Het drijfzand van didactische leeftijdsequivalenten

Er bestaan ernstige bezwaren tegen het gebruik van verhoudingsnormen. Bij intelligentietests worden ze bijna niet meer toegepast. In het onderwijs en de diagnostiek van leerachterstanden daarentegen, heeft het gebruik van 'didactische leeftijdsequivalenten' (DLE's) de laatste decennia juist een grote vlucht genomen. Namens de Cotan brengen Evers en Resing tien ernstige bedenkingen naar voren bij het gebruik van DLE's.

Arne Evers en Wilma Resing

Schooltoetsen zijn niet meer weg te denken uit het huidige onderwijs en de onderwijsbegeleiding. Of het nu gaat om toetsafnames in het kader van een leerlingvolgsysteem, een toets aan het eind van het basisonderwijs, of een toets ten behoeve van indicatiestelling, elke leerling en leerkracht komt met het fenomeen 'toets' gedurende zijn of haar schoolloopbaan meermalen in aanraking. Het is uiteraard belangrijk dat goed geconstrueerde, betrouwbare, valide en goed genormeerde toetsen worden gebruikt, mede omdat het resultaat dat een leerling op een toets heeft behaald grote consequenties kan hebben. Ook het gebruik van een in principe goed geconstrueerde toets kan echter ongewenste gevolgen hebben. Zelfs wanneer het met de betrouwbaarheid en validiteit van een toets wel goed zit, kan de manier van interpretatie van de toetsscores leiden tot onjuiste conclusies en verkeerde beslissingen. Naar de mening van de Commissie Testaangelegenheden Nederland (COTAN) is die kans zeker aanwezig wanneer bij de interpretatie van test- en toetsscores gebruik wordt gemaakt van het zogenaamde didactisch leeftijdsequivalent (DLE), een werkwijze die frequent wordt gehanteerd in de diagnostische praktijk. In deze publicatie zal deze stellingname worden toegelicht.

Normsystemen

Wanneer de toets[1] die een leerling heeft gemaakt wordt gescoord, levert dit in eerste instantie een zogenaamde ruwe score op. Vaak zal dit de som zijn van het aantal goed gemaakte opgaven. In het algemeen biedt de ruwe score weinig interpretatiemogelijkheden en krijgt deze pas betekenis door vergelijking met een norm. Drenth en Sijtsma (2006) onderscheiden vier typen normen. Het eerste type betreft *absolute normen*. Hierbij wordt de norm direct afgeleid uit een omschrijving van het domein van vaardigheden of leerstof die men dient te beheersen. Dit type normgebruik wordt ook wel domeingerichte interpretatie genoemd. Bij domeingerichte normering heeft men geen gegevens van een referentie- of normgroep nodig, dat wil zeggen de score van een individu wordt geïnterpreteerd onafhankelijk van die van anderen. Voor het gebruik van dit soort normen dient daarentegen wel een goede analyse te zijn gemaakt van het proces dat men wil evalueren. Een voorbeeld is de vaste cesuur die is gekozen om te slagen voor het staatsexamen Nederlands als tweede taal (NT2) (Staatsexamencommissie, 2007). Dit is de score die is vastgesteld door experts en die overeenkomt met het vaardigheidsniveau dat minimaal noodzakelijk wordt geacht om een vruchtbare start te maken met een opleiding of beroep. Absolute normen zullen in dit artikel verder niet ter sprake komen. De andere drie typen normsystemen maken in tegenstelling tot absolute normen wel gebruik van een referentiegroep. Het betreft normen gebaseerd op rangordening, standaardscores en verhoudingsnormen. De bekendste vorm van *normen gebaseerd op rangordening* zijn percentielscores of percentielen. Maar ook bijvoorbeeld decielen, kwartielen en de niveauscores A tot en met E die door het Cito worden gebruikt, zijn van dit type. In feite geeft men, door gebruik van een dergelijke norm, de rang aan die de getoetste persoon in-

neemt binnen de referentiegroep. Zo geeft een percentielscore 20 aan dat een leerling het, op basis van de ruwe score die hij of zij heeft behaald, even goed of beter doet dan 20% van de referentiegroep, maar dat 80% een hogere score behaalt. Indien ruwe scores worden omgerekend in standaardscore-eenheden wordt gesproken van *standaardscores*. Hierbij wordt gebruik gemaakt van het gemiddelde en de spreiding van de scores in de referentiegroep (deze doen bij rangordenormen niet ter zake). Bij standaardscores wordt ervan uitgegaan dat de scores normaal zijn verdeeld en wordt in feite de afstand tot het midden van de verdeling aangegeven. Hierdoor wordt er rekening mee gehouden dat in het midden van de ruwe-scoreverdeling de scores meestal op een kluitje liggen en dat scores meer naar de extremen van de verdeling toe dunner zijn gezaaid. Omdat rangordescores, zoals percentielen, hier géén rekening mee houden, is het nadeel van dit type scores onder andere dat kleine scoreverschillen in het midden van de verdeling tot een overschatting van hun betekenis leiden en scoreverschillen aan het uiteinde van de verdeling tot een onderschatting. Dit is bij standaardscores niet het geval. Bekende voorbeelden van standaardscores zijn T-scores (gemiddelde 50, standaardafwijking 10), stanines (gemiddelde 5, standaardafwijking 1.96), C-scores (gemiddelde 5, standaardafwijking 2) en 'moderne' deviatie-IQ-schalen (gemiddelde 100, standaardafwijking 15). Bij standaardscores wordt in feite bepaald hoe ver de ruwe score af ligt van het gemiddelde van de eigen (leeftijds-, leerjaar-, niveau-) groep. Ook bij 'moderne' IQ-schalen is dit het geval; daarom wordt hiervoor ook wel de term deviatie-IQ gebruikt.

