1. Uitgangspunten van de toetsconstructie

Bij onderstaande beoordeling van de kwaliteitsaspecten met bijbehorende codes van het voornoemde beoordelingskader worden passages uit de wetenschappelijke verantwoording en de Handleiding veelal letterlijk vermeld. De wetenschappelijke verantwoording heeft betrekking op de uitgangspunten van de toetsconstructie, de normen, de betrouwbaarheid en meetnauwkeurigheid en de validiteit. De Handleiding heeft betrekking op het gebruik van de toets, communicatie over de toetsgegevens en de inhoudsverantwoording.

Algemeen

Het Cito Volgsysteem primair en speciaal onderwijs beoogt de vorderingen van individuele leerlingen, groepen leerlingen en het onderwijs op school van groep 1 tot en met groep 8 te volgen en te evalueren. De DMT (Drie-Minuten-Toets) is een onderdeel van het Cito Volgsysteem voor primair en speciaal onderwijs en is bedoeld voor leerlingen in groep 3 tot halverwege groep 8. De DMT is een papieren toets, bedoeld om het vaardigheidsniveau en de ontwikkeling van leerlingen op het gebied van technisch lezen in kaart te brengen. Onderstaande beschrijving is gebaseerd op de Handleiding.

Meetpretentie

Binnen het leesonderwijs op de basisschool wordt een onderscheid gemaakt tussen technisch lezen en begrijpend lezen. Bij technisch lezen gaat het erom of leerlingen in staat zijn om geschreven woorden correct en vlot te verklanken.
Bij begrijpend lezen staat het begrip van geschreven teksten centraal. Technisch lezen is geen doel op zich, maar wordt gezien als een voorwaardelijke activiteit voor begrijpend lezen. Het neemt in het basisonderwijs, in ieder geval in de onder- en middenbouw, een volwaardige plaats in op het lesrooster. De DMT is bedoeld om vast te stellen hoe leerlingen technisch kunnen lezen en hoe deze vaardigheid zich ontwikkelt in de loop der jaren.

Doelgroep

De toets is bedoeld voor leerlingen van groep 3 tot en met halverwege groep 8 van het primair en speciaal onderwijs. Voor leerlingen in het speciaal onderwijs zijn geen aparte normen vastgesteld. Leerlingen met een visuele beperking mogen langer over het lezen doen dan andere leerlingen en de leeskaarten kunnen worden vergroot. Voor leerlingen met leesproblemen en dyslexie geldt dit niet. Voor leerlingen die de Nederlandse taal onvoldoende beheersen en voor leerlingen met ernstige spraakmoeilijkheden is de DMT niet geschikt.

Gebruiksdoel en functie

Het hoofddoel van de DMT is tweeledig: enerzijds het in kaart brengen van het vaardigheidsniveau en anderzijds het zichtbaar maken van de ontwikkeling van leerlingen op het gebied van technisch lezen. Daarnaast biedt de DMT diagnostische informatie, zodat inzicht geboden kan worden in het leesgedrag van de leerling.

Inhoudelijke theoretische inkadering:

Hoewel technisch lezen in de vakliteratuur niet altijd op precies dezelfde manier wordt omschreven, is het mogelijk een min of meer universele, formele definitie van de vaardigheid te geven. Deze houdt in dat technisch lezen gelijk staat aan het correct en vlot ontsleutelen van geschreven woorden, al dan niet in context, en het herkennen van deze woorden als dragers van betekenis.

Inhoud van het toetspakket
Het toetspakket DMT bestaat uit de volgende documenten:

- Handleiding, deze bevat informatie over:
  - de afname van de toets (hfdst. 2),
  - rapportages (hdfst. 3),
  - interpretatie en analyse op leerling- en groepsniveaus (hfdst 4),
  - algemene aandachtspunten voor het schoolplan (hfdst 5),
  - inhoudsverantwoording (hfdst 6),
  - communiceren over toetsresultaten met leerling en ouders (hfdst 7),
  - achtergrondinformatie en veelgestelde vragen (hfdst 8) en
  - enkele bijlagen
- Toetsmateriaal:
  - Afname-instructies DMT
  - Score-instructies DMT
  - Leeskaarten (3 leeskaarten met elk 150 woorden. De woordtypen verschillen per leeskaart. Van elke kaart is een A en B-versie/parallelversie)
  - scoreformulieren
- Tabellen
  - Tabel Leeskaart 1 en 2 op M3, A t/m E
  - Tabel Leeskaart 1 en 2 op M3, I t/m V
  - Tabel Leeskaarten 1, 2, 3 op M3 t/m M8, A t/m E
  - Tabel Leeskaarten 1, 2, 3 op M3 t/m M8, I t/m V
  - Tabel Leeskaart 2 en 3 op M5 t/m M8, A t/m E
  - Tabel Leeskaart 2 en 3 op M5 t/m M8, I t/m V

Er zijn 2 uitgaven van de DMT: een voor het onderwijs en een voor zorgaanbieders. Deze beoordeling heeft betrekking op de DMT voor het onderwijs.

