De validiteit van de Cito-spellingtoets gefalsifieerd[1]

B. Schraven[2]

Samenvatting

In 2006 introduceerde het Cito nieuwe spellingtoetsen. In de handleidingen staat dat deze bedoeld zijn om de vaardigheid in het omzetten van woorden in schriftbeelden te toetsen op individueel, groeps- en schoolniveau. Bij een deel van de toetsen zijn echter de dictee-opgaven vervangen door meerkeuzeopgaven. Bosman en collega's hebben op theoretische gronden betoogd, dat meerkeuzeopgaven niet geschikt zijn om het spellingniveau correct vast te stellen en de gebreken in de spellingvaardigheid van een leerling correct op te sporen. Daarnaast hebben zij de stelling van het Cito dat meerkeuzeopgaven hetzelfde meten als dictee-opgaven weerlegd: de scores van de twee typen opgaven komen niet overeen en uit de antwoorden op de meerkeuzeopgaven kan niet worden afgeleid of leerlingen een woord correct kunnen spellen en welke spelfouten zij daarbij maken. De door het Cito en de Cotan geleverde kritiek op het onderzoek van Bosman e.a. is niet terecht. Zo werd de omvang van de gebruikte toets door het Cito zelf bepaald en volgens de handleiding was deze geschikt om scores te berekenen. De hogere scores van de dicteeopgaven kunnen niet verklaard worden uit een leereffect. Dezelfde leerlingen hebben zowel de dicteeopgaven gemaakt als de meerkeuzeopgaven op het moment zoals voorgeschreven in de handleiding. De spellingtoets moet op individueel niveau geldige uitspraken opleveren over de spellingvaardigheid en spellingproblemen van onderzochte leerlingen signaleren. Dat blijkt niet het geval. De kritiek van het Cito dat de omvang van de onderzoeksgroep te klein zou zijn is voor deze conclusie niet relevant. De bewering van het Cito, dat uit hun eigen onderzoeken (met veel respondenten) blijkt, dat er met de spellingtoetsen niets mis, is niet terecht. De opzet van de onderzoeken van het Cito is namelijk ongeschikt om de validiteit van de spellingtoetsen vast te stellen; de Cito onderzoeksresultaten hebben uitsluitend betrekking op de betrouwbaarheid en niet op de validiteit. Ondanks de kritiek van het Cito op het werk van Bosman en collega's, erkent het Cito inmiddels wel dat de meerkeuzeopgaven iets anders meten dan de dicteeopgaven, dat meerkeuzeopgaven weinig diagnostische mogelijkheden bieden en dat meerkeuzeopgaven gevoelig zijn voor contexteffecten. Desondanks handhaaft het Cito de meerkeuzeopgaven. Waarom de Cotan het Cito hierin steunt en de onderwijsinspectie evenmin consequenties verbindt aan de aangetoonde ondeugdelijkheid van de Cito-spellingtoetsen blijft een raadsel.

Over de auteur

Drs. Ben Schraven studeerde sociale psychologie aan de Katholieke Universiteit Nijmegen en heeft zich vele jaren bezig gehouden met de beoordeling van de (methodologische) kwaliteit van voorstellen voor onderzoek en van uitgevoerd onderzoek, zowel gericht op theorieontwikkeling als gericht op analyse van concrete problemen: in hoeverre laten de opzet, aanpak en uitvoering van een onderzoek, zoals voorgenomen of gerealiseerd, de beoogde of getrokken conclusies toe? Zijn werkzaamheden verrichtte hij aan de Katholieke Universiteit (nu Radboud Universiteit) in Nijmegen, Vrije Universiteit te Amsterdam, de Sociale Verzekeringsraad in Amsterdam en laatstelijk bij het Ministerie van Sociale Zaken.

[1] Dit artikel werd gepubliceerd in Orthopedagogiek: Onderzoek en Praktijk (2013), 52, 459-475.
[2] Met dank aan I. Bazelier, A. Bosman, J. Schraven, C. van Luytelaer en H. van Pelt voor hun commentaar op een eerdere versie van deze tekst

Inleiding

Sinds 2006 brengt het Cito nieuwe spellingtoetsen op de markt. In publicaties van Bosman, Schraven en Van Eekhout (2010a, 2010b, Schraven, Bosman & van Eekhout, 201) wordt op theoretische en empirische gronden kritiek geleverd op het gebruik van meerkeuzeopgaven voor de meting van spellingvaardigheid in deze nieuwe Cito-spellingtoetsen. Zij komen tot de conclusie, dat meerkeuzeopgaven niet geschikt zijn om het spellingniveau van een leerling correct vast te stellen. Evenmin zijn meerkeuzeopgaven geschikt om een correcte diagnose te stellen van tekorten in de spellingvaardigheid van een leerling, zo concluderen zij.

Inmiddels hebben zowel het Cito (de Wijs 2010a, 2010b) als de COTAN (Cotan, 2010) op deze kritiek gereageerd. Het Cito (de Wijs, 2010b, p. 376) concludeert, dat hun "argumentatie voor het zich ondubbelzinnig uitspreken vóór het dictee en tegen meerkeuzeopgaven op losse gronden is gebaseerd". Als reden hiervoor geeft het Cito, dat "hun onderzoek, helaas, bij een veel te kleine groep leerlingen is uitgevoerd" (de Wijs, 2010b, p. 376). Ook de COTAN beweert (Cotan, 2010), dat "de resultaten uit dit onderzoek weinig overtuigend zijn", omdat "het gebaseerd is op een steekproef (bestaande uit) een homogene groep van slechts 18 leerlingen". Daarnaast suggereert het Cito (de Wijs, 2010a), dat uit hun onderzoek bij een veel groter aantal leerlingen zou blijken, dat "met de kwaliteit van de toetsen Spelling niets mis is!".

In hun kritiek op het onderzoek van Schraven et al. (2010; hierna aangeduid als: Kofschiponderzoek) gaan het Cito en de COTAN voorbij aan het karakter van dat Kofschiponderzoek en van het Cito-onderzoek. Daardoor hebben zij niet in de gaten, dat:

- het Kofschiponderzoek met de gekozen opzet juist wel geschikt is om de daarin centraal staande probleemstelling te onderzoeken, ook al is dat uitgevoerd bij een klein aantal leerlingen;
- de onderzoeken van het Cito niet geschikt zijn om de hierboven weergegeven conclusie van het Cito te trekken, ondanks hun omvang.

Ondanks hun kritiek op het Kofschiponderzoek neemt het Cito wel alle conclusies uit dat onderzoek (Schraven et al., 2010) over:

- de vaardigheid in het spellen (het omzetten van woorden in schriftbeelden) kan niet getoetst worden door middel van meerkeuzeopgaven;
- meerkeuzeopgaven bieden weinig diagnostische mogelijkheden;
- het antwoord op een meerkeuzeopgave wordt beïnvloed door de andere vet gedrukte woorden in de opgave.

Het positieve oordeel van de COTAN over de nieuwe Cito-spellingtoetsen en het negatieve oordeel van de COTAN over het Kofschiponderzoek (Cotan, 2010) worden niet overtuigend onderbouwd.

Schraven et al. (2010) waren de eersten die onderzoek naar de kwaliteit van de nieuwe Cito-spellingtoetsen publiceerden. De resultaten van de Cito-onderzoeken werden pas na hun artikel in 2010 gepubliceerd. Dat was vier jaar na de introductie van de nieuwe Cito-spellingtoetsen. Beslissingen over (het onderwijs aan) leerlingen in het basisonderwijs werden dus genomen op basis van toetsen waarvan niet door middel van onderzoek, hoe groot of klein ook, was aangetoond, dat die daarvoor geschikt zijn. Nu is dus duidelijk, dat die beslissingen worden genomen op basis van toetsen, die daarvoor niet geschikt zijn. Dat de Onderwijsinspectie bij scholen aandringt op het gebruik van de nieuwe Cito-spellingtoetsen is dan ook vreemd.

Wat beogen de (nieuwe) Cito-spellingtoetsen te meten?