Bij *verhoudingsnormen* worden testscores gedeeld door een andere variabele, zoals leeftijd in het geval van de bekende 'mental age'-scores die vroeger bij bijvoorbeeld de Stanford-Binet tests werden gebruikt. Het bekendste voorbeeld van verhoudingsnormen is het 'ouderwetse' intelligentiequotiënt of IQ, waarbij de mentale leeftijd wordt gedeeld door de werkelijke leeftijd en wordt vermenigvuldigd met 100. De mentale leeftijd wordt bepaald door in een tabel op te zoeken voor welke leeftijd de behaalde ruwe score het meest kenmerkend is. Bij mentale-leeftijdscores wordt het kind derhalve vergeleken met kinderen van een (heel) andere leeftijd, terwijl bij gebruik van standaardscores naar de afwijking in scores van het kind ten opzichte van de eigen leeftijd wordt gekeken. Ook bij het 'ouderwetse' intelligentiequotiënt (en bij verhoudingsnormen in het algemeen) wordt dus gebruik gemaakt van een normgroep. De testscore wordt echter *niet* geïnterpreteerd in termen van de afwijking ten opzichte van het gemiddelde van de eigen leeftijdsgroep zoals bij standaardscores, maar ten opzichte van een leeftijdsschaal en de veronderstelde vaardigheid die bij elke leeftijdsschaal hoort. Omdat de mentale leeftijd vervolgens door de chronologische leeftijd wordt gedeeld, wordt hiervoor ook wel de term ratio-IQ gebruikt.

Reeds in 1990 stelden Oud en Mommers een aantal bezwaren met betrekking tot het gebruik en de bruikbaarheid van verhoudingsnormen aan de orde. In hun bespreking van het ratio-IQ concluderen zij: 'Vanwege de bezwaren die in dit artikel centraal staan, zijn sinds enkele decennia alle serieuze intelligentietests van Mentale Leeftijd en ratio-IQ overgestapt op een of andere vorm van standaardscores' (Oud & Mommers, 1990, p. 446). Nog wat verder achteromkijkend concluderen Drenth en Sijtsma (2006, p. 176) '[...] verhoudingsnormen hebben vooral historische betekenis [...]' Ook deze auteurs plaatsen vervolgens een aantal kritische kanttekeningen bij verhoudingsnormen in het algemeen en bij het klassieke IQ-begrip in het bijzonder.

De bijdragen van Moelands, Mommers en Oud (1990) en Oud en Mommers (1990) hadden niet alleen betrekking op het gebruik van de maten voor ratio-IQ en Mentale Leeftijd bij het gebruik van intelligentietests, maar ook op het DLE en het leerquotiënt, aangezien deze ook tot de categorie verhoudingsnormen behoren. Inmiddels lijkt het gebruik en de interpretatie van verhoudingsnormen bij intelligentietests teruggedrongen, zo niet uitgestorven te zijn. Dit geldt echter niet voor gebruik en interpretatie van toets- en toetsscores in het onderwijs en de diagnostiek van leerachterstanden in termen van DLE's. Integendeel, het gebruik van DLE's heeft de laatste decennia een grote vlucht genomen (Melis, 2006). Deze 'DLE-revival' heeft mede als oorzaak dat voor de leerlinggebonden financiering en voor de toelaatbaarheid van leerlingen tot het speciaal onderwijs een tijdlang rapportage van leerachterstanden in termen van DLE's gangbaar is geweest en zelfs van overheidswege verplicht is gesteld. Inmiddels is deze wijze van interpretatie echter losgelaten en wordt vergelijking met de prestaties van leeftijdgenoten vereist (Ministerie van O, C en W, 2006), de toelating tot het leerwegondersteunend en het praktijkonderwijs in het voortgezet onderwijs uitgezonderd, waarvoor OCW nog steeds rapportage in DLE's verplicht stelt (Ministerie van O, C en W, 2003). Ook in de alledaagse onderwijspraktijk worden DLE's nog frequent berekend en gebruikt om de individuele achterstand van een kind in een of meer leerdomeinen te beschrijven. De vraag is of voor deze toepassingen het gebruik van DLE's niet zou moeten worden afgeschaft, omdat de bezwaren tegen verhoudingsnormen onverkort ook gelden voor het gebruik van DLE's en leerquotiënten. In het vervolg van dit artikel zullen deze bezwaren uitgebreid worden besproken. In het navolgende zullen eerst de begrippen DLE en leerquotiënt worden toegelicht.

Didactische-leeftijdquivalentscores

De didactische leeftijd (DL) van een leerling is het aantal maanden onderwijs dat deze leerling vanaf begin groep 3 heeft gevolgd, waarbij elk leerjaar telt als tien maanden. Aan het begin van groep 3 heeft een leerling dus een DL van 0, aan het eind van groep 8 heeft een leerling – als deze niet is blijven zitten – een DL van 60. Het DLE wordt bepaald door in een tabel op te zoeken voor welke DL de ruwe toetsscore van een leerling het meest kenmerkend is. 'Kenmerkend' betekent in dit geval: dié didactische leeftijdgroep die de betreffende ruwe score als gemiddelde score heeft. Bijvoorbeeld: de DL van een leerling aan het eind van

