De Cito-spellingtoets: onze bezwaren nader toegelicht

Een reactie op 'Kritiek op toetsen Spelling steunt op losse gronden'

Anna M.T. Bosman, José L.M. Schraven, & Truus van Eekhout

SAMENVATTING

In deze reactie op De Wijs (2010) in het vorige nummer van dit tijdschrift bespreken wij nogmaals onze bedenkingen ten aanzien van de keuze van het Cito om een meerkeuzetoetsvorm op te nemen in het nieuwe Leerling Onderwijs Volg Systeem (LOVS) als maat voor spellingvaardigheid. Op basis van de samenvatting van het oorspronkelijke empirische onderzoek laten we zien dat er wel degelijk grote problemen zijn met de validiteit van de Cito-spellingtoets. Ook de stelling van het Cito dat er passieve spellingkennis gemeten moet worden, middels de meerkeuzetoetsvorm, wordt kritisch beschouwd. Opnieuw wordt er gewezen op een belangrijk didactisch probleem, namelijk systematische blootstelling aan foutgespelde woorden. We maken ook duidelijk dat er problemen zijn met de betrouwbaarheid en dat de diagnostische mogelijkheden van de meerkeuzespellingtoets zeer beperkt zijn. De conclusie is dan ook dat al onze oorspronkelijke bezwaren niet zijn weggenomen. We blijven hopen dat het Cito terugkomt op haar besluit om spelling te toetsen met behulp van meerkeuzeopgaven.

Aanleiding

In februari 2010 bespraken wij (Schraven, Bosman, & van Eekhout, 2010) in het *Tijdschrift voor Orthopedagogiek* (vereniging O & A) de resultaten van een onderzoek naar de Cito-spellingtoets groep 4. De aanleiding waren de vele vragen uit het werkveld over de nieuwe toetswijze. Het doel was om te onderzoeken of de keuze van het Cito om een deel van de spellingtoets te vervangen door een meerkeuzetoets verantwoord was. Daartoe kregen 18 leerlingen precies zoals in de handleiding staat de toets aangeboden. Het eerste deel bestond uit een dictee van 25 woorden (Startwoorden), het tweede deel (Vervolg 2) bestond uit 25 meerkeuzeopgaven. Hier moesten de leerlingen bepalen welk van de vier vetgedrukte woorden in vier verschillende zinnen fout gespeld was. Om te bepalen of deze meerkeuzetoets een adequate vervanging is van een dictee, lieten we dezelfde leerlingen vervolgens alle vetgedrukte woorden opschrijven middels een dictee. De algemene conclusie uit ons onderzoek luidde dat een meerkeuzetoets iets anders meet dan een dictee. We constateerden dit op basis van de volgende feiten:

— De samenhang tussen de prestaties op het dictee en die op de meerkeuzetoets bleek

zwaar onvoldoende; de correlatie tussen de scores op de twee toetsen bedroeg .45.
- Gemiddeld genomen presteerde deze groep leerlingen significant beter op het dictee (85% correct) dan op de meerkeuzetoets (79%).[1]
- Slechts drie leerlingen hadden identieke scores op de meerkeuzetoets en het dictee. Dus bij 15 leerlingen was er een verschil en bij drie leerlingen bleek dit verschil onacceptabel groot te zijn. Zo had leerling 2 op de meerkeuzetoets 60% correct en op het dictee 92%; leerling 12 had 68% correct op de meerkeuzetoets en 88% op het dictee, terwijl leerling 14 slechts 40% correct had op de meerkeuzetoets, maar op het dictee 76% van de woorden correct spelde.
- Er bleek een discrepantie van 25.8% tussen de spelling van het dictee en die van de meerkeuzetoets; in 16.0% van de gevallen was de spelling goed op het dictee, en fout op de meerkeuzetoets en in 9.8% fout op het dictee en goed op de meerkeuzetoets
- Een analyse van de antwoorden van vier meerkeuzeopgaven en een vergelijking daarvan met de prestaties op het dictee doet vermoeden dat de leerlingen in verwarring zijn geraakt door de context waarin de opgaven stonden. Het lijkt erop dat niet alleen kennis van de spelling de score op de meerkeuzetoets heeft beïnvloed, maar ook woordkennis, leesvaardigheid en taakopvatting.

Deze resultaten kunnen eenduidig geïnterpreteerd worden. Naast het feit dat dezelfde woorden in het dictee en in de meerkeuzetoets werden aangeboden, werden ze immers ook bij dezelfde steekproef van 'proefpersonen' afgenomen. De ontstane verschillen kunnen dus niet veroorzaakt zijn door testmateriaal, leerkrachtverschillen, verschil in methoden, streek, achtergrond of eigenschappen van leerlingen (zoals intelligentie) enzovoort. Op basis van onze resultaten concludeerden wij dat de validiteit van de Cito-spellingtoets in het geding is en dat daarmee de meerkeuzetoets geen adequate vervanging is van het woorddictee.[2]

Reactie van het Cito

In het vorige nummer van *Orthopedagogiek: Onderzoek en Praktijk*[3] (2010) reageert De Wijs op ons artikel. Wij zijn verheugd dat het Cito het gesprek met ons wil aangaan. Het onderwijs heeft er recht op te horen welke argumenten er zijn ten aanzien van keuzes voor toetsen en toetsingswijzen. In tegenstelling tot de handleiding behorende bij Spelling Groep 4 (De Wijs, Krom, & van Berkel, 2006) verantwoordt De Wijs nu, vier jaar na de invoering van de toets, wel de keuze van het Cito voor de verandering. In dezelfde reactie gaat zij in op ons onderzoek en worden de resultaten van een studie die het Cito alsnog heeft uitgevoerd, twee jaar na invoering van de toets, besproken.