De nieuwe Cito-spellingtoetsen dienen ter vervanging van de SVS-toetspakketten. Bij een deel van de toetsen zijn dicteeopgaven vervangen door meerkeuzeopgaven. Als redenen voor de introductie van nieuwe spellingtoetsen worden in de handleiding (de Wijs, Krom & van Berkel, 2006: p. 7) alleen de wijzigingen in de spellingregels en de afstemming op nieuwe spellingmethoden gegeven. De vervanging van dicteeopgaven door meerkeuzeopgaven wordt niet gemotiveerd. Spellen is een concrete, direct waarneembare activiteit. De meting van spellingvaardigheid kan daardoor rechtstreeks/direct plaats vinden. Een indirecte meting van spellingvaardigheid zal altijd minder zuiver/valide zijn dan een directe meting. Er is ook geen reden de meting van spellingvaardigheid indirect uit te voeren, zoals, bijvoorbeeld wel het geval is, bij de vaardigheid in het besturen van een straalvliegtuig, waarbij het verstandig is die eerst indirect te meten in een simulator vóór de definitieve proef op de som in een echt vliegtuig.

Uit de handleiding (de Wijs et al., 2006, p. 7) en de productinformatie (Cito, 2010) blijkt, dat de nieuwe spellingtoetsen in hun geheel, dus inclusief de daarin opgenomen meerkeuzeopgaven, bedoeld zijn ter vervanging van de SVS-toetspakketten (zie ook: (de Wijs, 2010b, p. 370). Deze zijn dus niet als een aanvulling daarop bedoeld. In de handleiding wordt ook niet vermeld, dat de meerkeuzeopgaven bedoeld zijn om een andere vaardigheid te meten dan de dicteeopgaven. Integendeel, in alle handleidingen bij de toetsen Spelling staat namelijk de volgende passage (de Wijs et al., 2006, p. 11): “de verschillende opgaventypen in de toetsen Spelling (woorddictee, zinsdictee, meerkeuzeopgave) .... zeggen iets over dezelfde spellingvaardigheid. De vaardigheidsscores van leerlingen die een toets Spelling gemaakt hebben, kunnen dus altijd onderling vergeleken worden, ook al hebben de leerlingen niet allemaal dezelfde soort opgaven gemaakt.” In strijd hiermee, beweert het Cito nu (de Wijs, 2010b, p. 317), dat “het Cito niet beweert, dat je alle dicteeopgaven kunt vervangen door meerkeuzeopgaven, of andersom, zonder het toetsresultaat te beïnvloeden”. Hoe is deze opvatting van het Cito te rijmen met de passage in de handleidingen van de spellingtoetsen.

Ook uit de instructies voor het gebruik van de spellingtoetsen blijkt, dat het Cito ervan uitgaat, dat met de meerkeuzeopgaven dezelfde vaardigheid wordt gemeten als met de dicteeopgaven , want:

- voor de berekening van de score dienen de goede antwoorden op de dictee- en meerkeuzeopgaven bij elkaar opgeteld te worden (de Wijs et al., 2006, p. 29);
- voor een nadere analyse van een fout antwoord op een meerkeuzeopgave worden in het hulpboek dicteeopgaven aangeboden (de Wijs et al., 2006, p. 42).

Ook uit het feit dat de Eindtoets Basisonderwijs louter uit meerkeuzeopgaven bestaat, blijkt, dat het Cito destijds helemaal niet de bedoeling had om met de introductie van meerkeuzeopgaven een ander aspect van de spellingvaardigheid te meten dan met de dicteeopgaven. Hieruit kan alleen maar geconcludeerd worden, dat deze nieuwe spellingtoetsen in hun geheel bedoeld waren om hetzelfde te meten als de SVS-toetspakketten, namelijk de vaardigheid in spellen. Door het Cito (de Wijs et al., 2006, p. 9) wordt van 'spellen' de volgende definitie gegeven: “Bij spellen gaat het erom woorden om te zetten in schriftbeelden”. In de productinformatie (Cito, 2010) wordt 'spelling' door het Cito omschreven als “het foutloos schrijven van teksten”.

De nieuwe spellingtoetsen zijn een onderdeel van het 'Leerling- en onderwijsvolgsysteem' (LOVS) (de Wijs et al., 2006, p. 7; de

Wijs et al 2010b: p. 370). Volgens de handleiding (de Wijs et al., 2006, p. 7) en de productinformatie (Cito, 2010) kunnen met het LOVS de vorderingen van individuele leerlingen, groepen leerlingen en het onderwijs op een school gevolgd worden. De gegevens van het LOVS bieden, volgens de handleiding (de Wijs et al., 2006, p. 7) en de productinformatie (Cito, 2010), de mogelijkheid het onderwijs op individueel niveau, groepsniveau en schoolniveau te evalueren. Volgens de handleiding (de Wijs et al., 2006, p. 39), is het een belangrijke taak van de leerkracht na te gaan wat de gevonden resultaten voor diens onderwijs betekenen.

Dat betekent, dat iedere leerkracht op elke basisschool overal in Nederland bij iedere leerling afzonderlijk ervan uit mag gaan, dat deze toetsen geschikt zijn om:

- het spellingniveau van die leerling correct vast te stellen,
- de gebreken in de spellingvaardigheid van die leerling correct op te sporen.

De leerkracht mag er dus vanuit gaan, dat de nieuwe spellingtoetsen (met meerkeuzeopgaven) een goede indicatie opleveren van het spellingniveau en van de problemen met spelling bij iedere leerling.

Afgaande op de officiële productinformatie over en instructies voor het gebruik van de nieuwe Cito-spellingtoetsen gingen Schraven et al., (2010) in hun onderzoek uitgevoerd in 2008 er terecht vanuit, dat ook de meerkeuzeopgaven in de nieuwe spellingtoetsen, evenals de dicteeopgaven, bedoeld zijn om de vaardigheid in het foutloos schrijven van woorden en de problemen daarmee vast te stellen, en dat dat bij iedere leerling, waarvoor de gebruikte toets bedoeld is, een correcte (geldige) indicatie daarvan oplevert.

Dat het Cito (de Wijs 2010a, p. 370; 2010b, p. 372) in 2010, na publicatie van bijna alle nieuwe spellingtoetsen en na publicatie van de kritiek daarop van Bosman, Schraven en Van Eekhout (2010a, 2010b), van “mening” is, dat “spelling meer aspecten heeft dan het zelf foutloos schrijven van woorden”, en dat ook “het opsporen van spelfouten in het onderwijs aan bod zou moeten komen” doet daaraan niets af. Uit niets blijkt, dat deze 'mening' destijds aanleiding was voor de vervanging van dicteeopgaven door meerkeuzeopgaven.

Het Kofschiponderzoek naar de validiteit van een Cito-spellingtoets

Doel van het Kofschiponderzoek

In 2008 is in het Kofschiponderzoek (Schraven et al., 2010) onderzocht, of de vervanging van dicteeopgaven in de oude SVS-toetspakketten door meerkeuzeopgaven in (een deel van) de nieuwe LOVS-spellingtoetsen gerechtvaardigd/verantwoord is.

Het Kofschiponderzoek was dus een onderzoek ter toetsing van de stelling van het Cito (de Wijs et al., 2006, p. 11), dat een toets met meerkeuzeopgaven de spellingvaardigheid en problemen daarin van iedere leerling evengoed meet als een toets opgebouwd uit dicteeopgaven[3]. Dit onderzoek was dus niet bedoeld als een onderzoek naar de eigenschappen van de Cito-spellingtoetsen, zoals De Wijs (de Wijs, 2010b, p. 370), ten onrechte, beweert.

Om de juistheid van die stelling te toetsen zijn in het Kofschiponderzoek de antwoorden op de meerkeuzeopgaven direct vergeleken met de wijze waarop dezelfde leerlingen de gemarkeerde woorden in de meerkeuzeopgaven opschrijven. Of leerlingen in staat zijn de bij de meerkeuzeopgaven gemarkeerde woorden correct te

[3] Het Kofschiponderzoek was dus een hypothesetoetsend onderzoek en niet een instrumenteel-nomologisch onderzoek naar de psychometrische kenmerken van de Cito-spellingtoetsen hypothesetoetsend onderzoek en niet een instrumenteel-nomologisch onderzoek naar de psychometrische kenmerken van de Cito-spellingtoetsen

spellen, werd zo direct vastgesteld.