2. Beoordeling van de kwaliteitsaspecten

De beoordeling vindt plaats volgens het 'Beoordelingskader voor de psychometrische aspecten van (reeksen van) toetsen uit leerlingvolgsystemen (LOVS)', zoals opgesteld door de Expertgroep Toetsen PO. De Expertgroep Toetsen PO wordt gevormd door Prof. Dr. Cees Van der Vleuten (voorzitter), Prof. Dr. Cees Glas (psychometrisch expert), Dr. Desirée Joosten-Ten Brinke (onderwijskundig expert) en mevrouw Paulyn K. Berding-Oldersma MSc (secretaris).

De kwaliteit van de steekproef

S1.1. Is de steekproef representatief?

Bevindingen
In paragraaf 4.2 van de wetenschappelijke verantwoording wordt aandacht besteed aan de kenmerken die relevant zijn om de relatie tussen de normeringssteekproeven en de populatie in kaart te brengen. Bij de normering van de DMT is rekening gehouden met de achtergrondvariabelen regio (regio Noord, regio Oost, regio West en regio Zuid), urbanisatiegraad (niet tot matig verstedelijkt (platteland) en sterk tot zeer sterk verstedelijkt (stad)), schooltype (percentage achterstandsleerlingen [0, 0,15] en percentage achterstandsleerlingen [0,15, 1]) en sekse (jongens en meisjes), waarbij de

eerste drie variabelen op schoolniveau zijn gedefinieerd en de laatste variabele op leerlingniveau. Een controle op de representativiteit werd uitgevoerd door de populatieverdelingen van gegevens uit DUO en CBS te vergelijken met de steekproefverdelingen van M3 t/m E8. In de tabellen 4.2 - 4.11 (pag. 30-36) worden de resultaten van de representativiteitsanalyses gepresenteerd.

De representativiteitsanalyse bestond uit twee onderdelen: ten eerste is per achtergrondvariabele een vergelijking gemaakt tussen de verdeling van leerlingen in een steekproef en de verdeling van leerlingen in de populatie via een chi-kwadraat toetsing. Voor de interpretatie is echter vooral gekeken naar de effectgrootte φ, omdat significantie niet zoveel zegt bij grotere steekproeven. Een φ van 0.10 kan aangemerkt worden als een klein effect, een φ van 0.30 als een gemiddeld effect, en een φ van 0.50 als een groot effect (Cohen, 1988). Ten tweede is nagegaan of de prestaties van groepen leerlingen betekenisvol van elkaar verschilden. In het regressiemodel was de toetsscore (totaalscore over de afgenomen leeskaarten) de afhankelijke variabele (criteriumvariabele) en waren regio, urbanisatiegraad, schooltype en sekse de verklarende variabelen (predictoren). De regressieparameters zijn geïnterpreteerd in termen van significantie (z) en relevantie (d). De z-waarden zijn bepaald door de regressieparameters te delen door de standaardfout en de effectgroottes d door de regressieparameters te delen door de residuele standaarddeviatie. We spreken van een klein effect als d = 0.20, van een gemiddeld effect als d = 0.50, en van een groot effect als d = 0.80 (Cohen, 1988).

Voor afnamemoment M3 bestond de steekproef uit 564 leerlingen, voor afnamemoment E3 bestond de steekproef uit 655 leerlingen, voor afnamemoment M4 bestond de steekproef uit 532 leerlingen, voor afnamemoment E4 bestond de steekproef uit 820 leerlingen, voor afnamemoment M5 bestond de steekproef uit 585 leerlingen, voor afnamemoment E5 bestond de steekproef uit 720 leerlingen, voor afnamemoment M6 bestond de steekproef uit 576 leerlingen, voor afnamemoment E6 bestond de steekproef uit 736 leerlingen, voor afnamemoment M7 bestond de steekproef uit 590 leerlingen, voor afnamemoment E7 bestond de steekproef uit 671 leerlingen en voor afnamemoment M8 bestond de steekproef uit 604 leerlingen.