Doel van dit artikel

In deze bijdrage willen wij nader toelichten waarom de Cito-spellingtoets geen adequaat toetsinstrument is. We zullen eerst, in de paragraaf over de validiteit, opnieuw uitleggen waarom de meerkeuzetoets van de Cito-spellingtoets iets anders meet dan een dictee. Daarna gaan we in op het feit dat het Cito besloten heeft om passieve spellingkennis te gaan meten. Vervolgens bespreken we het probleem van de geringe betrouwbaarheid van de Cito-spellingtoets. We vervolgen met de problemen die zijn ontstaan met de beperkte diagnostische mogelijkheden van de meerkeuzespellingtoets, een thema dat het Cito systematisch onderbelicht. We sluiten ten slotte af met de hoop dat het Cito terugkomt op haar besluit om spelling te toetsen met behulp van meerkeuzeopgaven.

Validiteit van de Cito-spellingtoets is onvoldoende

In de klassieke testtheorie staan twee concepten centraal, betrouwbaarheid en validiteit. De betrouwbaarheid van een test zegt iets

over de mate waarin de test een volgende keer dezelfde score oplevert. Als een meetlat keer op keer aangeeft dat een tafel 115 cm breed is, dan kunnen we stellen dat deze meetlat de lengte van de tafel betrouwbaar meet. Of de lengte van de tafel werkelijk 115 cm is, kan alleen bepaald worden als we een norm hebben, een standaardlengte, waaraan we dat kunnen afmeten. In dit geval is dat een afgesproken eenheid van lengte (een prototype). Als de lengte van onze meetlat overeenstemt met de lengte van het prototype noemen we de meetlat valide.

Validiteit geeft dus aan of de test meet wat deze beoogt te meten. Als je wilt weten hoe warm het is, neem je een thermometer en geen hygrometer. Als je daarentegen wilt weten wat de luchtvochtigheid is, dan neem je een hygrometer. In de natuurkunde is daar nauwelijks discussie over. Het bepalen van de validiteit van een psychologische of didactische test is een stuk lastiger. Dat komt omdat psychologische begrippen of eigenschappen veel minder concreet zijn (denk aan persoonlijkheidseigenschappen of intelligentie) dan zaken als lengte of massa. Spellingvaardigheid lijkt een vrij concreet begrip. Desondanks is uit onze discussie met het Cito gebleken dat er belangrijke verschillen bestaan in de definiering ervan. Dat is zorgwekkend. Eenduidigheid over de definitie van spellingvaardigheid en vervolgens over de wijze waarop deze dan het beste onderwezen en vervolgens gemeten kan worden is immers van groot belang voor de leerkracht (remedial teacher, dyslexiebehandelaar). In het onderwijs moet men erop kunnen vertrouwen dat de uitslag van de toets een beeld geeft van de spellingvaardigheid van elk kind in de klas.

Toen wij ons onderzoek uitvoerden (januari 2008), hebben wij ons gebaseerd op de volgende passage uit de handleiding van de Cito-spellingtoets groep 4 (p. 9, De Wijs e.a., 2006; cursivering door ons aangebracht):

> "Bij spellen gaat het erom woorden om te zetten in schriftbeelden. Daarbij maken we onderscheid tussen klankzuivere en niet-klankzuivere woorden, De eerste fase van het spellingonderwijs richt zich op het correct leren *schrijven* van klankzuivere woorden: je *schrijft* op wat je hoort. Al snel daarna komen de niet-klankzuivere woorden aan de orde, de woorden waarbij er geen eenduidige relatie is tussen klank en letter, zoals bij 'bomen', 'trein' of 'begin'. Om die goed te kunnen *schrijven* moeten de leerlingen regels kunnen toepassen of een woord naar analogie van een ander woord kunnen *schrijven*."

Deze definitie en haar toelichting lezen wij als, we herhalen het hier nog een keer: dat spellen een vaardigheid is waarbij men uit het hoofd de orthografische vorm van een woord opschrijft (dan wel intypt); het woord *schrijven* komt vier keer voor in de geciteerde passage. Er kan dus nauwelijks misverstand zijn over deze definitie. Bovendien kunnen wij van harte instemmen met deze definitie. Daar zit de pijn dus niet.