In het Kofschiponderzoek ging het dus om de vraag, of meerkeuzeopgaven de spellingvaardigheid van een leerling evengoed meten als dicteeopgaven. Dit is een vraag naar de validiteit van die meerkeuzeopgaven: zijn op basis van meerkeuzeopgaven geldige (= juiste/valide) uitspraken te doen over de spellingvaardigheid van een leerling?

Dan is het opmerkelijk, dat het Cito in zijn kritiek op het Kofschiponderzoek het steeds heeft over de “betrouwbaarheid” van de onderzochte spellingtoets (de Wijs, 2010b, p. 372, 373). Het lijkt erop dat hier het begrip ‘validiteit’ van een toets verward wordt met de ‘betrouwbaarheid’.

De betrouwbaarheid van een toets is relevant, want een toets hoort 'stabiel' te zijn en de uitslag van de toets (de score) dient onafhankelijk te zijn van het tijdstip en de plaats van afname: in principe dient bij herhaling van de afname bij dezelfde personen maar op een ander tijdstip de uitslag hetzelfde te zijn. Wanneer een toets bij iedere afname een heel andere score oplevert (dus onbetrouwbaar is), dan is het zeker, dat deze het beoogde kenmerk niet goed meet (dus niet valide is). Een zeer betrouwbare toets kan echter toch volstrekt invalide zijn (zie: Bosman et al. 2010b). Betrouwbaarheid is, namelijk, wel een noodzakelijke maar geen voldoende voorwaarde voor validiteit van een toets.

Het Kofschiponderzoek is in 2008 uitgevoerd bij alle leerlingen in groep 4 van basisschool 'Het Kofschip'. Bij hen is op het aangewezen moment precies volgens de handleiding de toets 'Spelling medio groep 4' afgenomen. Het Kofschiponderzoek was niet bedoeld om een uitspraak te doen over de spellingvaardigheid van alle leerlingen in groep 4 van het basisonderwijs in Nederland. Met het Kofschiponderzoek werd beoogd een uitspraak te doen over de geschiktheid van meerkeuzeopgaven voor de meting van de spellingvaardigheid van een leerling. Het object van het Kofschiponderzoek was dus niet 'leerlingen van groep 4 van het basisonderwijs in Nederland'. Het onderwerp van dat onderzoek was de meerkeuzemodule uit de toets 'Spelling medio groep 4'. De deelnemers aan dit onderzoek hoeven daarom geen “afspiegeling van de totale Nederlandse populatie van groep 4-leerlingen” te zijn, zoals het Cito (de Wijs, 2010b, p. 372-373) beweert. Iedere leerling uit de doelgroep van de toets is op zich geschikt voor toetsing van die gesuggereerde validiteit van die toetsmodule. De deelnemers aan dit onderzoek vormen dan ook geen 'steekproef' maar zijn 18 herhalingen (replicaties) van hetzelfde onderzoek. De COTAN had dat kennelijk niet in de gaten, want zij duidt de groep, waarbij het onderzoek is uitgevoerd, ten onrechte aan als “steekproef” (Cotan, 2010).

Resultaten van het Kofschiponderzoek

Nadat de toets 'Spelling medio groep 4' was afgenomen, zoals voorgeschreven in de handleiding, zijn alle gemarkeerde woorden uit de meerkeuzeopgaven van deze toets in dictee-vorm aan dezelfde leerlingen aangeboden. Daardoor was die vergelijking van de antwoorden op de meerkeuzeopgaven met de spelling van dezelfde toetswoorden door iedere leerling mogelijk.

De dictee-opgaven zijn dus niet door Schraven, Bosman en van Eekhout ontworpen, zoals het Cito beweert (2013), maar zijn de door het Cito zelf ontworpen items die in dicteevorm in plaats van meerkeuzevorm zijn aangeboden. De resultaten van het Kofschiponderzoek konden niet vergeleken worden met de psychometrische kenmerken (zoals de betrouwbaarheid) van de toets, omdat het Cito daarover toen nog niets gepubliceerd had.

Spellingniveau

Het spellingniveau van een leerling kan uitgedrukt worden in een score die aangeeft hoeveel woorden een leerling correct kan

spellen. In het Kofschiponderzoek (Schraven et al., 2010, p. 79, 80) is bij slechts 3 (van de 18) leerlingen de score vastgesteld met de meerkeuzeopgaven hetzelfde als de score op basis van de dicteeopgaven. Bij de meeste leerlingen geeft de score op de meerkeuzeopgaven dus geen goede indicatie van hun spellingniveau (Schraven et al., 2010, p. 80):

- bij de 15 leerlingen met ongelijke scores verschilt de score op basis van de meerkeuzeopgaven gemiddeld 3.2 van de score op basis van de dicteeopgaven;
- bij alle 18 leerlingen verschilt de score op basis van de meerkeuzeopgaven gemiddeld 2.6 van de score op basis van de dicteeopgaven.

De meerkeuzetoets levert dus geen correcte (valide) meting van het spellingniveau op, want dan had bij alle leerlingen de score vastgesteld met de meerkeuzeopgaven, hetzelfde moeten zijn als de score vastgesteld met de dicteeopgaven. De meerkeuzeopgaven zijn dus niet geschikt om een correcte inschatting te maken van het spellingniveau van een leerling.

Deze uitkomst van het Kofschiponderzoek betekent dus ook een ontkrachting van de bewering in de handleidingen van alle nieuwe spellingtoetsen (de Wijs et al., 2006, p.11), dat “de vaardigheidsscores van leerlingen die een toets Spelling gemaakt hebben, altijd onderling vergeleken kunnen worden, ook al hebben de leerlingen niet allemaal dezelfde soort opgaven (woorddictee, zinsdictee, meerkeuzeopgave) gemaakt”.

Ten onrechte wordt door De Wijs (de Wijs, 2010b, p. 374) aan Schraven et al., (2010) de conclusie toegeschreven, dat de spellingniveauscores niet consistent zijn. Natuurlijk hadden de scores op basis van de meerkeuzeopgaven en op basis van de dicteeopgaven hetzelfde moeten zijn om van een valide meting van het spellingniveau door middel van de meerkeuzeopgaven te kunnen spreken.

Daarnaast verwijt het Cito (de Wijs, 2010b, p. 373) Schraven et al. (2010), dat zij een score berekenen voor de (gehele) module M4 Vervolg 2, want die module “bevat slechts 25 opgaven en dat is erg weinig om een betrouwbaar beeld te kunnen geven van iemands spellingvaardigheid”. Dit verwijt is niet terecht, want:

- niet Schraven et al. (2010) zijn verantwoordelijk voor de omvang van die module maar het Cito;
- in de handleiding wordt datzelfde aantal opgaven wel geschikt gevonden om 'zwakkere spellers' van 'betere spellers' te onderscheiden (de Wijs et al., 2006, p. 16);
- in het eigen onderzoek werkt het Cito met scores op basis van slechts 12 of 13 (de Wijs, 2010b, p. 373).

Het lijkt erop dat het Cito niet in de gaten heeft, dat het in het Kofschiponderzoek niet gaat om de vaststelling van de spellingvaardigheid van leerlingen die aan het onderzoek deelnamen, maar om de vaststelling of de score op basis van de meerkeuzeopgaven hetzelfde is als de score op basis van dicteeopgaven met dezelfde toetswoorden.

Diagnose van de spellingvaardigheid

Bij de diagnose van de spellingvaardigheid gaat het erom vast te stellen:

- of een leerling een bepaald woord wel of niet correct kan spellen,
- wat de aard van een eventuele spelfout is.

Voor een correcte (valide) diagnose van de spellingvaardigheid van een leerling kunnen meerkeuzeopgaven alleen dan gebruikt worden, wanneer er sprake is van het volgende consistente antwoordpatroon:

- een leerling die een fout gespeld woord in een meerkeuzeopgave aanstreept,

schrijft dat woord ook correct op bij het dictee

- een leerling die het fout gespelde woord in een meerkeuzeopgave niet aanstreept, schrijft dat woord incorrect op bij het dictee.