groep 4 is 20. Deze leerling behaalt op een rekentoets een ruwe score van 25. Wanneer nu blijkt dat leerlingen aan het eind van groep 3 gemiddeld 25 sommen goed maken op deze toets, dan krijgt deze leerling een DLE van 10. In dit geval is het DLE kleiner dan het DL en het leerquotiënt (DLE/DL = 0.50) kleiner dan 1.00. Dit zou kunnen wijzen op leerachterstand. Wanneer een leerling gedurende zijn hele schoolloopbaan scores haalt die precies overeenkomen met het gemiddelde van leerlingen met eenzelfde didactische leeftijd, dan zal zijn DLE steeds gelijk zijn aan zijn DL en is het leerquotiënt steeds gelijk aan 1.00. Loopt een leerling achter bij het gemiddelde van zijn groep dan is zijn DLE kleiner dan zijn/haar DL en het leerquotiënt kleiner dan 1.00; behaalt een leerling een hogere score dan het gemiddelde van zijn groep dan is het omgekeerde het geval. De gelijkenis met Mentale Leeftijd of ratio-IQ's is groot: DLE kan worden vervangen door mentale leeftijd en DL door werkelijke leeftijd. Soms wordt ook de vermenigvuldigingsfactor van 100 meegenomen.

Om DLE's te kunnen berekenen, is het nodig te beschikken over de gemiddelde ruwe scores op een toets die door de verschillende didactische leeftijdgroepen worden behaald, waarbij het een voorwaarde is dat de toets landelijk is genormeerd (overigens geldt de voorwaarde van landelijke normering ook voor vrijwel alle toepassingen van toetsen die met rangordenormen of met standaardscores werken). Het liefst zou men voor elke maand onderwijs een dergelijke normgroep ter beschikking hebben, maar dit is in de praktijk niet haalbaar. De minimale voorwaarde om DLE's te kunnen berekenen is dat van ten minste twee groepen leerlingen met verschillende didactische leeftijden normgegevens beschikbaar zijn. Men neemt aan dat vervolgens via interpolatie de ruwe scores van de tussenliggende didactische leeftijden kunnen worden geschat. Hierbij wordt uitgegaan van een *lineaire groeisnelheid* voor de verwerving van vaardigheden die de toets beoogt te meten voor de gemiddelde leerling in de tussenliggende periode. Wanneer van bovengenoemde rekentoets aan het eind van groep 3 het gemiddelde 25 is en aan het eind van groep 4 een gemiddelde van 40 wordt behaald, dan wordt dus aangenomen dat de gemiddelde leerling elke maand 1.5 punt beter wordt (het verschil van 40 en 25 gedeeld door 10). Een ruwe score van bijvoorbeeld 31 levert in dit geval een DLE van 14 op (25 + 4 × 1.5, is gelijk aan DLE 10 + 4 maanden). Soms wordt ook geëxtrapoleerd. Dit betekent dat DLE's worden berekend voor groepen die buiten de leerjaren vallen waar men gegevens voor heeft. In het bovenstaande voorbeeld kan dat betekenen dat men de DLE-tabel doortrekt naar eind groep 5 (of zelfs tot eind groep 8), onder de onderstelling dat de toename in score ook na eind groep 4 steeds 1.5 punt per maand zal bedragen. Een voorbeeld is de toets Begrijpend Lezen 678 (Aarnoutse & Kapinga, 2006) die is afgenomen in de groepen 6, 7 en 8. Bij deze toets lopen de DLE's echter naar beneden door tot 15 (de score 15 staat gelijk aan het niveau medio groep 4 volgens de DLE-idee).

Waarom is het gebruik van DLE-scores populair? Men kan er de leerachterstand in maanden mee aangeven, zo is het idee erachter, en deze gekwantificeerde leerachterstand zou goed bruikbaar en inzichtelijk kunnen zijn voor het onderwijsveld. Een kind zou leerstof aangeboden kunnen krijgen op het goede didactische niveau. Ook zou in het kader van leerlingvolgsystemen het individuele leerrendement aan de hand van DLE-scores in kaart kunnen worden gebracht (Melis, 2006). Bovendien zouden DLE's het gemakkelijk maken om de beheersing op verschillende leerstofgebieden met elkaar te vergelijken. Er kleven echter ernstige bezwaren aan het DLE-concept. Deze zullen hieronder worden besproken.

Bezwaren tegen DLE's

Bij dit overzicht van bezwaren tegen DLE-scores is gebruik gemaakt van de volgende publicaties: Angoff (1971), LVS-nieuws (Cito, juni 2001), Drenth & Sijtsma (2006), Moelands, Mommers & Oud (1990), Oud & Mommers (1990) en Stanley (1964). Een aantal bezwaren is ook al vermeld in Evers et al. (2002) en Resing et al. (2005). In totaal zullen tien verschillende problemen met DLE's aan de orde worden gesteld. Deze kritiekpunten en de consequenties voor de praktijk worden vervolgens samengevat in tabelvorm (zie Tabel 1).