Het probleem ontstaat pas bij de wijze waarop de veronderstelde vaardigheid geoperationaliseerd wordt. Als je wilt weten of een leerling in staat is om een woord uit het hoofd op te schrijven dan is de meest zuivere test een dictee. Er is dus een prima toetsmiddel voorhanden om de gedefinieerde vaardigheid te meten. Toch koos het Cito ervoor om naast een dictee een tweede toetsvorm op te nemen in haar Leerling Onderwijs Volg Systeem (LOVS), namelijk een meerkeuzeherkenningstoets. Nergens in de handleiding wordt deze verandering inhoudelijk verantwoord. De enige reden die gegeven wordt is dat de meerkeuzevorm afgestemd is op de Entreetoets en Eindtoets (zie De Wijs e.a, 2006, p. 7).

Vier jaar nadat de spellingtoets op de markt kwam, stelt De Wijs (2010) in haar reactie dat er ook helemaal niet gezegd is dat een dictee hetzelfde meet als een meerkeuzetoets. Het dictee zou de actieve spellingkennis meten en de meerkeuzetoets de passieve spelling. Met dit onderscheid kunnen we het theoretisch van harte eens zijn. Of het praktisch handig dan wel relevant is, is een vraag die we verder-

op negatief zullen beantwoorden. Helaas lost deze verantwoording achteraf het probleem niet op. Immers de test is zo geconstrueerd dat de leerkracht uitsluitend een score voor de spellingvaardigheid kan bepalen door een optelling van de score op het dictee en op dat van de meerkeuzetoets. Dit is, om in termen van Koning (2008) te spreken, het optellen van appels en peren (p. 14). Om weer even de vergelijking met de thermometer en de hygrometer als onderdelen van de weersgesteldheid te gebruiken. Om de weergesteldheid te bepalen, zouden we 7 graden Celsius en 18 grammen luchtvochtigheid optellen en dan concluderen dat de weersgesteldheid 25 is. De optelling zegt niets en er gaat op deze manier relevante informatie verloren.

Vier jaar nadat de spellingtoets op de markt gebracht werd rapporteert De Wijs (zie p. 373, 2010) over een onderzoek naar de samenhang tussen de meerkeuzetoets en het dictee. Ter herinnering, in ons onderzoek bij 18 leerlingen vonden wij een correlatie van .45 tussen de dicteeopgaven en de meerkeuzeopgaven. Het Cito vond bij een grote groep van 1318 leerlingen een correlatie van .80. Met de wet der grote aantallen aan hun zijde krijgen ze uiteraard een verhoging van de correlatiecoëfficient als indicatie voor de validiteit, met een onderzoeksopzet die op belangrijke punten minder zuiver was dan die van ons. Dit is overigens nog altijd geen erg hoge waarde voor een spellingtoets. Immers de hoeveelheid gedeelde variantie tussen meerkeuze- en dicteeopgaven is daarmee slechts 64%.Dat wil dus zeggen dat 36% van de onderlinge dicteeverschillen niet verklaard kan worden uit de meerkeuzeverschillen of omgekeerd.[4]

We komen nog even terug op de kritiek van De Wijs dat een aantal van 18 leerlingen die ook nog eens tot de betere spellers behoren te weinig zou zijn voor een goed onderzoek naar validiteit (zie ook B. Schraven, 2010). Ze stelt (in navolging van een opmerking van de COTAN) dat de validiteit van een test afgenomen bij een homogene groep (van goede spellers, overigens afkomstig van vooral lagere sociaaleconomische milieus) negatief beïnvloed kan worden. De Wijs wil toch niet verdedigen dat het toelaatbaar is dat een test bij een homogene groep minder valide is? Er zijn immers vele scholen die een relatief homogene populatie hebben, denk aan zwarte en witte scholen of scholen in relatief dure wijken en scholen in achterstandswijken.

Overigens is het vreemd dat de correlatie tussen dictee en meerkeuzetoets juist in onze groep van betere spellers lager uitviel dan bij het grotere onderzoek van het Cito. Immers de meerkeuzetoets (Vervolg 2) is nou juist bedoeld voor de betere spellers, de zwakkere spellers krijgen volgens de standaardprocedure een dictee als vervolg op de startwoorden. In het Cito-onderzoek deed een grote heterogene groep leerlingen mee aan de meerkeuzetoets met andere woorden dan de dicteewoorden, en ook deden de minder goede spellers uit groep 4 mee. Hoe verklaart De Wijs dit resultaat?

Zonder technisch te worden willen we nog kort iets zeggen over de uitspraak dat een analyse van de dictee- en meerkeuzeopgaven middels het One Parameter Logistic Model (OPLM) zou hebben laten zien dat de onderliggende vaardigheid dezelfde is (zie p. 370, De Wijs, 2010). Tegelijkertijd wordt opgemerkt dat de dictee- en meerkeuzeopgaven niet identiek zijn. Dit is statistisch mogelijk verantwoord, maar we vergeten dat een meetmodel ons geen enkel inzicht geeft in de aard van de te meten variabele, factor, eigenschap, of vaardigheid. In een meetmodel wordt gerekend met getallen. Waar deze getallen voor staan kan het model (i.c., het computerprogramma) nog altijd niet zelf bepalen. Er is een mens voor nodig, de onderzoeker, om er betekenis aan te geven. Dus dan blijft de vraag staan: welke onderliggende vaardigheid wordt er nu feitelijk gemeten?

