearce, C. 2007. The assessment of entry-level students' academic literacy: does it matter? *Ensovoort* 11(2): 33-48.
- Cronbach, L.J. & Meehl, P.E. 1955. Construct validity in psychological tests. *Psychological Bulletin* 52: 281-302.
- Davies, A. & Elder, C. (Reds.) 2004. *The handbook of applied linguistics*. Victoria, Australië: Blackwell.
- Davies, A. & Elder, C. 2005. Validity and validation in language testing. In: Hinkel, E. (Red.) 2005. *Handbook of research on second language teaching and learning*. Mahwah, New Jersey: Lawrence Erlbaum Associates. pp. 795-813.
- Davies, A. 1990. *Principles of language testing*. Cambridge: Basil Blackwell.

- French, J.W. & Michael, W.B. 1968. The nature and meaning of validity and reliability. In: Gronlund, N.E. (Red.) 1968. *Readings in measurement and evaluation*. New York: MacMillan. pp. 165-172
- Fulcher, G. & Davidson, F. 2007. *Language testing and assessment: an advanced resource book*. Abingdon, Oxon: Routledge.
- Gronlund, N.E. (Red.) 1968. *Readings in measurement and evaluation*. New York: MacMillan.
- Hinkel, E. (Red.) 2005. *Handbook of research on second language teaching and learning*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Hughes, A. 2003. *Testing for language teachers*. Cambridge: Cambridge University Press.
- Jordan, R.R. 1997. *English for academic purposes. A guide and resource book for teachers*. Cambridge: Cambridge University Press.
- Kelley, T.L. 1927. *Interpretation of educational measurements*. New York: Macmillan.
- Lado, R. 1961. *Language testing*. Londen: Longman.
- Linn, R.L. (Red.) 1989. *Educational measurement*. New York: MacMillan.
- Magnussen, D. 1966. *Test theory*. Ontario: Don Mills.
- McNamara, T. & Roever, C. 2006. *Language testing: The social dimension*. Oxford: Blackwell.
- McNamara, T. 1996. *Measuring second language performance*. Londen: Longman.
- Messick, S. 1980. Test validity and the ethics of assesment. *American Psychologist* 35(11): 1012-1027.
- Messick, S. 1989. Validity. In: Linn, R.L. (Red.) 1989. *Educational measurement*. New York: MacMillan. pp. 3-103.
- Messick, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23(2): 13-23.
- Messick, S. 1996. Validity and washback in language testing. *Language Testing* 13: 242-256.
- Miller, M.D., Linn, R.L. & Gronlund, N.E. 2009. *Measurement and assessment in teaching*. Upper Saddle River, New Jersey: Pearson Education.

- Schuurman, E. 2005. *The technological world picture and an ethics of responsibility: Struggles in the ethics of technology*. Sioux Center, Iowa: Dordt College Press.
- Sluifster, C. 1998. *Toetsen en beslissen: Toetsing bij doorstroombeslissingen in het voortgezet onderwijs*. Ongepubliceerde proefskrif. Arnhem: Cito.
- Sundberg, N.D. 1977. *Assessments of persons*. New Jersey: Prentice Hall.
- Van der Walt, J.L. & Steyn, H.S. (Jnr). 2007. Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort* 11(2): 138-153.
- Van Dyk, T.J. 2010. *Konstitutiewe voorwaardes vir die ontwerp van 'n toets van akademiese geletterdheid*. Ongepubliceerde doktorsale tesis. Bloemfontein: Universiteit van die Vrystaat.
- Weideman, A.J. 2006. Transparency and accountability in applied linguistics. *Southern African Linguistics and Applied Language Studies* 24(1): 71-86.
- Weideman, A.J. 2007a. The redefinition of applied linguistics: Modernist and postmodernist views. *Southern African Linguistics and Applied Language Studies* 25(4): 589-605.
- Weideman, A.J. 2007b. A responsible agenda for applied linguistics: Confessions of a philosopher. *Per Linguam* 23(2): 29-53.
- Weideman, A.J. 2009a. Constitutive and regulative conditions for the assessment of academic literacy. *Southern African Linguistics and Applied Language Studies* 27(3): 235-251.
- Weideman, A.J. 2009b. Uncharted territory: A complex systems approach as an emerging paradigm in applied linguistics. *Per Linguam* 25(1): 61-75.
- Weideman, A.J. 2009c. *Beyond expression: A systematic study of the foundations of linguistics*. Grand Rapids, Michigan: Paideia Press.
- Weir, C.J. 1993. *Understanding and developing language tests*. Hertfordshire: Prentice Hall Europe.
- Yeld, N. 2001. *Assessment, equity and language of learning: Key issues for higher education selection in South Africa*. Ongepubliceerde doktorsale tesis. Kaapstad: University of Cape Town.
-