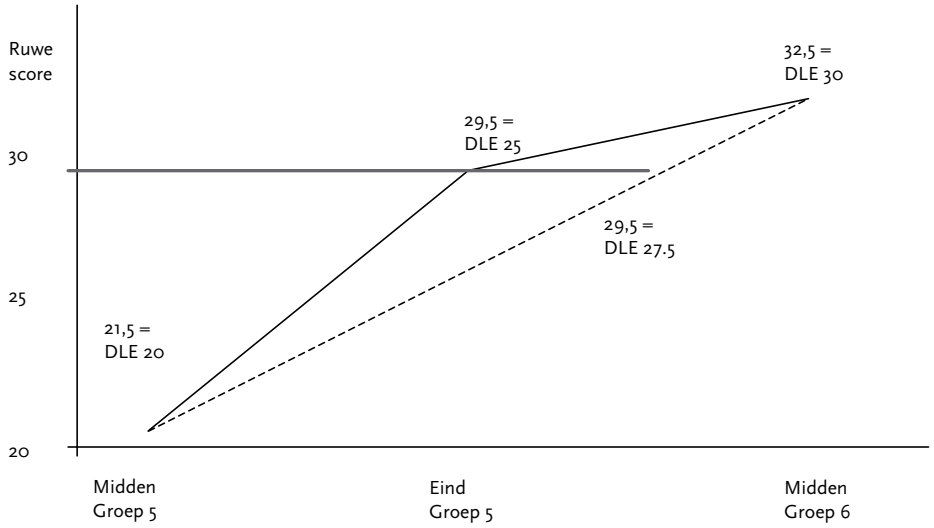
Een eerste probleem is dat men aanneemt dat scores op een toets in de tijd lineair zullen toenemen en voor alle leerlingen eenzelfde verloop zullen hebben. De onderstelling van een lineaire groeisnelheid is een sterke onderstelling die in de praktijk vaak niet houdbaar blijkt. Kinderen leren niet volgens een regelmatige stabiele curve en zeker niet volgens een strikt lineair patroon. Dat houdt in dat de DLE's die op grond van interpolaties (en eventueel extrapolaties) worden berekend, aanzienlijk kunnen afwijken van de ruwe scores die werkelijk bij een bepaalde didactische leeftijd horen. Een voorbeeld hiervan zijn de scores op de Toetsen Begrijpend Lezen (Staphorsius & Krom, 1998). Op deze toetsen bedraagt de gemiddelde ruwe score Midden groep 5 21.5 (didactische leeftijd 25 maanden), de gemiddelde score Eind groep 5 is 29.5 (didactische leeftijd 30 maanden) en de gemiddelde score Midden groep 6 is 32.5 (didactische leeftijd 35 maanden). Bij de eerste 5 maanden is er een scoreverschil van 8 punten, bij de laatste 5 maanden is er een scoreverschil van slechts 3 punten. Wanneer men slechts één meetpunt per jaar zou hebben, zou dit totaal van 11 scorepunten gelijkmatig over het hele jaar worden uitgesmeerd met over deze hele periode van 12 maanden een overschatting van leerprestaties als gevolg. De maximale overschatting bedraagt in dit voorbeeld 2.5 scorepunten, hetgeen overeenkomt met bijna drie maanden (zie Figuur 1)! Echter, ook binnen een ontwikkelingsspanne van een half jaar zijn dit soort groepsfluctuaties – afgezien van individuele verschillen in leercurves – nog mogelijk. In het algemeen geldt: hoe verder de meetpunten uit elkaar liggen, des te groter de kans op fouten.

Een tweede, verwant probleem heeft te maken met bodem- en plafondeffecten. Deze effecten komen bij veel toetsen voor. Vaak zal een toets voor de jongste leerlingen

Bezwaar categorie	Feitelijk probleem	Consequentie
Onjuiste fundamentele onderstelling (1) (10) (7)	Onterechte onderstelling van lineaire toename	Willekeurige over- of onderschatting
	Dezelfde DLE's van leerlingen op verschillende leeftijden zijn onderling niet vergelijkbaar	Het aanbieden van dezelfde leerstof en begeleiding leidt niet tot verbetering
	DLE-score ten onrechte gelijkstellen aan aantal maanden leerachterstand	Leerachterstand in aantal maanden wordt te letterlijk genomen
Tekortkoming van de toets in relatie tot DLE's (2) (8) (9)	Geen rekening houden met bodem- en/of plafondeffecten	Onderschatting of overschatting van DLE's afhankelijk van niveau van de leerling
	Onzuiverheid van normgroepen	Vertekening van scores
	Toetsen bevatten te weinig items in verhouding tot aantal DLE-klassen	Verschillen tussen leerlingen worden overschat
Statistisch probleem inherent aan DLE's (3) (6)	Ontbreken van standaardmeetfout in relatie tot kleine scoreverschillen	Te veel waarde hechten aan leerachterstand van enkele maanden
	Spreiding is op verschillende leeftijden ongelijk	Veranderingen in de mate van achterblijven of in de progressie van leerlingen worden foutief ingeschat
Uitwas (4) (5)	Onterecht extrapoleren buiten meetpunten	Willekeurige over- of onderschatting, waarbij het risico van grote fouten bestaat
	Rapportage van onmogelijke DLE-scores (scores buiten DLE-range 0 – 60)	In tegenspraak met uitgangspunten van het DLE-concept zelf, risico van grote fouten

Noot: De getallen tussen haakjes verwijzen naar de puntsgewijze bespreking in de tekst.

Tabel 1. Categorisering van de bezwaren tegen DLE's inclusief de praktische consequenties



Figuur 1. Geïnterpoleerde ruwe scores voor verschillende didactische leeftijden bij gebruik van drie meetpunten (onderbroken lijn) en twee meetpunten (stippellijn).

iets te moeilijk zijn en/of voor de oudste leerlingen iets te gemakkelijk. Dit betekent dat de scoreverdeling aan de onder- en bovenkant afvlakt en er dus geen lineaire interpolatie kan plaatsvinden. De DLE's van leerlingen met een lage respectievelijk hoge vaardigheid zullen in dat geval in aanzienlijke mate onder- respectievelijk overschat worden. Dit is des te ernstiger omdat diagnostische besluitvorming vooral plaatsvindt bij kinderen met een laag vaardigheidsniveau.