Samengevat blijft onze conclusie ten aanzien van de validiteit van de Cito-spellingtoets gehandhaafd: deze is ernstig in het geding. De belangrijkste reden is dat de meerkeuzetoets naast spellingkennis ook factoren als woordkennis, leesniveau en taakopvatting

meet. Het meest cruciale uitgangspunt van de klassieke testtheorie wordt hier geschonden namelijk dat een toets zo zuiver mogelijk de eigenschap moet meten die ze beoogt te meten. Terug naar onze thermometer. Een thermometer die niet alleen beïnvloed wordt door de temperatuur, maar ook door de luchtvochtigheid is onbruikbaar. We kunnen dan wel een waarde aflezen, maar wat die dan aangeeft is de vraag.

Spellingvaardigheid en foute spellingen

Zoals uit het voorgaande blijkt kan spellingvaardigheid op verschillende manieren worden gemeten. Naast een dictee en een meerkeuzetoets zou ook een opstel gebruikt kunnen worden voor het vaststellen van de spellingvaardigheid. Om een verantwoorde keuze te maken voor een spellingtoets moeten we het erover eens zijn wat we nu precies willen vaststellen. En omdat het Cito een belangrijke toetsinstantie is – veel scholen gebruiken immers het LOVS – draagt het Cito een grote verantwoordelijkheid in deze. In haar reactie stelt De Wijs (2010, p. 371) dat het belangrijk is dat leerlingen naast actieve spellingkennis, dus zelf woorden correct kunnen spellen, ook passieve spellingkennis opdoen. Passieve spellingkennis is het herkennen van een foute (dan wel correcte) spelling.

Het Cito is zich bewust van haar rol ten aanzien van de onderwijspraktijk. In de handleiding staat nadrukkelijk dat het Cito de mogelijkheid biedt om met behulp van toetsen het onderwijs (op individueel, groeps- en schoolniveau) te evalueren (p. 7). Dit blijkt ook uit het feit dat ze wil aansluiten bij de spellingcategorieën die gehanteerd worden in de meest gebruikte methodes en de omschrijving van wat men onder spelling verstaat. Wat de ontwikkeling van de nieuwe spellingtoets betreft lijkt er iets niet helemaal goed te zijn gegaan. Immers, de opname van een meerkeuzetoets als meting van de passieve spellingkennis in het LOVS sluit helemaal niet aan bij de methoden. Het ontwikkelen van passieve spellingkennis komt immers helemaal niet systematisch aan bod in het onderwijs. En daar zijn heel goede redenen voor.

Ook dit herhalen we nog maar eens een keer. Uit internationaal onderzoek is gebleken dat blootstelling aan foute spellingen van anderen de spellingvaardigheid negatief kan beïnvloeden (Brown, 1988; Dixon & Kaminska, 1997; Jacoby & Hollingshead, 1990). Het veroorzaakt foute woordbeelden met alle gevolgen van dien. Leerlingen kunnen dan niet meer vergelijken of het geschreven woord klopt met een schriftbeeld. Iedere leerkracht weet hoe moeilijk het is om leerlingen die letters omkeren dit af te leren.

Uit de reactie van De Wijs (2010, p. 371) leiden wij overigens af dat ze het met ons eens is wat die blootstelling aan foute spelling betreft, er staat namelijk:

> "De toetsen Spelling worden twee keer per jaar afgenomen. In totaal krijgt een gemiddeld vaardige leerling via de toetsen dus maximaal twee keer per jaar gedurende een half uur een aantal zinnen onder ogen waarin een fout gespeld woord staat... Het gaat hier om een geïsoleerde toetservaring, waarbij het de leerling duidelijk is dat de aangeboden woorden niet geleerd hoeven te worden. De kans dat woorden tijdens de toetsafname foutief worden ingeprent door leerlingen die de juiste schrijfwijze van het woord blijkbaar eerder niet hadden kunnen onthouden lijkt ons dan ook klein. (Ging het leren maar zo makkelijk!)"

Maar helaas, dat is niet de werkelijkheid. Leerkrachten hebben nu al gemerkt dat hun leerlingen slechter presteren op de meerkeuzetoetsvorm. Om de leerlingen goed voor te bereiden wordt er geoefend met het maken van opgaven waarbij ze foute van goede spellingen moeten onderscheiden. Er bestaan al door uitgeverijen gemaakte oefenbladen met

fout en goed gespelde woorden (e.g., kopiën en kopieën). Een ander probleem dat zich aandient met de komst van de meerkeuzetoets is dat leerlingen die niet alleen actieve spellingkennis moeten verwerven middels een dictee, maar ook passief de spelling moeten oefenen, minder dan nu al het geval is en wenselijk wordt geacht (dagelijks) een dictee maken (Bosman, 2007, J.L.M. Schraven, 2004).