In het Kofschiponderzoek komt dit consistente antwoordpatroon bij slechts één woord (stank) voor (Schraven et al., 2010, p. 82). Bij twee andere woorden (hond, klank) vormt 1 leerling een uitzondering op dit consistente antwoordpatroon. Alleen bij deze drie woorden leidt het antwoord op de meerkeuzeopgave bij (bijna) alle leerlingen tot een correcte conclusie over hun vermogen om het betreffende toetswoord correct te spellen, terwijl dat bij alle (25) toetswoorden het geval had moeten zijn.

Bij de overige 22 meerkeuzeopgaven geeft een antwoord van 2 tot 11 (van de 18) leerlingen geen goede indicatie of een leerling het betreffende toetswoord correct kan spellen. De misverstanden die dat kan opleveren blijken uit de volgende voorbeelden (Schraven et al., 2010, p. 82):

- Bij opgave 6 kiezen 14 (78%) van de (18) leerlingen het woord 'aarrecht', terecht, als onjuist gespeld woord; slechts 7 van hen (50%) blijken het woord correct te kunnen spellen.
- Bij opgave 1 kiezen 14 leerlingen (78%) 'swak', ten onrechte, niet als onjuist gespeld woord; 11 van hen (79%) blijken dat woord wel correct te kunnen spellen. In het hulpboek wordt er echter van uitgegaan, dat alle leerlingen die in opgave 1 'swak' niet gekozen hebben als onjuist gespeld woord, de spelling van dat woord niet beheersen.

Het Kofschiponderzoek *laat zien* dat

- de meerkeuzeopgaven niet geschikt zijn om een correcte inschatting te maken of een leerling een woord (uit een bepaalde spellingcategorie) correct kan schrijven en wat de aard van een eventuele spellingprobleem is;
- het hulpboek aan de leerkracht bij veel leerlingen een onjuiste interpretatie van een onjuist antwoord op een meerkeuzeopgave presenteert.

Uit het Kofschiponderzoek *blijkt* dat

- de meeste leerlingen meer toetswoorden correct spellen bij de dicteeopgaven dan fout gespelde woorden aanstrepen bij de daaraan voorafgaande meerkeuzeopgaven (Schraven et al., 2010, p. 78);
- de meeste toetswoorden door meer leerlingen correct worden gespeld dan daaraan voorafgaand als fout gespeld woord aangestreept (Schraven et al., 2010, p. 82).

Het zijn dezelfde leerlingen die beide opgaven (eerst meerkeuze-, later dictee-) hebben gemaakt. Toch kunnen deze uitkomsten niet verklaard worden met een *leereffect*. Voor zover bekend, is er, namelijk, geen leertheorie die voorspelt, dat het zien van een fout gespeld woord de kans vergroot, dat datzelfde woord vervolgens wel correct wordt gespeld. Het is echter niet uitgesloten, dat er sprake was van een leereffect van de beantwoording van de meerkeuzeopgaven op de antwoorden op de dicteeopgaven. De leerlingen werden eerst geconfronteerd met de fout gespelde woorden in de meerkeuzeopgaven. Pas daarna kregen zij de opdracht om die woorden correct op te schrijven. Een eventueel leereffect kan dan alleen maar geleid hebben tot meer fout gespelde toetswoorden op het dictee dan zonder voorafgaande meerkeuzeopgaven. Dat zou betekenen dat het eigenlijke verschil in de totaalscore en in de 'antwoorden' per toetswoord tussen de meerkeuzeopgaven en de dicteeopgaven nog groter is dan vastgesteld in het Kofschiponderzoek.

Volgens het Cito “lijkt de kans klein dat woorden tijdens de toetsafname foutief

worden ingeprent door leerlingen die de juiste schrijfwijze van het woord blijkbaar eerder niet hadden kunnen onthouden" (de Wijs, 2010b, p. 371, 372). Welke leertheorie de COTAN (Cotan, 2010) op het oog had, toen zij als kritiek op het Kofschiponderzoek formuleerde, dat "niet valt uit te sluiten, dat leereffecten een rol hebben gespeeld", is dan ook een raadsel.

Kenmerken van de leerlingen

Ook de kenmerken van de leerlingen bieden geen alternatieve verklaring voor de uitkomsten van het Kofschiponderzoek. De afname van zowel de meerkeuzeopgaven als de dicteeopgaven met dezelfde toetswoorden bij dezelfde leerlingen biedt een belangrijk methodologisch voordeel. Daarmee wordt uitgesloten, dat verschillen in antwoorden tussen meerkeuzeopgaven en dicteeopgaven een gevolg zijn van verschillen in onderwijservaring, leerkracht, gebruikte lesmethode, intelligentie, achtergrond tussen de leerlingen die de meerkeuzeopgaven hebben gemaakt, en de leerlingen die de dicteeopgaven hebben gemaakt.

Zowel het Cito (de Wijs, 2010b, p. 372) als de COTAN (Cotan, 2010) plaatsen bij de deelnemers aan het Kofschiponderzoek de kritische kanttekening, dat deze bestaat uit een homogene groep van alleen maar 'betere spellers'. Daaruit blijkt, dat het de COTAN en ook het Cito ontgaan is, dat in groep 4, volgens de handleiding, de meerkeuzeopgaven alleen aan 'betere spellers' voorgelegd mogen worden. De COTAN (Cotan, 2010) legt niet uit, hoe die "homogeniteit" van de onderzoeksgroep een alternatieve verklaring voor de gevonden discrepantie tussen antwoorden op de meerkeuzeopgaven en de dicteeopgaven biedt.

Het Kofschiponderzoek is uitgevoerd bij *één groep op één basisschool*. Bij deze groep bleek, dat de antwoorden van de leerlingen op de meerkeuzeopgaven niet gebruikt mogen worden voor de bepaling van hun spellingniveau en voor de diagnose van de gebreken in hun spellingvaardigheid.

Uitgaande van de handleiding zou de gebruikte meerkeuzetoets daarvoor wel geschikt zijn, want de deelnemers aan het onderzoek voldeden aan de vereiste kenmerken en de afname vond plaats op het voorgeschreven moment. Volgens de handleiding is het dan mogelijk om het onderwijs op leerling- en groeps- (en school)niveau te evalueren met de gebruikte toets.

Uit het Kofschiponderzoek moet geconcludeerd worden, dat de stelling van het Cito (de Wijs et al. 2006, p. 11), dat een toets met meerkeuzeopgaven de spellingvaardigheid van iedere leerling evengoed meet als een toets opgebouwd uit dicteeopgaven, verworpen moet worden. Ook al is dit onderzoek bij slechts één groep van één school uitgevoerd, toch is dat voldoende voor deze conclusie, want bij deze groep was dat niet het geval en had dat volgens die stelling wel moeten zijn.

Zoals één zwarte zwaan voldoende is om de theorie 'alle zwanen zijn wit' te verwerpen (niet geldig te verklaren), volgens het falsificatieparadigma, zo is de ongeschiktheid van de Cito-spellingtoets met meerkeuzeopgaven bij één groep (4) van één basisschool voldoende om de bewering van het Cito, dat "met de kwaliteit van de toetsen Spelling niets mis is!" te ontkrachten. Dat de COTAN (Cotan, 2010) het Kofschiponderzoek toch "weinig overtuigend" vindt, geeft aan, dat de COTAN niet vertrouwd is met de principes van op falsificatie gericht hypothese-toetsend onderzoek.

Het Cito (de Wijs, 2010b, p. 373) beweert, dat het Kofschiponderzoek vanwege zijn omvang "de toets der kritiek niet kan doorstaan". Alleen bij afname van de toets bij een groot aantal leerlingen die "een representatieve afspiegeling vormen van de totale Nederlandse populatie van groep 4-leerlingen" kunnen, volgens het Cito (de Wijs, 2010b, p. 272), uitspraken over de

toets gedaan worden. Het Kofschiponderzoek was echter niet bedoeld om in zijn algemeenheid de psychometrische kenmerken van de toets 'Spelling medio groep 4' vast te stellen maar om de stelling te toetsen, dat meerkeuzeopgaven de spellingvaardigheid evengoed meten als dicteeopgaven. Omdat de Cito-spellingtoetsen juist bedoeld zijn om op individueel leerlingniveau uitspraken daarover te doen (de Wijs et al. 2006, p. 7), is er helemaal geen representatieve steekproef nodig om deze stelling van het Cito te toetsen, als de deelnemers uit het onderzoek maar behoren tot "de populatie waarvoor die toets bestemd is".