Ten derde, bodem- en plafondeffecten kunnen er ook de oorzaak van zijn dat de gemiddelde scores van normgroepen op twee opeenvolgende metingen relatief weinig van elkaar verschillen. Kleine scoreverschillen hoeven zich echter niet noodzakelijk tot de uiteinden van de verdeling te beperken. Wanneer dergelijke kleine groepsverschillen voorkomen, zullen ook kleine verschillen tussen leerlingen leiden tot aanzienlijke verschillen in DLE's. Zit er bijvoorbeeld tussen twee afnamemomenten één jaar, terwijl de gemiddelde scores voor die twee afnamemomenten maar twee punten verschillen, dan zullen leerlingen met slechts twee punten verschil in ruwe scores tien maanden verschil in DLE vertonen. Een verschil van twee scorepunten is echter nietszeggend, omdat de ruwe score op een toets wordt vertekend door meetfouten. Dergelijke kleine scoreverschillen liggen in het algemeen royaal binnen de standaardmeetfout van de toets (bijvoorbeeld bij een betrouwbaarheid van .84 en een standaardafwijking van 10 is de standaardmeetfout van de ruwe scores gelijk aan 4). Toetsconstructeurs rapporteren in het algemeen geen standaardmeetfouten voor DLE-scores. Mede daardoor gaan gebruikers er, ten onrechte, van uit dat DLE-scores geen standaardmeetfout zouden hebben en interpreteren ze deze scores dienten-

gevolge als absolute waarheden. Het berekenen van standaardmeetfouten voor DLE's komt de auteurs overigens voor als een statistisch onmogelijke opgave. Voor de berekening van de standaardmeetfout gaat men immers uit van de standaardafwijking en de betrouwbaarheid bij één normgroep. DLE's komen tot stand door interpolatie op basis van gegevens van minstens twee normgroepen. Vanuit de standaardmeetfout van de ruwe scores is echter wel een schatting te maken. Als voorbeeld is het onderdeel Technisch Lezen uit het Drempelonderzoek (Kapinga, 2006) genomen. De standaardmeetfout ligt bij deze toets in groep 6 en groep 7 rond de 5. Het verschil in gemiddeldes tussen groep 6 en groep 7 is ongeveer 10 punten, dat wil zeggen precies 1 ruwe-scorepunt per maand. De onbetrouwbaarheid van de meting in termen van DLE's (gedefinieerd als het 68%-betrouwbaarheidsinterval) bedraagt hier derhalve ongeveer plus of min 5 maanden (en het 90%-interval zelfs +/- 8 maanden).

Ten vierde, worden bij extrapolaties scores berekend voor (didactische) leeftijden die vóór het eerste en/of ná het laatste meetmoment liggen. Omdat voor deze leeftijden feitelijk geen gegevens voorhanden zijn (bij interpolaties heeft men ten minste nog een vast begin- en eindpunt), is dit een sprong in het duister. Daarbij wordt de claim van lineairiteit bij extrapolatie extra dubieus, omdat extrapolaties aan het begin en/of het eind van het basisonderwijs zullen plaatsvinden en de kans op bovengenoemde bodem- en plafondeffecten juist daar het grootst is.

Een vijfde probleem is dat DLE's vooral zijn ontworpen om leerachterstand in kaart te brengen, waardoor de differentiatiemogelijkheid aan de bovenkant van de schaal beperkt is. Hoger scoren dan de gemiddelde leerling aan het eind van groep 8 kan niet in DLE's worden uitgedrukt. Sommige toetsauteurs hebben daar iets op gevonden door de DLE-schaal boven de 60 door te laten lopen. Zo loopt de DLE-schaal in het onderdeel Technisch Lezen van de Toets voor Plaatsing in het Voortgezet Onderwijs (A-VISION, 2006) door tot DLE 70. Het betreft in dit geval een extrapolatie, want gegevens van leerlingen in het eerste jaar van het voortgezet onderwijs worden niet vermeld. Verder behoort technisch lezen niet tot het lesprogramma van het voortgezet onderwijs. Dus zelfs als deze gegevens wel verzameld zouden zijn, dan zouden deze leerlingen geen tien maanden extra les in technisch lezen hebben gehad waardoor het lastig zou worden hun DL te bepalen. Zou men nu werkelijk menen dat een leerling eind groep 8 die een DLE van 70 behaalt een leervoorsprong van tien maanden heeft? Gaat het niveau van technische leesvaardigheid nog zo omhoog in het voortgezet onderwijs? Het omgekeerde gebeurt in het onderdeel Inzichtelijk Rekenen Internetversie (Verweij & Lubbers, 2006) van dezelfde testserie waarbij ruwe scores van vmbo-leerlingen worden teruggerekend naar een DLE-schaal van 0 tot 60. In dit geval is het volstrekt onduidelijk welke DL zou moeten worden gehanteerd en bovendien is het onjuist om scores van leerlingen in het voortgezet onderwijs met normgegevens van leerlingen in het basisonderwijs te vergelijken. Ook komt het voor dat negatieve DLE-scores (lopend tot -20) worden gerapporteerd. Zie bijvoorbeeld Melis (2006) waarin voor tests uit de Cito-serie voor kleuters negatieve DLE's worden gerapporteerd. Dit mogen uitwassen lijken, het illustreert duidelijk hoe onverantwoord er in de alledaagse onderwijs- en toetspraktijk met DLE's wordt omgesprongen.