Correct leren schrijven is de eerste doelstelling binnen ons huidige spellingonderwijs. Bij spellingonderwijs gaat het toch vooral om een houding aan te kweken die leerlingen ertoe aanzet om onmiddellijk foutloos te spellen. Dat is al moeilijk genoeg op de basisschool. Pas als je iets goed beheerst of er voldoende mee vertrouwd bent, is het mogelijk om er op effectieve wijze op te reflecteren. Hiermee suggereren wij allerminst dat leerkrachten hun leerlingen niet zouden kunnen vragen om hun opstel na te kijken op fouten. Dat dit tot een belangrijke reductie van het aantal fouten kan leiden in zowel het regulier als in het speciaal basisonderwijs is wel gebleken uit eerder onderzoek van Willemen en collega's (de Haan, 2010; Willemen, Bosman, & Van Hell, 2000, 2002). Het meest interessante resultaat uit dit onderzoek was dat de leerlingen zich tijdens het schrijven van hun opstel zo bewust waren van het feit dat ze straks hun opstel op spelfouten moesten nakijken, dat ze reeds tijdens het schrijven na gingen denken over de juiste spelwijze. De substantiële daling van het aantal spelfouten was dus het resultaat van de bewustwording van de leerling dat ze tijdens het schrijven goed op de spelling diende te letten en niet van het achteraf nakijken.

Betrouwbaarheid van de Cito-spellingtoets is onvoldoende

Betrouwbaarheid zegt iets over de mate waarin de test een volgende keer dezelfde score oplevert. Als een meetlat keer op keer aangeeft dat een tafel 115 cm breed is, dan kunnen we stellen dat deze meetlat de lengte van de tafel betrouwbaar meet. Voor het meten van de betrouwbaarheid van psychologische en didactische toetsen kan de meting of de toets vaak niet, zoals bij natuurkundige fenomenen, keer op keer worden afgenomen, omdat degenen bij wie de toets wordt afgenomen ervan leert, waardoor de toets een steeds betere score oplevert. Om dit probleem te omzeilen heeft men een alternatief bedacht. Door een toets uit meerdere opgaven te laten bestaan, kan elke opgave gezien worden als een herhaalde meting van de vorige. De mate van overeenstemming (consistentie) tussen de antwoorden op de toets wordt dan gezien als een van maat van de betrouwbaarheid. Dit wordt de interne consistentie genoemd, ook wel aangeduid met Cronbachs alfa.

Als de interne consistentie 1.0 bedraagt, dan is de test perfect betrouwbaar. Dit komt in de psychologie echter niet voor. Lees- en spellingtoetsen bereiken wel vaak een waarde van .90. We spreken dan van een zeer hoge mate van betrouwbaarheid. Persoonlijkheidstesten komen vaak niet boven de .50 of .60 uit. Een vuistregel van het NIP/COTAN (Evers, Lucassen, Meijer, & Sijtsma, 2010) is dat een test bedoeld voor individueel gebruik een minimale waarde van .80 moet hebben.

In ons onderzoek bleek de meerkeuzetoets (Vervolg 2), die 25 opgaven kent, een alfawaarde van .75 te hebben. Onvoldoende dus als we het NIP/COTAN-criterium serieus nemen. De Wijs (2010) veronderstelt dat een aantal van 25 opgaven te weinig is om de betrouwbaarheid betrouwbaar vast te stellen. Wij kennen geen vuistregel voor een minimum aantal items ter bepaling van de betrouwbaarheid. Wat we wel weten is dat naarmate er meer items nodig zijn om een test betrouwbaar te maken, deze minder eenduidig een onderliggend concept (trait, latente variabele enz.) meet. Anders gezegd, een test die veel items nodig heeft om betrouwbaar te zijn, toont juist de zwakte van het meetinstrument aan.[5]

Het Cito heeft nadat de test op de markt was gebracht alsnog onderzoek gedaan naar de betrouwbaarheid. De test werd uitgevoerd op de 25 items M4 Startdictee en de 25 items Vervolg 1 (ook een dictee) en op de 25 items van M4 Startdictee in combinatie met de 25 items Vervolg 2 (de meerkeuzetoets) De waarden van de betrouwbaarheden bedroegen respectievelijk .90 en .91. Een prima uitslag gegeven het criterium van het NIP/COTAN. Om na te gaan wat de betrouwbaarheid was van de zogenoemde totaaltest binnen onze steekproef hebben we deze in navolging van het Cito-onderzoek ook bepaald. Precies zoals het Cito beweert, als je maar genoeg items neemt (zelfs al zijn ze niet verkregen met eenzelfde type meting en dezelfde woorden) dan stijgt deze naar een statistisch acceptabel niveau van .84. De betrouwbaarheid heeft echter niet de waarde van .91, die het Cito vond bij een steekproef van 782 leerlingen.

Hier zit nu precies het probleem. Met de wet van de grote aantallen, zowel wat het aantal items betreft als het aantal proefpersonen, is dit helemaal geen verrassing. Een leerkracht, remedial teacher, dyslexiebehandelaar heeft echter met individuele leerlingen van doen waarvan hij in de toets een correcte spellingvaardigheid wil vaststellen. Hier sluit het volgende probleem aan dat de COTAN met ons onderzoek meent te bespeuren.