Samenvattend leveren de representativiteitsanalyses het volgende beeld op:

- De normeringssteekproeven hebben met elkaar gemeen dat de verdelingen naar regio en urbanisatiegraad afwijken van de populatieverdeling. De zuidelijke provincies en de plattelandsgebieden zijn oververtegenwoordigd.
- In de meeste normeringssteekproeven is er sprake van afwijkingen in de verdeling naar schooltype. Met uitzondering van afnamemomenten medio 4, medio 5, einde 7 en medio 8 zijn de scholen met weinig achterstandsleerlingen oververtegenwoordigd ten opzichte van de scholen met veel achterstandsleerlingen.
- In vrijwel alle normeringssteekproeven is de verdeling naar sekse in overeenstemming met de populatieverdeling. Op dit punt zijn de normeringssteekproeven dus representatief te noemen. Afnamemoment medio 5 vormt een uitzondering: daar zijn de meisjes licht oververtegenwoordigd ten opzichte van de jongens.
- De regressieanalyses laten geen consistent beeld zien over de afnamemomenten. Soms lijkt er een relatie te bestaan tussen een bepaalde variabele en de prestaties van leerlingen, maar een duidelijk patroon over de afnamemomenten ontbreekt. Er lijkt eerder sprake te zijn van toeval dan van een systematisch effect.

Op grond van de uitkomsten van de representativiteitsanalyses (af en toe sprake van een onder- of overrepresentatie van een bepaalde groep leerlingen én een betekenisvol prestatieverschil) is het twijfelachtig of op basis van de beschikbare gegevens een landelijke representatieve normering opgesteld kan worden voor de DMT. Daarom is besloten om via *sampling* per afnamemoment een representatieve steekproef te trekken uit de beschikbare data (*N* minimaal 400) en de scoreverdelingen van de subsamples te vergelijken met de scoreverdelingen van de volledige steekproeven. Uit een vergelijking tussen de cumulatieve scoreverdelingen van de volledige steekproeven en de *subsamples* voor alle afnamemomenten M3, E4, M4, E4, M5, E5, M6, E6, M7, E7 en M8 (zie Fig. 4.2 op pag. 38-39) blijkt dat deze volledig met elkaar overlappen. Dit betekent dat het voor de normering weinig uitmaakt of deze gebaseerd is op de volledige steekproeven of de subsamples. Daarom is ervoor gekozen om de normering te baseren op de volledige steekproeven en bij de normering dus alle beschikbare data te gebruiken.

Conclusie
De representativiteit van de steekproef is onderzocht met betrekking tot regio, urbanisatiegraad, schooltype en sekse. Op aspect S1.1 wordt de toets DMT als '**voldoende**' beoordeeld.

S1.2. In geval van een onvolledig dataverzamelingsdesign: is het design adequaat?

Bevindingen
Na een screening en try-out (uit een resulterende set van 13.000 inhoudswoorden is aselect een groot aantal woorden getrokken) zijn de herziene versies van de leeskaarten (versie A, versie B en versie C) in een proeftoets voorgelegd aan leerlingen in de eerste twee leerjaren van het primair onderwijs. Versie A en B bevatten dezelfde woorden, maar in een andere volgorde en zijn dus parallelversies. De leeskaarten van versie C (uitgave voor zorgaanbieders) bevatten andere woorden dan versie A en B, maar zijn qua opzet volledig vergelijkbaar. Bij iedere afname worden drie leeskaarten elk gedurende één minuut hardop voorgelezen door een toetsleider aan de leerlingen, waarbij elke leeskaart 150 woorden bevat. Periode van afname was het E-moment (mei/juni 2015). De leeskaarten werden niet in de hogere groepen afgenomen, omdat eventuele struikelblokken als onverwacht vaak fout gelezen woorden het beste in de laagste leerjaren opgespoord konden worden.

Figuur 3.1 op pag. 25 laat gedetailleerd het *proeftoetsdesign* zien voor jaargroep 3 en 4. De verschillende versies van de DMT leeskaarten zijn volgens een "balanced incomplete block" design aan subgroepen toegewezen, hetgeen inhoudt dat alle versies even vaak, in alle mogelijke combinaties, zijn afgenomen. Er zijn in totaal 270 leerlingen geworven per jaargroep (10 leerlingen x 27 subgroepen), zodat van elke leeskaart van de DMT minstens 90 waarnemingen beschikbaar waren per jaargroep (180 in totaal). Zoals blijkt uit Tabel 3.7 op pag. 26 (aantal leerlingen in de proeftoets uitgesplitst naar afnamemoment en leeskaart) is deze doelstelling ruimschoots gerealiseerd. Na afloop van de proeftoets zijn de scores op de verschillende leeskaarten geanalyseerd en is ervoor gezorgd, net zoals bij de try-out, dat woorden die werden vervangen door andere overeenkwamen voor wat betreft woordtype en leesmoeilijkheden.