Een zesde probleem van DLE's is de onderstelling dat de spreiding van de scores binnen de normgroepen op verschillende leeftijden gelijk is. Ook deze onderstelling blijkt in de praktijk zelden houdbaar, meestal wordt de spreiding van groep 3 tot groep 8 groter omdat de verschillen in vaardigheden tussen leerlingen toenemen. Leerlingen die beneden het gemiddelde presteren en steeds precies dezelfde positie in de normgroep innemen, lijken dan volgens hun DLE's achteruit te gaan. Voor boven het gemiddelde presterende leerlingen geldt het omgekeerde. Maar ook voor leerlingen waarvan de relatieve positie in de groep verandert, leveren DLE's vertekening van leerprestaties op. Dit wordt door Geelhoed en Reitsma (1999) geïllustreerd met behulp van de scores van een leerling op het PI-dictee. Deze leerling maakte het PI-dictee achtereenvolgens halverwege groep 3, 4, 5 en 6. Zijn ruwe scores zijn respectievelijk 6, 42, 70 en 92 en deze komen overeen met percentielscores van respectievelijk 5, 15, 21 en 28 binnen de verschillende jaargroepen. Deze leerling is dus zwak begonnen, maar zijn prestaties trekken geleidelijk steeds meer bij en hij weet zijn relatieve positie steeds te verbeteren. Zijn DL op de vier afnamemomenten is respectievelijk 5, 15, 25 en 35 en zijn DLE respectievelijk 1, 11, 21 en 31. Deze leerling heeft dus steeds een 'leerachterstand' van vier maanden en in termen van DLE's lijkt het alsof er geen verbetering van leerprestaties heeft plaatsgevonden, terwijl deze leerling juist een opmerkelijk positieve prestatie heeft geleverd.

Een zevende probleem is de misvatting dat een leerling bij een leerachterstand van bijvoorbeeld vier maanden dan ook vier maanden extra leertijd nodig zou hebben om deze achterstand in te halen. DLE's worden berekend op basis van groepsgemiddelden en laten geen interpretatie in benodigde leertijd voor individuele kinderen toe. Zo kan het voorkomen dat een leerling die voorafgaand aan de toets een belangrijke les heeft gemist, een leerachterstand heeft van vier maanden, maar deze na het inhalen van de les weer vrij gemakkelijk inloopt (Sanders & Sluijter, 2002).

Een achtste probleem heeft te maken met de zuiverheid van de normgroepen die worden gebruikt. Elke normgroep bij een toets dient te bestaan uit kinderen met eenzelfde didactische leeftijd. Scores van doubleurs moeten worden uitgesloten alvorens de gemiddelde score te bepalen. Dat gebeurt in de praktijk zelden of nooit. De rechtstreeks bepaalde DLE's kunnen dus al min of meer vertekend zijn.

Een negende, praktisch, probleem dat bij veel toetsen voorkomt, is het feit dat toetsen over het algemeen te weinig items bevatten om alle 60 DLE-klassen te kunnen vullen. Een toets zou dus minimaal uit 60 items moeten bestaan om aan elke score een DLE-waarde te kunnen koppelen. Aangezien de groep met didactische leeftijd 1 echter niet de gemiddelde ruwe score 1 zal halen maar een hogere score, en de groep met didactische leeftijd 60 niet de ge-

middelde ruwe score 60 zal halen maar een lagere score (op een toets van 60 items) is de effectieve scorerange ook in dat geval een stuk kleiner dan 60. Ook zouden liefst meerdere scores tot hetzelfde DLE moeten leiden, anders heeft men bij één item extra fout meteen een maand leerachterstand. Maar zelfs een aantal van 120 items, zoals bij het onderdeel Technisch Lezen van de Testserie voor Plaatsing in het Voortgezet Onderwijs (A-VISION, 2006) is niet genoeg om alle DLE-klassen te vullen. Dit betekent dat een verschil van één scorepunt (één item meer of minder goed) in veel gevallen een DLE-verschil van twee maanden oplevert. Overigens geldt dit probleem ook voor rangordescores en standaardscores die uit veel klassen bestaan, zoals percentielen en T-scores.

Het tiende en laatste probleem dat hier wordt besproken heeft te maken met de inhoud van de leerstof die als basis dient voor de toetsinhoud. Ten eerste moet men er bij de interpretatie op bedacht zijn dat DLE's in verschillende leerjaren wat betreft inhoud van de leerstof niet echt vergelijkbaar zijn, wanneer van verschillende toetsen gebruik wordt gemaakt. Een DLE van 25 behaald op bijvoorbeeld een spellingtoets voor groep 3, betekent iets anders dan een DLE van 25 behaald op een toets voor groep 5, omdat de toets op andere leerstof betrekking heeft. Bovendien kan men zich sowieso afvragen of de prestaties van deze twee leerlingen wel vergelijkbaar zijn en beide niet een geheel andere aanpak behoeven. Een tweede punt is dat scholen kunnen verschillen in de tijdstippen waarop ze bepaalde leerstof aanbieden en hoe ze die aanbieden. Ten slotte, gaat men er bij het gebruik van DLE's van uit dat de leerstofinhoud in de tijd gelijk of vergelijkbaar blijft. De inhoud van het onderwijs en de onderwijsmethoden kunnen echter veranderen waardoor in het verleden bepaalde DLE's een vertekend beeld geven of niet meer van toepassing zijn.

Conclusie

Oud en Mommers (1990, p. 446) stelden reeds meer dan vijftien jaar geleden met klem met betrekking tot IQ-meting en, analoog daaraan met betrekking tot toetsgebruik in het algemeen '[...] Mentale Leeftijd en ratio-IQ zijn ongeschikt om uitspraken betreffende de achterstand of voorsprong ten opzichte van andere personen te doen.' Naar aanleiding van bovenstaande tien punten van statistische, instrumentele en inhoudelijke kritiek kan ook onze conclusie met betrekking tot toetsgebruik waarbij interpretatie met behulp van DLE's aan de orde is luiden: *DLE en leerquotiënt zijn ongeschikt om uitspraken te doen betreffende de achterstand of voorsprong ten opzichte van andere leerlingen*. Bij het gebruik van DLE's is sprake van schijnexactheid. Hoewel het gebruik van DLE's voor leerkrachten en diagnostici inzichtelijk is en naar ouders gemakkelijk communiceerbare kwantitatieve gegevens oplevert, kunnen deze gegevens leiden tot uitspraken en beslissingen die niet te rechtvaardigen zijn. Leerlingen die in werkelijkheid niet in vaardigheid verschillen, kunnen grote verschillen in DLE vertonen. Dit geldt zowel voor het herhaald gebruik in leerlingvolgsystemen als bij momentopnamen bijvoorbeeld ten behoeve van de indicatiestelling voor het speciaal onderwijs. DLE's blijken drijfzand; oppervlakkig ziet het er prachtig uit, maar de ondergrond is uitermate onbetrouwbaar.