In de beoordeling van de COTAN wordt onder het kopje *Begripsvaliditeit* ook een opmerking gemaakt over de betrouwbaarheid (We raden de COTAN aan dit gescheiden te houden, het kan tot verwarring leiden). De COTAN stelt dat bij het onderzoek van Schraven et al. (2010) een aantal kanttekeningen te plaatsen zijn: 'Ten eerste is het gebaseerd op een steekproef van slechts 18 leerlingen. Ten tweede lijkt het om een homogene groep leerlingen te gaan ("alle leerlingen zijn betere spellers"), en dan is het niet zo gek dat de betrouwbaarheden lager uitvallen'. Het verbaast ons dat de COTAN het accepteert dat een test mogelijk minder betrouwbaar is bij een homogene groep (zie ook B. Schraven, 2010). Een betrouwbare meetlat meet de lengte van de verzameling koffietafels net zo nauwkeurig als die van keukentafels, bijzettafels, picknicktafels, tafels van arme mensen, tafels van rijke mensen en de tafels in uw huis. De betrouwbaarheid van een grote steekproef van allerlei verschillende tafels mag dan wel voldoende zijn, voor het individuele geval is dat blijkbaar niet het geval. Een individu heeft niets aan de mededeling dat het bij een grote groep wel geldt.

Ten slotte gaan we nog even kort in op de vraag van De Wijs waarom wij geen aandacht hebben besteed aan de lage betrouwbaarheden van het Startdictee (.66), de woorden uit de A-zinnen (.65), B-zinnen (.40), C-zinnen (.27) en D-zinnen (.39). Overigens bleek opnieuw dat de wet van de grote getallen (toets over 100 items van de woorden uit de A t/m D-zinnen) ons helpt om de betrouwbaarheid te doen stijgen tot een bijna aanvaardbaar niveau van .79. In ons oorspronkelijke artikel hadden we behalve een gebrek aan ruimte ook vooral het doel om de problemen met de meerkeuzetoets naar voren te brengen. Nu De Wijs er ons echter naar vraagt, willen we er kort op ingaan. Wij schrokken enorm van die lage betrouwbaarheden van de onderscheiden dictees en wilden laten zien dat ook met de gebruikte items in de meerkeuzetoets iets fundamenteels mis is. Nadere inspectie van de items liet zien dat van minimaal 16 van de 100 dikgedrukte woorden niet verwacht wordt dat deze tot de spellingkennis van de leerlingen behoren. Het argument dat de betere speller ook uitgedaagd moet worden, is onverantwoord ten aanzien van de zwakkere spellers, die er door in verwarring kunnen raken. Als we de betere spellers willen uitdagen kan dit zonder problemen met een spellingtoets van een hoger niveau. Inhoudelijk hebben we de items niet verder geanalyseerd, daar zouden ook nog wel eens redenen kunnen worden gevonden voor de lage betrouwbaarheden van de afzonderlijke dictees.

Samengevat, de Cito-spellingtoets is onbetrouwbaar gebleken in onze steekproef. Een school moet erop kunnen vertrouwen dat de spellingtoets voor elke leerling betrouwbaar

meet. Scholen met vooral zwarte leerlingen en scholen met uitsluitend witte leerlingen, scholen met voornamelijk leerlingen uit hogere en scholen met vooral leerlingen uit lagere sociaaleconomische milieus, scholen voor hoogbegaafden en scholen voor minder begaafde leerlingen enzovoort. Leerkrachten kopen niets voor het excuus dat de betrouwbaarheid over een grote groep leerlingen wel in orde is. Ze hebben niet te maken met groepen, ze hebben individuele leerlingen in de klas (cf. p. 14, Koning, 2008) bij wie ze de spellingvaardigheid afzonderlijk willen vaststellen.

Diagnostische waarde van de Cito-spellingtoets is onvoldoende

De spellingtoets van het Cito maakt deel uit van LOVS. De toetsen zijn er in eerste instantie voor de leerkracht om na te gaan hoe het gesteld is met de ontwikkeling van de vaardigheid die deze probeert te onderwijzen. Het doel is niet alleen om vast te stellen op welk niveau elke leerling zit. Het feit dat een leerling een C-niveau heeft wat spelling betreft, geeft de leerkracht geen informatie over hoe te handelen. De ene C-leerling is gebaat bij instructie over de open en gesloten lettergreep, terwijl de andere leerling moet oefenen met de eind-d. Anders gezegd, een test moet de leerkracht informatie geven over wat de leerling wel en wat deze niet beheerst (zie ook handleiding pag. 40).