Nu uit het Kofschiponderzoek blijkt, dat de meerkeuzeopgaven leiden tot onjuiste uitspraken over de spellingvaardigheid en problemen daarin van een leerling, moet geconcludeerd worden, dat de nieuwe Cito-spellingtoetsen met meerkeuzeopgaven niet geschikt zijn om de spellingvaardigheid en de gebreken daarin van een leerling in een groep van een basisschool waar dan ook in Nederland vast te stellen. Het was dus niet gerechtvaardigd/verantwoord de dicteeopgaven uit de SVS-toetspakketten in de nieuwe LOVS-spellingtoetsen te vervangen door meerkeuzeopgaven.

Onderzoek en conclusies van en over het Cito

Cito-onderzoeken

Om aan te tonen, dat er "met de kwaliteit van de toetsen Spelling niets mis is!" verwijst het Cito (de Wijs, 2010b, p. 373-374) naar een aantal door hen uitgevoerde onderzoeken (hierna aangeduid als: Cito-onderzoeken).[4] Deze Cito-onderzoeken zijn uitgevoerd bij enkele tientallen scholen en daaraan hebben, respectievelijk, 782, 1.318 en 1.928 leerlingen deelgenomen. Het lijkt erop, dat die Cito-onderzoeken bedoeld waren om de psychometrische kenmerken van de nieuwe spellingtoetsen vast te stellen. Dan is het noodzakelijk, dat deze uitgevoerd zijn bij een steekproef die representatief is voor alle leerlingen waarvoor de spellingtoetsen bedoeld zijn. In een dergelijk onderzoek zal voor iedere relevante deelcategorie, op basis van, onder andere, de kenmerken die door het Cito (de Wijs, 2010b, p. 272-273) worden vermeld, onderzocht moeten worden of de toets de spellingvaardigheid voldoende betrouwbaar en valide meet. Daarvoor dient iedere relevante deelcategorie in voldoende mate in de steekproef vertegenwoordigd te zijn. Daardoor is bij een dergelijk onderzoek een grote steekproef van (scholen en) leerlingen noodzakelijk.

Het is nog maar de vraag, of het aantal deelnemers aan de vermelde Cito-onderzoeken wel groot genoeg was om de psychometrische kenmerken van de nieuwe Cito-spellingtoetsen vast te stellen. In ieder geval, heeft ook het Cito onderzoek gedaan naar psychometrische kenmerken van dezelfde toets als de toets die centraal stond in het Kofschiponderzoek, 'Spelling medio groep 4'. Belangrijker dan de omvang van het Cito-onderzoek is de opzet daarvan.

Spelling is een direct waarneembare activiteit. De *validiteit* van de antwoorden van leerlingen op de meerkeuzeopgaven (of die antwoorden overeenkomen met de vaardigheid in het correct spellen van die woorden door die leerlingen) kan dus direct worden vastgesteld door de antwoorden op de meerkeuzeopgaven te vergelijken met de wijze waarop dezelfde leerlingen de gemarkeerde woorden opschrijven. Dat gebeurt echter niet in het Cito-onderzoek.

Uit de beschrijving van de opzet van het Cito-onderzoek (de Wijs, 2010b, p. 373) blijkt namelijk:

[4] Van deze onderzoeken is geen algemeen toegankelijke/beschikbare publicatie aangetroffen.

- een kwart van de onderzoekdeelnemers heeft de module met meerkeuzeopgaven (door het Cito aangeduid als: M4 Vervolg 2) helemaal niet gemaakt en alleen de toetswoorden in dicteeopgaven aangeboden gekregen ('groep 1'), zodat hun vaardigheid in het opschrijven van de toetswoorden niet vergeleken kan worden met hun antwoorden op de meerkeuzeopgaven;
- een kwart van de onderzoekdeelnemers heeft alleen maar die module gemaakt ('groep 4') en dezelfde toetswoorden niet in dicteevorm aangeboden gekregen, zodat hun antwoorden niet vergeleken kunnen worden met het feitelijk opschrijven van de betreffende woorden;
- de andere helft van de onderzoekdeelnemers heeft slechts de helft van de module 'M4 Vervolg 2' gemaakt en daarnaast een dictee met de toetswoorden uit de andere dan de door hen gemaakte helft van de module 'M4 Vervolg 2', zodat ook hun antwoorden bij de meerkeuzeopgaven niet vergeleken kunnen worden met hun vaardigheid in het schrijven van dezelfde toetswoorden.

Door deze opzet van het Cito-onderzoek is het dus niet mogelijk na te gaan, of op basis van (een toets met) de module met meerkeuzeopgaven geldige uitspraken gedaan kunnen worden over het spellingniveau en problemen in de spellingvaardigheid van een leerling.

Daarnaast is het opvallend, dat het Cito de meerkeuzeopgaven (uit module 'M4 Vervolg 2') ook heeft voorgelegd aan leerlingen die als 'zwakke spellers' gekwalificeerd werden (de Wijs, 2010b, p. 374), terwijl dat in strijd is met de handleiding.

Welke probleemstelling het Cito met de gekozen opzet wel heeft willen onderzoeken is een raadsel. In ieder geval, presenteert het (de Wijs, 2010b) geen uitkomsten die in strijd zijn met de conclusie uit het Kofschip-onderzoek (Schraven et al., 2010), dat uitspraken over het spellingniveau en over problemen in de spellingvaardigheid van leerlingen op basis van een toets met meerkeuzeopgaven niet geldig zijn.

Misschien vanwege de onmogelijkheid in het Cito-onderzoek om conclusies te trekken over de geldigheid van de meting van spellingvaardigheid door middel van meerkeuzeopgaven, besteedt het Cito (de Wijs, 2010b) veel aandacht aan de *betrouwbaarheid* van de metingen en aan correlaties tussen scores van leerlingen.

Opvallend is, dat het Cito voor de vaststelling van de betrouwbaarheid van de toetsen gekozen heeft voor een op allerlei veronderstellingen berustende, indirecte, statistische benadering. Het Cito had die ook direct vast kunnen stellen door middel van een herhaalde afname van een toets, want de kans op een leereffect acht het Cito in zo'n geval klein (de Wijs, 2010b, p. 371, 372). Informatie over de betrouwbaarheid van een toets, hoe dan ook vastgesteld, zegt echter weinig over de geschiktheid van die toets voor gebruik in de praktijk. Die hangt niet af van de betrouwbaarheid maar van de validiteit daarvan.

Het Cito vindt het kennelijk (de Wijs, 2010b, p. 374) ook van belang, dat scores op basis van dicteeopgaven hoog correleren met scores op basis van meerkeuzeopgaven (met andere toetswoorden!). Deze hoge correlatie geeft echter alleen maar aan, dat de antwoorden op deze verschillende opgaventypen een overeenkomstig onderliggend kenmerk hebben. Het lijkt voor de hand te liggen, dat dat kennis van spellingregels is. Die hoge correlatie bewijst niet, dat met die verschillende opgavetypen dezelfde vaardigheid gemeten wordt.

Zeven jaar na publicatie van de eerste nieuwe Cito-spellingtoetsen heeft het Cito in geen enkele publicatie onderzoekresulta-

ten gepubliceerd waaruit blijkt dat deze spellingtoetsen wel valide zijn, zelfs niet in de reacties op onderzoeken van anderen waaruit blijkt dat deze niet valide zijn. Ook in de meest recente poging van het Cito (2013) om te bewijzen dat er met de Cito-spellingtoetsen niets mis is, worden geen onderzoekresultaten vermeld waaruit blijkt dat deze wel valide zijn.

Conclusies van het Cito

Ook al blijkt dat niet uit het eigen Cito-onderzoek, toch erkent het Cito nu (de Wijs, 2010b, p. 370, 371, 375), dat de vaardigheid in "het *zelf foutloos schrijven*" *alleen* getoetst kan worden *door middel van dictee-opgave***n** en niet door middel van meerkeuzeopgaven. Daarmee onderschrijft het Cito dus de belangrijkste uitkomst van het Kofschiponderzoek. Daarmee erkent het Cito vooral, dat de stelling in de handleidingen (de Wijs et al., 2006, p. 11), meerkeuzeopgaven geschikt zijn om de vaardigheid in het "omzetten van woorden in schriftbeelden" en problemen daarbij correct te meten, niet juist is.