Wat is het alternatief? Leerachterstand kan, net als andere meetinhouden van tests of toetsen (bijvoorbeeld intelligentie of een ernstige stoornis), het best worden uitgedrukt in een maat waarin de relatieve positie *ten opzichte van didactische leeftijdsgenoten en/of klasgenoten* duidelijk wordt, dat wil zeggen in standaardscores of in rangordescores. Dit wil zeggen dat scores behaald door een leerling vergeleken dienen te worden met de normgegevens van de leerjaargroep waarin hij gezien zijn leeftijd thuis zou horen. Voor leerlingvolgsystemen is het aan te bevelen de toets(en) te baseren op de item-responstheorie (IRT), maar dat is niet per se noodzakelijk.

Natuurlijk dienen ook normeringen op grond waarvan standaardscores en rangordescores zijn bepaald aan specifieke eisen te voldoen. Er dient sprake te zijn van landelijke steekproeven, standaardmeetfouten dienen te worden vermeld en de range van normscores dient in ieder geval niet groter te zijn dan de range van ruwe scores. Deze typen scores kennen echter veel minder problemen dan DLE's omdat er sprake is van een fundamenteel ander uitgangspunt. Een probleem dat standaardscores en rangordescores net als DLE's overigens wel kennen is dat de toets in de normgroep meestal in bepaalde periodes van het jaar is afgenomen. Leerlingen die in een andere periode van het jaar worden getest, kunnen niet goed met deze normgroep worden vergeleken. Tegenwoordig zijn hiervoor normeringsprogramma's beschikbaar die gebruik maken van fit-procedures (zie bijvoorbeeld Snijders, Tellegen & Laros, 1988) of een zogenaamde leerdagcorrectie toepassen (zoals bij de Nederlandse Differentiatie Testserie; Van Hoorn, Van der Kamp & Den Brinker, 2004), waarmee een, statistisch, alternatief wordt geboden voor de lineaire interpolaties zoals die bij DLE's worden toegepast. Een bezwaar dat door DLE-gebruikers frequent wordt geuit tegen de niveauscores van het Cito is, dat de onderste klasse (klasse E, dat wil zeggen de laagst scorende 10%) te breed is en te weinig differentieert, bijvoorbeeld tussen leerlingen die in principe thuishoren op het praktijkonderwijs en leerlingen die in aanmerking komen voor leerwegondersteunend onderwijs. Dit probleem is eenvoudig op te lossen door bijvoorbeeld gebruik te maken van stanines (daarbij wordt onderscheid gemaakt tussen de 4% laagst scorenden – stanine 1 – en de 6.5% die daarboven scoort – stanine 2), C-scores (daarbij wordt ook de onderste 4% nog in tweeën gedeeld) of de E-niveau klasse te splitsen en een F-klasse toe te voegen of alle percentielscores tussen 1 en 10 te vermelden. Echter, ook in dit laatste geval dient men te waken voor schijnexactheid.

Consequenties voor de COTAN-beoordeling van normen

De COTAN probeert verantwoord testgebruik te bevorderen door het verstrekken van informatie over de kwaliteit van

tests. Door middel van dit artikel wil de COTAN nogmaals stelling nemen tegen het onverantwoord maar onverminderd voortgaand gebruik van DLE's. De COTAN heeft reeds in een eerder stadium in haar beoordelingssysteem het criterium opgenomen dat toetsen die uitsluitend in termen van DLE's rapporteren een 'onvoldoende' krijgen voor normen. Dit laat echter onverlet dat gebruik en interpretatie van DLE's in psychologische rapportages maar ook in diverse onderwijsgremia nog steeds aan de orde van de dag, soms zelfs voorgeschreven, is. Weliswaar worden nu bij de meeste toetsen die DLE-georiënteerd zijn, al dan niet in een bijlage, ook standaardscores of rangordescores gerapporteerd, maar de algehele interpretatievoorschriften in de handleidingen bij sommige van deze toetsen blijven dermate eenzijdig op DLE's gericht dat de gebruiker nauwelijks keus lijkt te hebben. Naar de mening van de COTAN dient rapportage in standaardscores of rangordescores voorop te staan, vergezeld van een gedegen uitleg van deze systemen ten behoeve van de gebruikers. Indien vervolgens ook DLE's worden gerapporteerd – het zal duidelijk zijn dat de COTAN rapportage in termen van DLE's het liefst geheel ziet verdwijnen – dient de gebruiker in elk geval in de handleiding uitdrukkelijk voor de beperkingen ervan te worden gewaarschuwd. Toetsen waarbij dit niet het geval is, zullen met een 'onvoldoende' voor normen worden beoordeeld teneinde de gebruiker te waarschuwen dat gebruik van de betreffende toets tot onverantwoorde beslissingen voor de leerling kan leiden.

Dr. A. Evers is werkzaam als universitair hoofddocent bij de Programmagroep Arbeids- & Organisatiepsychologie van de Universiteit van Amsterdam, Roetersstraat 15, 1018 WB Amsterdam. E-mail: <a.v.a.m.evers@uva.nl>.