In haar reactie stelt De Wijs dat de context waarin een fout of goed gespeld woord (i.c., zomer) voorkomt wel degelijk invloed heeft op de kans dat deze gedetecteerd wordt. Het feit dat de context mede de keuze bepaalt, wordt dus door het Cito erkend. Desalniettemin stelt het Cito dat dit minder erg is dan het lijkt. Immers de spellingvaardigheid wordt door meer dan een opgave bepaald en er is maar één opgave met het foutgespelde woord 'somer', en dat woord is voor elk kind gelijk. Het probleem is nu juist dat dit niet voor elke leerling hetzelfde is. De ene leerling heeft problemen met het s-z onderscheid, terwijl de ander de open-lettergreepregel niet kent. In beide gevallen is het woord zomer moeilijk te spellen, maar in het geval de foute spelling somer gepresenteerd wordt, zal leerling A een probleem hebben en leerling B niet. In het geval dat de foute spelling zoomer gepresenteerd wordt heeft leerling B een probleem en leerling A niet. Anders gezegd, hoe weet de leerkracht wie welk probleem heeft?

In haar reactie stelt De Wijs dat meerkeuzeopgaven toch diagnostische mogelijkheden hebben, via de omgekeerde bewijsvoering. Dat komt erop neer dat wanneer je een fout over het hoofd hebt gezien, afgezien van het feit dat je een goed gespeld woord als fout gespeld hebt aangemerkt, je de spellingcategorie van het feitelijk fout gespelde woord niet beheerst. Anders had de leerling die fout wel ontdekt. Dit is een prachtige redenering en we zouden het ermee eens kunnen zijn, ware het niet dat ons onderzoek laat zien dat deze zogenaamde logische analyse in de praktijk niet opgaat. We illustreren dit aan een van de items.

In meerkeuzeopgave 1 moeten leerlingen kiezen uit VLAG, WRAK, GRAP en SWAK. In de meerkeuzeopgave geeft 67% (12 van de 18 leerlingen) aan dat WRAK het fout gespelde woord is. Op basis van de redenering van De Wijs moet de leerkracht er dus vanuit gaan dat 12 leerlingen uit een klas van 18 een probleem hebben met de categorie s/z, want SWAK wordt niet herkend als het fout gespelde woord. Uit het dictee blijkt echter dat 11% (slechts 2 van de 18 leerlingen) ZWAK met een S spelt, de rest doet dit gewoon goed. Kortom, de omgekeerde bewijslast klopt niet en zou de leerkracht een verkeerd beeld hebben gegeven, op basis waarvan deze verkeerde hulp zou hebben gegeven.

Dat het Cito ook wel weet dat de diagnostiek en remediatie vanuit het meerkeuzeformat niet werkt, blijkt uit het feit dat in de handleiding van de Cito-spellingtoets staat dat een nadere analyse met een (controle)*dictee* uit het zogenoemde 'hulpboek' de meeste informatie oplevert.

Conclusie

De uitkomst van onze nadere analyse van het probleem met de Cito-spellingtoets is identiek aan die van ons oorspronkelijke onderzoek (Schraven e.a., 2010). Er zijn fundamentele problemen met de nieuwe Cito-spellingtoets. De validiteit, de betrouwbaarheid en de diagnostische waarde van de toets zijn onvoldoende voor de dagelijkse onderwijspraktijk. In deze bijdrage hebben we onze argumenten nogmaals op een rij gezet en nader toegelicht. Deze resultaten bevestigen de bedenkingen die veel leerkrachten en intern begeleiders naar voren brachten.

We moeten helaas dan ook concluderen dat de slotconclusie van De Wijs *niet* juist is: 'Dergelijke opgaven (bedoeld worden meerkeuzeopgaven) zijn namelijk niet alleen betrouwbaar, maar ook efficiënt, objectief en praktisch' (2010, p. 377). Ons onderzoek laat zien dat meerkeuzeopgaven noch betrouwbaar, noch objectief (ze zijn immers niet valide), noch praktisch (voor de diagnostiek hebben we er niets aan) zijn. Ook de efficiëntie laat zeer te wensen over. Meerkeuzetoetsvormen laten immers niet eenduidig zien met welke spellingcategorie een leerling moeite heeft. Om dat precies te weten te komen moet er alsnog een controledictee worden afgenomen. Als er altijd een dictee afgenomen zou worden en de leerkracht analyseert zelf het foutenpatroon, dan weet deze in veel gevallen onmiddellijk welke spellingcategorie of regel nog geoefend moet worden, zonder enige omweg te maken.

NOTEN

1 We willen hier de opmerking van de Wijs rechtzetten, namelijk dat 18 leerlingen een te klein aantal zou zijn om überhaupt *t*-waarden over te berekenen. In het boek Ferguson, 1976 staat op pagina 487 een tabel met kritische *t*-waarden. Zonder gebruik te maken van SPSS kun je zelf de gevonden *t*-waarde uitrekenen. Vervolgens kan de gevonden *t*-waarde vergeleken worden, aan de hand van het aantal vrijheidsgraden, met de kritische *t*-waarde om zo te bepalen of deze wel of niet in het verwerpingsgebied ligt. In deze tabel staan zelfs kritische *t*-waarden met slechts één vrijheidsgraad. Waarom 17 (als $n = 18$) dan te weinig zou zijn, is niet duidelijk.

2 De eerste auteur van dit artikel stelt zich op het standpunt dat als we dan toch uitgaan van de juistheid van de klassieke testtheorie, we ons dan ook aan haar principes moeten houden. Het probleem met de huidige vorm van testen is echter veel fundamenteler. Uit het werk van Peter Molenaar (hoogleraar aan Penn State University in de VS) blijkt dat tests gebaseerd op groepen feitelijk niets over het individu kunnen zeggen (Molenaar, 2004, 2007, 2008).