Of de meerkeuzeopgaven wel geschikt zijn om "fout gespelde woorden te herkennen", zoals het Cito nu beweert (de Wijs, 2010b, p. 370, 371, 375, 376), is nog maar de vraag. Bij een meerkeuzeopgave gaat het om een sorteer/keuzetaak: het kiezen van een fout gespeld woord uit de gemarkeerde woorden, wetend dat (zeker en uitsluitend) één van die woorden onjuist gespeld is. Het opsporen van spellingfouten in een doorlopende tekst, niet-wetend of er fout gespelde woorden in staan, is geen sorteer/keuzetaak maar een vigilantietaak.

Het "omzetten van een woord in een schriftbeeld", 'het kiezen van het fout gespelde woord uit meerdere woorden, wetend dat er (slechts) één daarvan fout gespeld is' en 'het opsporen van een fout gespeld woord in een doorlopende tekst, zonder te weten of er een fout gespeld woord tussen staat' (zoals bij het nakijken van een zelf geschreven tekst) zijn niet verschillende "opgaven waarmee je dezelfde vaardigheid wil meten", zoals het Cito beweert (de Wijs, 2010b, p. 371), maar zijn verschillende taken waarin kennis van de spellingregels gemeenschappelijk is. De uitvoering van die verschillende taken vergt verschillende vaardigheden in het toepassen van die spellingregels. Het lijkt erop, dat het Cito 'kennen' en 'kunnen' door elkaar haalt, wanneer het stelt, dat "de onderliggende vaardigheid dezelfde is" (de Wijs, 2010b, p. 371). Het Cito haalt 'kennen' en 'kunnen' ook door elkaar, wanneer het ter rechtvaardiging van het onderscheid in "actieve" en "passieve spelling", nu zegt (de Wijs, 2010b, p. 372), dat het "spelling wil omschrijven als: weten wat de juiste schrijfwijze van een woord is". Daarbij realiseert het Cito zich ook niet, deze "omschrijving" afwijkt van de definitie in de handleiding (de Wijs et al., 2006, p. 9): "Bij spellen gaat het erom woorden om te zetten in schriftbeelden". Ook al blijkt dat niet uit hun onderzoek, toch erkent het Cito inmiddels:

- de meerkeuzeopgaven bieden *weinig mogelijkheden tot diagnostiek* (de Wijs, 2010b, p. 371);
- vrees dat leerkrachten foute beslissingen nemen op basis van de meerkeuzeopgaven is niet ongegrond (de Wijs, 2010b, p. 376).

Het Cito beweert echter ook, dat "de meerkeuzeopgaven toch diagnostische mogelijkheden kennen via een omgekeerde bewijsvoering" (de Wijs, 2010b, p. 371). Deze omgekeerde bewijsvoering houdt in, dat "als een leerling een meerkeuzeopgave fout beantwoordt, ... het de echte fout niet heeft gezien .... die in een bepaalde spellingcategorie valt die de leerling blijkbaar nog niet volledig beheerst, want anders had hij of zij die fout wel ontdekt" (de Wijs, 2010b, p. 371).

Uit de uitkomsten van het Kofschiponderzoek (Schraven et al., 2010, p. 82: Tabel 4)

blijkt echter, dat deze 'theorie van de omgekeerde bewijsvoering' niet opgaat, want veel leerlingen die een fout gespeld woord bij de meerkeuzeopgaven niet aanstreepten, blijken dat woord wel correct te kunnen spellen bij het dictee. De uitkomsten van het Kofschiponderzoek laten juist zien, dat de meerkeuzeopgaven niet geschikt zijn voor diagnose van de spellingvaardigheid van een leerling. Het is opmerkelijk, dat dat niet door het Cito erkend wordt. Sterker nog, het blijkt dat het Cito aan Tabel 4 (Schraven et al., 2010, p. 82) de verkeerde conclusie verbindt. De inconsistentie in de antwoordpatronen die blijkt uit Tabel 4, geeft namelijk geen antwoord op de vraag of “je het spellingniveau van een leerling kunt bepalen met” meerkeuzeopgaven, zoals het Cito beweert (de Wijs, 2010b, p. 374). Het antwoord op die vraag is te vinden in de Tabellen 2 en 3 (Schraven et al., 2010, p. 79, 80). Uit Tabel 4 volgt de conclusie (Schraven et al., 2010, p. 81): “Dit betekent voor de diagnostiek dat een leerkracht op basis van de meerkeuzetoets per woord bij gemiddeld ruim een kwart van de leerlingen een onjuist beeld krijgt van de spellingvaardigheid van een leerling”.

Ook al blijkt dat ook niet uit het Citoonderzoek, toch erkent het Cito (de Wijs, 2010b, p. 375) nu, dat bij de meerkeuzeopgaven het antwoord ook beïnvloed wordt door de set van daarin opgenomen vet gedrukte woorden en niet alleen door het herkennen van de fout in het fout gespelde woord. Het Cito erkent nu, dat “de *context* van de andere woorden in een meerkeuzeopgave *wel degelijk invloed* heeft op de moeilijkheid van de opgave” (de Wijs, 2010b, p. 375). Het Cito vindt dit echter “minder erg dan het lijkt” en “wil het beeld dat een dergelijke invloed zonder meer ongewenst is graag nuanceren” (de Wijs, 2010b, p. 375). Het Cito vindt het geen probleem, dat leerlingen een voor hen onbekend woord als fout gespeld aanstrepen, terwijl dat goed gespeld is, omdat “in teksten wel vaker woorden staan die leerlingen niet kennen” en opname van deze woorden in een toets nodig is om “de goede speller van de minder goede speller te onderscheiden”.

Met deze 'nuancering' geeft het Cito aan, dat zij het helemaal geen probleem vinden, dat een spellingtoets iets anders meet dan de spellingvaardigheid en problemen daarin. Die 'nuancering' houdt namelijk in, dat het Cito het geen probleem vindt, dat door de beïnvloeding van het antwoord door, onder andere, de andere vet gedrukte woorden, het antwoord op die meerkeuzeopgave geen goede indicator is voor de vaardigheid in het spellen van het betreffende toetswoord en de eventuele gebreken daarin. Kennelijk vindt het Cito het niet van belang, dat een spellingtoets valide is.

Ook heeft het Cito niet in de gaten, dat hun acceptatie van de invloed van de context van andere woorden op de moeilijkheid van de opgave in strijd is met hun bewering, dat “via een omgekeerde bewijsvoering” een incorrect antwoord “toch diagnostische mogelijkheden kent” (de Wijs, 2010b, p. 371). Als de onbekendheid met een correct gespeld woord de reden kan zijn van het ten onrechte niet aanstrepen van het onjuist gespelde woord, dan kan een onjuist antwoord niet geïnterpreteerd worden als aanwijzing, dat een leerling de spelling van het toetswoord niet kent.

Dat het Cito dit niet in de gaten heeft blijkt ook uit hun bespreking van de eerste opgave van module 'M4 Vervolg 2' (de Wijs, 2010b, p. 375-376). Door de Kofschiponderzoekers wordt niet ontkend, dat de leerlingen die terecht 'swak' als onjuist gespeld woord aanstrepen, de juiste spelling van dat woord kennen. Door hen wordt beweerd, dat uit de meerkeuzeopgave “ten onrechte wordt geconcludeerd dat de leerlingen de spelling van ZWAK niet beheersen” (Schraven et al., 2010, p. 81), omdat ook een groot deel van de leerlingen die dat woord

niet aanstreepten, de spelling daarvan toch blijkt te kennen.

Dat differentiatie in moeilijkheidsgraad tussen verschillende opgaven nodig is om goede en minder goede spellers van elkaar onderscheiden, staat buiten kijf. Dat is echter niet nodig “om betrouwbaar te kunnen meten”, zoals het Cito beweert (de Wijs, 2010b, p. 375). Die differentiatie in moeilijkheidsgraad dient echter niet veroorzaakt te worden door de woorden die tegelijkertijd met een toetswoord worden aangeboden maar door verschil in moeilijkheidsgraad tussen de aangeboden toetswoorden zelf.