Mw prof.dr. W.C.M. Resing is werkzaam als hoogleraar bij de Sectie Ontwikkelings- en Onderwijspsychologie van de Universiteit Leiden, Wassenaarseweg 52, 2333 AK Leiden. E-mail: <resing@fsw.leidenuniv.nl>.

Noten

De in deze publicatie gekozen stellingname is te beschouwen als het formele standpunt van de Commissie Test Aangelegenheden Nederland (COTAN); beide auteurs zijn lid van de COTAN. Hierbij danken de auteurs alle andere COTAN-leden voor hun waardevolle en constructieve commentaar op een eerdere versie van dit artikel.

1. Voor het gemak wordt in deze publicatie steeds van een toets gesproken, maar hier kan ook elke andere testvorm worden gelezen.

Literatuur

Aarnoutse, C. & Kapinga, T. (2006). *Begrijpend lezen 345678*. Ridderkerk: 678 Onderwijs Advisering.

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (p. 508-600). Washington, DC: American Council on Education.

A-VISION (2006). *Testserie voor Plaatsing in het Voortgezet Onderwijs. Technisch Lezen (TPVO-TL). Handleiding*. Ugchelen: A-VISION.

Cito groep (2001). DLE, hoe zit het daarmee? *Lvs Nieuws*, *7* (5). Arnhem: Cito.

Drenth, P.J.D. & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* (4de, herziene druk). Houten: Bohn Stafleu van Loghum.

Evers, A., Vliet-Mulder, J.C. van, Resing, W.C.M., Starren, J.C.M.G. & Boxtel, H. van. (2002). *COTAN Testboek voor het onderwijs*. Amsterdam: NDC/Boom.

Geelhoed, J. & Reitsma, P. (1999). *PI-dictee handleiding*. Lisse: Swets & Zeitlinger.

Hoorn, W. van, Kamp, L. van der & Brinker, W. den. (2004). *Nederlandse Differentiatie Testserie. Handleiding*. Amsterdam: Harcourt Test Publishers.

Kapinga, T. J. (2006). *Drempelonderzoek. Didactische plaatsbepaling binnen VMBO en praktijkonderwijs*. Ridderkerk: 678 Onderwijs Advisering.

Melis, G. (2006). *DLE boek*. Amsterdam: Boom.

Ministerie van Onderwijs, Cultuur en Wetenschap (2003). Besluit van 27 mei 2003, houdende regels inzake regionale verwijzingscommissies, een regionaal zorgbudget en praktijkscholen met declaratiebekostiging in het voortgezet onderwijs en houdende wijzigingen van besluiten in verband met onder meer de bekostiging van leerwegondersteunend en praktijkonderwijs. *Staatsblad van het Koninkrijk der Nederlanden*, *262*, 1-53.

Ministerie van Onderwijs, Cultuur en Wetenschap (2006). Besluit van 23 juni 2006, houdende wijziging van het Besluit leerlinggebonden financiering in verband met de vaststelling van criteria voor toelaatbaarheid van leerlingen tot het speciaal onderwijs. *Staatsblad van het Koninkrijk der Nederlanden*, *327*, 1-22.

Moelands, A.H.J., Mommers, M.J.C. & Oud, J.H.L. (1990). Leerlingvolgsystemen verklaard en vergeleken. *School en Begeleiding*, *26*, 19-28.

Oud, J.H.L. & Mommers, M.J.C. (1990). De valkuil van het didactisch leeftijdsequivalent. *Tijdschrift voor Orthopedagogiek*, *29*, 445-459.

Resing, W.C.M., Evers, A., Koomen, H.M.Y., Pameijer, N.K. & Bleichrodt, N. (2005). *Indicatiestelling speciaal onderwijs en leerlinggebonden financiering. Condities en instrumentarium*. Amsterdam: Boom.

Sanders, P. & Sluijter, C. (2002, januari). *DLE's, validiteit, beoordeling CAT's*. Presentatie COTAN, Utrecht, 21-01-2002.

Snijders, J.Th., Tellegen, P.J. & Laros, J. A. (1988). *Snijders-Oomen Niet-verbale Intelligentietest SON-R 5.5-17. Verantwoording en Handleiding*. Groningen: Wolters-Noordhoff.

Staatsexamencommissie (2007). *Staatsexamen NT2*. Gevonden op 11 april 2007 op http://www.expertisecentrumnt2.nl/staat/

Stanley, J.C. (1964). *Measurement in today's school*. Englewood Cliffs, NY: Prentice Hall.

Staphorsius, G. & Krom, R. (1998). *Toetsen Begrijpend Lezen. Handleiding*. Arnhem: Cito.

Verweij, A.C. & Lubbers, W.J. (2006). *Testserie voor Plaatsing in het Voortgezet Onderwijs. Inzichtelijk Rekenen. Internetversie. Handleiding*. Ugchelen: A-VISION.

summary

The quicksand of grade equivalent scores

A. Evers, W.C.M. Resing

In the Netherlands grade equivalent scores go through a remarkable revival. This revival is at least partly dependent on governmental policy, as for the allowance of funds for special education assessment of educational retardiness in terms of grade equivalent scores is required. The authors discuss ten disadvantages of this type of scores. Their conclusion is that grade equivalent scores and educational quotients are unsuitable for the assessment of educational retardedness. Use of these scores can possibly lead to non-justifiable educational decisions and recommendations. Based on this conclusion, the Dutch Committee on Testing (COTAN) has decided to rate the quality of the norms of tests that primarily make use of grade equivalent scores as 'insufficient'.