3 *Orthopedagogiek: Onderzoek en Praktijk* is de voortzetting van het Tijdschrift voor Orthopedagogiek, het huisorgaan van de Vereniging O & A.

4 Het kwadraat van de correlatiecoëfficiënt vermenigvuldigd met 100 is het percentage verklaarde variantie van de ene variabele uit de andere op basis van een veronderstelde *lineaire* samenhang tussen de variabelen: (.80 * .80) * 100 = 64%. We zullen hier niet ingaan op het probleem van de veronderstelling dat de samenhang lineair is. Dit is namelijk een stilzwijgende assumptie waarvan we helemaal niet weten of die waar is.

5 Niemand accepteert het dat een meetlat of een thermometer 50 keer gebruikt moet worden om enigszins betrouwbaar de lengte respectievelijk de temperatuur te meten.

GERAADPLEEGDE LITERATUUR

Bosman, A.M.T. (2007). Zo leer je kinderen lezen en spellen. *Tijdschrift voor Orthopedagogiek, 46,* 451-465.

Brown, A.S. (1988). Encountering misspellings and spelling performance: Why wrong isn't right. *Journal of Educational Psychology, 80*, 488-494.

Dixon, M., & Kaminska, Z. (1997). Is it misspelled or is it mispelled? The influence of fresh orthographic information on spelling. *Reading and Writing: An interdisciplinary Journal, 9*, 483-498.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests (gewijzigde herdruk mei 2010)*. Amsterdam: NIP/COTAN.

Haan, M. de (2010). Spelling, ik heb het geweten! Meesterstuk, Master SEN, interne begeleiding.

Jacoby, L.L. & Hollingshead, A. (1990). Reading student essays may be hazardous to your spelling: Effects of reading incorrectly and correctly spelled words. *Canadian Journal of Psychology, 44*, 345-358.

Koning, L. (2008). *Cito Spelling; wat is er mis mee?* Lekkerkerk: Uitgeverij Pravoo.

Molenaar, P.C.M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2,* 201-218.

Molenaar, P.C.M. (2007). Consequences of the ergodic theorems for classical test theory, factor analysis, and the analysis of developmental processes. In: S.M. Hofer & D.F. Alwin (Eds.), *Handbook of cognitive aging: Interdisciplinary perspectives* (pp. 90-104). Thousand Oaks, CA: Sage.

Molenaar, P.C.M. (2008). On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. *Developmental Psychobiology, 50,* 60-69.

Schraven, B. (2010). *De omvang van een onderzoek en de validiteit van de Cito-Spellingtoets*. http://www.annabosman.eu/of http://www.zoleerjekinderenlezenenspellen.nl/

Schraven, J.L.M. (2009). *Zo leer je kinderen lezen en spellen (Handleiding methodiek)*. Zutphen: Stichting TGM/ www.zoleerjekinderenlezenenspellen.nl.

Schraven, J.L.M., Bosman, A.M.T., & van Eekhout, T. (2010). De nieuwe Cito-spellingtoets ter discussie. *Tijdschrift voor Orthopedagogiek (O en A), 49,* 75-86.

Vaughn, S., Schumm, J. S., & Gordon, J. (1993). Which motoric condition is most effective for teaching spelling to students with and without learning disabilities? *Journal of Learning Disabilities, 26,* 191-198.

Wijs, A. de (2010). Kritiek op toetsen Spelling steunt op losse gronden. *Orthopedagogiek: Onderzoek en Praktijk,49,* 000-000.

Wijs, de A., Krom, R., & van Berkel, S. (2006). *Leerling- en onderwijsvolgsysteem Spelling groep 4. Handleiding*. Arnhem: Cito.

Willemen, M., Bosman, A.M.T., & van Hell, J.G. (2000). Beter leren spellen tijdens het stellen. *Pedagogische Studiën, 77*, 173-182.

Willemen, M., Bosman, A.M.T., & van Hell, J.G. (2002). Leren stellen en niet vergeten correct te spellen. Het succes van de zelfcorrectietraining. *Tijdschrift voor Remedial Teaching, 10(1)*, 22-25.

OVER DE AUTEURS

Prof. dr. Anna M.T. Bosman is werkzaam aan de Radboud Universiteit Nijmegen bij de sectie Orthopedagogiek. Haar specialisme is leren lezen en spellen (www.annabosman.eu); *E-mail:* a.bosman@pwo.ru.nl. **Drs. José L.M. Schraven** is orthopedagoge en de auteur van 'Zo leer je kinderen lezen en spellen' (www.zoleerjekinderenlezenenspellen.nl); *E-mail:* zlkls@planet.nl. **Mw. Truus van Eekhout**, **B.Ed**. is remedial teacher en intern begeleider op openbare basisschool 'Het Kofschip' locatie Schrijvershof in Zevenaar (http://www.het-kofschip.nl); *E-mail*: teekhout@xs4all.nl