Conclusie over het Cito

Nu het Cito erkent, dat

- de vaardigheid in het spellen (het omzetten van woorden in schriftbeelden (de Wijs et al., 2006, p. 9) niet getoetst kan worden door middel van meerkeuzeopgaven,
- meerkeuzeopgaven weinig diagnostische mogelijkheden bieden,
- het antwoord op een meerkeuzeopgave wordt beïnvloed door de andere vet gedrukte woorden in de opgave,

verbaast het, dat het Cito toch vast blijft houden aan meerkeuzeopgaven in de vervolgmodules van de nieuwe spellingtoetsen. Daaruit ontstaat de indruk, dat het Cito de meerkeuzeopgaven destijds om andere redenen heeft ingevoerd dan nu door het Cito (de Wijs, 2010a, 2010b) worden genoemd. Belangrijker is echter dat het belang van de validiteit van een toets door het Cito niet erkend wordt, getuige het feit dat:

- het Cito helemaal geen aandacht besteedt aan de validiteit van de nieuwe spellingtoetsen in (hun beschrijving van) de Cito-onderzoeken;
- het Cito niet in de gaten heeft, dat voor de geschiktheid van een toets de validiteit daarvan bepalend is en niet de betrouwbaarheid;
- het Cito niet in de gaten heeft, dat de betrouwbaarheid van een toets en de correlatie tussen scores op toetsen met verschillende opgavetypen niets zegt over de validiteit van een toets;
- het Cito het begrip 'validiteit' niet kent, want het Cito gebruikt ook steeds de term 'betrouwbaarheid' waar 'validiteit' wordt bedoeld (de Wijs, 2010b, p. 372, 373, 375);
- het Cito zich niet realiseert, dat een spellingtoets bij iedere leerling uit iedere basisschoolgroep waarvoor die bedoeld is, tot een correcte uitspraak over de spellingvaardigheid van die leerling dient te leiden en niet alleen maar gemiddeld bij een representatieve steekproef van leerlingen;
- het Cito niet in de gaten heeft, dat bij een onderzoek naar de validiteit van een spellingtoets het niet gaat om de vaststelling van de spellingvaardigheid van de leerlingen die aan het onderzoek deelnemen, maar om de vaststelling of de toets (met de meerkeuzeopgaven) het beoogde kenmerk (spellingvaardigheid) goed meet;
- het Cito het verantwoord vindt om in een onderzoek naar psychometrische kenmerken van een toets die toets ook voor te leggen aan leerlingen waarvoor die niet bedoeld is.

Het is eigenaardig, dat dit geconstateerd moet worden voor de instantie die het monopolie heeft op de ontwikkeling van onderwijstoetsen in Nederland.

Beoordeling door de COTAN

Als belangrijk argument om aan de meerkeuzeopgaven in de vervolgmodules van de nieuwe spellingtoetsen vast te houden verwijst het Cito naar het oordeel van de

COTAN (de Wijs, 2010b, p. 377). De COTAN heeft (pas) in 2010 een positief oordeel gegeven over de nieuwe Cito-spellingtoetsen, terwijl die al sinds 2006 op de markt zijn. Het is onduidelijk, welke toetsen door de COTAN zijn beoordeeld: de toetsen voor de groepen 3 t/m 6 (zoals vermeld in de titel van de publicatie waarnaar verwezen), of de toetsen voor de groepen 3 t/m 8 (zoals vermeld in de aanduiding van de toets in de COTAN-documentatie). De COTAN verwijst naar een versie uit 2010, terwijl bijna alle nieuwe spellingtoetsen eerder zijn gepubliceerd, in ieder geval de toetsen die bij dat jaartal vermeld staan (voor de groepen 3 t/m 6).

Het lijkt er dus op, dat de COTAN haar oordeel slechts heeft gebaseerd op het Cito-rapport uit 2010 met de "wetenschappelijke verantwoording van de toetsen Spelling" en de toetsen zelf en daarbij horende handleidingen en hulpboeken niet heeft bestudeerd. Dit zou ook kunnen verklaren, dat de COTAN niet in de gaten heeft gehad dat:

- de nieuwe spellingtoetsen in het kader van het LOVS geen geheel nieuwe toetsen zijn maar dienen ter vervanging van de bestaande SVS-toetspakketten;
- in het kader van die vervanging die toetsen niet alleen zijn geactualiseerd maar dat daarin dicteeopgaven door meerkeuzeopgaven zijn vervangen, zonder dat dat gemotiveerd wordt, en er dus geen sprake was van de introductie van een nieuwe toets die uit twee elementen bestaat;
- de bewering dat met dicteeopgaven beoogd wordt 'actieve spelling' te meten en met meerkeuzeopgaven 'passieve spelling' in strijd is met definitie van de vaardigheid die volgens de handleiding centraal staat bij deze toetsen;
- de bewering dat met dicteeopgaven beoogd wordt 'actieve spelling' te meten en met meerkeuzeopgaven 'passieve spelling' niet voorkomt in de handleiding en zelfs in strijd is met de bewering in de handleiding (de Wijs et al., 2006, p.11), dat "de vaardigheidsscores van leerlingen die een toets Spelling gemaakt hebben, altijd onderling vergeleken kunnen worden, ook al hebben de leerlingen niet allemaal dezelfde soort opgaven (woorddictee, zinsdictee, meerkeuzeopgave) gemaakt";
- de bewering dat met dicteeopgaven beoogd wordt 'actieve spelling' te meten en met meerkeuzeopgaven 'passieve spelling' niet in overeenstemming is met
    - de scoringsvoorschriften in de handleiding,
    - de aanvullende opgaven in het hulpboek,
    - het parallelle gebruik van dicteeopgaven en meerkeuzeopgaven in de vervolgmodules voor groep 4 en 5,
    - het gebruik van alleen meerkeuzeopgaven in de eindtoets.

In de toelichting op het oordeel over de validiteit plaatst de COTAN de volgende kritische kanttekeningen (Cotan, 2010):

- het is "verrassend", dat "geen gegevens op itemniveau worden verstrekt" en "er geen onderzoek naar bijvoorbeeld itembias is gedaan",
- "het is niet duidelijk welke aspecten van modelpassing zijn onderzocht en er worden geen empirische gegevens verstrekt",
- "in de handleiding wordt geen evidentie gegeven dat een methode-effect niet in het geding is",
- "in het onderzoek naar de validiteit wordt geen onderscheid gemaakt tussen de groepen die een verschillende vervolgtoets hebben gemaakt", "op dat

punt schiet het onderzoek naar de validiteit wat tekort".

Ondanks deze kritiekpunten beoordeelt de COTAN de validiteit van de nieuwe spellingtoetsen toch als "voldoende" (Cotan, 2010). Dit oordeel past dus niet bij deze kritiekpunten. Dit oordeel wordt niet ondersteund door de Cito-onderzoeken, want op basis daarvan kunnen geen uitspraken over de validiteit van de spellingtoetsen worden gedaan (zie hierboven). Dit oordeel is ook in strijd met de uitkomsten van het Kofschip-onderzoek (Schraven et al., 2010) (zie hierboven), dat door de COTAN ten onrechte terzijde wordt geschoven vanwege de samenstelling en omvang van de groep deelnemers, terwijl daar gezien de aard van het onderzoek niets mis mee is, en een leereffect, dat nooit opgetreden kan zijn.

Het is merkwaardig, dat de COTAN suggereert dat er onderzoek naar en onderzoekresultaten betreffende de validiteit van de spellingtoetsen beschikbaar is, terwijl het Cito zelf die in geen enkele publicatie over de spellingtoetsen presenteert.

De COTAN geeft een oordeel over het gehele pakket van nieuwe spellingtoetsen, inclusief de op dat moment nog niet beschikbare toetsen (voor de groepen 7 en 8), terwijl uit de reactie van het Cito blijkt, dat voor slechts 2 toetsen onderzoek naar psychometrische kenmerken is gedaan. Een verantwoording door de COTAN voor zo'n grove veralgemenisering van hun oordeel ontbreekt.

Daarnaast valt op, dat de COTAN de volgende punten niet signaleert:

- In de meerkeuzeopgaven van de module 'M4 Vervolg 2' zijn 16 (van de 100) woorden, waarvan 2 (van de 25) toetswoorden, opgenomen die niet tot het te meten spellingniveau behoren (Schraven et al., 2010, p. 83). Als de spellingtoetsen bedoeld zijn om het spellingniveau vast te stellen, dan dienen daarin alleen maar woorden opgenomen te zijn die tot dat niveau behoren, want anders valt de score niet meer te interpreteren (in een toets naar de kennis van het Duits horen geen Franse woorden). Daarnaast is het onbegrijpelijk dat de spaarzame ruimte gebruikt wordt om niet relevante woorden aan te bieden in plaats van relevante, zeker nu het Cito zelf klaagt over de geringe mogelijkheid om alle spellingcategorieën voldoende aan bod te laten komen in een toets van 25 opgaven (de Wijs et al., 2006, p. 376).
- De fout gespelde woorden worden niet in een dusdanig patroon in de opeenvolgende opgaven aangeboden dat antwoordgeneigdheden die niets te maken hebben met de te meten vaardigheid (bijvoorbeeld, een voorkeur voor een bepaalde positie), geen rol kunnen spelen in de uitkomst.
- De fout gespelde woorden zijn niet gelijk verdeeld over de antwoordposities: op positie A: 8; op positie B: 5; op positie C: 8; op positie D: 4.
- Binnen één blok van 4 opeenvolgende opgaven komen de toetswoorden niet slechts één keer op een positie voor; binnen opgaven 1 t/m 4 D 2 keer en B 0 keer, binnen opgaven 5 t/m 8 A 2 keer en B weer 0 keer, binnen opgaven 9 t/m 12 C 4 (sic!) keer en A en C 0 keer en B weer (!) 0 keer, binnen opgaven 13 t/m 16 A 2 keer en C 0 keer, binnen opgaven 17 t/m 20 B 2 keer en D 0 keer, binnen opgaven 21 t/m 25 (# 5) A en B 2 keer en D 0 keer.
- Tussen blokken van vier opeenvolgende opgaven worden volgorde-effecten niet ongedaan gemaakt.
- Binnen één opgave worden vetgedrukte woorden aangeboden die niet van eenzelfde spellingtechnisch type zijn (tot dezelfde spellingcategorie horen).

- Binnen één opgave zijn de aangeboden vetgedrukte woorden niet van dezelfde moeilijkheidsgraad, zoals nu ook wordt erkend door het Cito (de Wijs, 2010b, p. 375), terwijl door de bank genomen leerlingen in de beoogde doelgroep de vier vet gedrukte woorden binnen één opgave evengoed/slecht moeten kunnen spellen.

Nu de COTAN in de toelichting op zijn beoordeling, zeker ten aanzien van de validiteit, voornamelijk kritische kanttekeningen plaatst en daarnaast een aantal relevante gebreken van de Cito-spellingtoetsen niet signaleert, blijft onduidelijk, op welke gronden de COTAN zijn positieve oordeel over de Cito-spellingtoetsen heeft geveld.

Rol van de Onderwijsinspectie

De nieuwe Cito-spellingtoetsen worden sinds 2006 op de markt gebracht (de Wijs, 2010b, p. 370). Pas in 2008 en 2009 heeft het Cito onderzoek gedaan ter bepaling van de psychometrische kenmerken van die toetsen (de Wijs, 2010b, p. 373-374), twee tot drie jaar na de introductie. Dit onderzoek heeft alleen betrekking op de toetsen 'Spelling M4' en 'Spelling M8' (de Wijs, 2010b, p. 373-374). Dit onderzoek naar de psychometrische kenmerken van de nieuwe spellingtoetsen lijkt ook inhoudelijk beperkt van karakter. Het lijkt erop, dat de resultaten van dit onderzoek pas sinds 2010 bekend zijn (de Wijs, 2010b, p. 377). Pas in 2010 heeft de COTAN een oordeel over de nieuwe Cito-spellingtoetsen gepubliceerd.

Deze toetsen waren dus al vier jaar op de markt, zonder dat er ook maar iets bekend was over de deugdelijkheid daarvan. Ook op dit moment is de informatie daarover beperkt. Omdat van de uitkomsten op de Cito-toetsen zowel voor leerlingen als voor scholen veel afhangt, lijkt dit onverantwoord.

Het is onbegrijpelijk, dat de Onderwijsinspectie bij scholen aandringt op het gebruik van toetsen waarvan eerst niet duidelijk was of die wel valide zijn en waarvan nu duidelijk is dat dat niet het geval is. Als instantie die verantwoordelijk is voor het toezicht op de kwaliteit van het onderwijs in Nederland (en hoe die wordt vastgesteld) mag een kritischer houding ten aanzien van de nieuwe Cito-spellingtoetsen verwacht worden. Ook zonder het nu beschikbare onderzoek hadden bij de Onderwijsinspectie al twijfels over de geschiktheid van deze toetsen moeten rijzen. Kennelijk vindt de Onderwijsinspectie de zorgvuldigheid die noodzakelijk wordt geacht bij de introductie van geneesmiddelen, niet nodig bij zo bepalende schooltoetsen.

Tot slot

Onduidelijk is, waarop de COTAN zijn positieve oordeel over (de validiteit van) de nieuwe Cito-spellingtoetsen baseert. Duidelijk is wel, dat op basis van het Cito-onderzoek geen conclusies getrokken kunnen worden over de geschiktheid van de meerkeuzeopgaven voor de meting van de spellingvaardigheid van leerlingen en de problemen daarin, ondanks het groot aantal leerlingen die daaraan hebben deelgenomen.

Ondanks de beperkte omvang van het Kofschiponderzoek (Schraven et al., 2010) kan daaruit wel de conclusie getrokken worden, dat de stelling dat met de daarin opgenomen meerkeuzeopgaven de spellingvaardigheid van leerlingen en problemen daarin correct vastgesteld kunnen worden, niet juist is. De conclusie van Schraven et al. (2010, p. 84), dat “de meerkeuzetoets iets anders blijkt te meten dan het dictee” is dus terecht. Hun advies aan de scholen om te “besluiten om geen meerkeuzetoets meer toe te passen ter bepaling van de spellingvaardigheid” (Schraven et al., 2010, p. 85) is daarmee in overeenstemming.

Nu het Cito erkent, dat met de meerkeuzeopgaven iets anders gemeten wordt dan met de dicteeopgaven, dat de meerkeuze-

opgaven weinig diagnostische mogelijkheden bieden en dat bij de beantwoording van de meerkeuzeopgaven sprake is van een contexteffect, is het onbegrijpelijk, dat het Cito het advies van Schraven et al. (2010, p.85) om “de spellingherkenningstoets zoals opgenomen in de eindtoets te veranderen in een spellingproductietoets, zodat de vaardigheid die men beoogt te toetsen ook daadwerkelijk getoetst wordt” niet overneemt. De bewering van het Cito, dat er “met de kwaliteit van de toetsen Spelling niets mis is”, is in ieder geval niet terecht.

Geraadpleegde literatuur

Bosman, A.M.T., Schraven, J.L.M., & van Eekhout, T. (2010a). De nieuwe Cito-spellingtoets. *JSW, 94(10),* 6-9

Bosman, A.M.T., Schraven, J.L.M., & van Eekhout, T. (2010b). *Nogmaals ons bezwaar tegen de Cito-spellingtoets*. www.jsw-online.nl: nieuws

Cito (2010). *Spelling voor groep 3 tot en met 8*. www.cito.nl. Onderwijs: primair en speciaal onderwijs: alle producten: spelling. Arnhem: Cito.

Cito (2013). *Psychometrische kanttekeningen bij de onderzoeken van Bosman en Schraven*. http://www.cito.nl/nl

Cotan (2010). *Spelling groep 3 t/m 8 LOVS*. www.cotandocumentatie.nl.

Wijs, A. de (2010a). *Reactie Cito op artikel over toetsen Spelling*; www.jsw-online.nl: nieuws.

Wijs, A. de (2010b). Kritiek op toetsen Spelling steunt op losse gronden. Een reactie op het artikel 'De nieuwe Cito-spellingtoets ter discussie'. *Orthopedagogiek: Onderzoek en Praktijk, 49,* 370-377.

Wijs, A. de, Krom, R. & van Berkel, S. (2006). *LOVS Spelling groep 4*. Arnhem: Cito.

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