In schooljaar 2015/2016 is een landelijk normeringsonderzoek in het reguliere basisonderwijs uitgevoerd ten behoeve van de normering van de DMT. Een cross-

sectioneel design is gebruikt waarin leerlingen in de groepen 3 t/m 7 op twee momenten in het schooljaar werden getoetst, namelijk in januari/februari (medio afname) en in mei/juni (einde afname). De leerlingen in groep 8 zijn alleen in januari/februari getoetst. Er is naar gestreefd om bij de medio en einde afname in een jaargroep dezelfde leerlingen te toetsen, zodat de voortgang van deze leerlingen in beeld kon worden gebracht. Bij de toetsafname kregen leerlingen drie leeskaarten van de DMT voorgelegd, met uitzondering van de leerlingen in groep 3. Zij lazen bij de medio afname alleen kaart 1 en 2, omdat de meeste leerlingen halverwege groep 3 nog niet toe zijn aan het lezen van meerlettergrepige woorden (bijv. groente, hartelijk en schoolvakantie). De woordtypen van kaart 1, 2 en 3 werden door elkaar aangeboden en waren afkomstig uit de (nieuw ontwikkelde) DMT. Figuur 4.1 op pag. 27 geeft gedetailleerd weer hoe het design eruit zag voor jaargroep 4 t/m 7. Het design voor jaargroep 3 en 8 was gebaseerd op dit design, maar week op enkele punten af. Voor jaargroep 3 gold dat alle versies van kaart 3 zijn weggelaten bij de medio-afname. Voor jaargroep 8 gold dat de eindafname kwam te vervallen.

Figuur 4.1 op pag. 27 laat zien dat elke jaargroep tijdens het normeringsonderzoek is opgedeeld in 27 subgroepen. Evenals bij de proeftoetsing zijn de verschillende versies van de DMT leeskaarten wederom volgens een "balanced incomplete block" design aan subgroepen toegewezen. Zoals reeds eerder gememoreerd betekent dit dat alle versies even vaak, in alle mogelike combinaties, zijn afgenomen, en dat leerlingen bij de tweede afname (eind) nooit dezelfde versie maakten als bij de eerste afname (medio). Bij het normeringsonderzoek is er naar gestreefd om elke subgroep te laten bestaan uit 20 leerlingen. Per afnamemoment zouden dan 540 leerlingen deelnemen (20 leerlingen x 27 subgroepen) en in totaal zou elke leeskaart van de DMT 1980 keer worden afgenomen (9 subgroepen x 20 leerlingen x 11 afnamemomenten (M3 t/m M8)). Deze doelstelling is ruimschoots gerealiseerd. Tabel 4.1 op pag. 28 toont het aantal leerlingen in de steekproef uitgesplitst naar afnamemoment en leeskaart, waaruit blijkt dat bij alle afnamemomenten het gerealiseerde leerlingaantal groter is dan het streefaantal van 540 en ook het aantal waarnemingen per leeskaart steeds groter is dan het streefaantal van 1980. Door deze opzet konden dus uitgebalanceerde sets leeskaarten aan leerlingen worden voorgelegd.

Tenslotte bleek uit het gepresenteerde kalibratieonderzoek in Hoofdstuk 4 van de wetenschappelijke verantwoording dat voor de DMT de data passen bij het Poisson-model als geassumeerd meetmodel nadat het model in een vervolgstap was verfijnd met drie latente klassen (voor het model zonder latente klassen waren zowel de kaartparameters σ als de persoonsparameters θ niet invariant). Het voorgaande betekent dat er sprake is van een eendimensionale vaardigheidsschaal waar kaarten en leerlingen op afgebeeld kunnen worden.

Conclusie
Op aspect S1.2 wordt de toets DMT als '**voldoende**' beoordeeld.

Normering

N1.2.1. Zijn de normgroepen groot genoeg?

Bevindingen
In Tabel 4.12 op pag. 48 wordt de normtabel gepresenteerd voor alle afnamemomenten

M3 t/m M8. Deze informatie wordt ook visueel weergegeven in Figuur 4.8 op pag. 49 in de vorm van cumulatieve vaardigheidsverdelingen. In Tabel 4.12 op pag. 48 worden vaardigheidsverdelingen gepresenteerd, d.w.z. gemiddelde score, standaarddeviatie en de percentielen P10, P20, P25, P40, P50, P60, P75, P80 en P90. Met behulp van deze percentielen kunnen de twee gebruikelijke Cito niveau-indelingen A-E en I-V opgesteld worden, welke een indicatie geven van het relatieve niveau van de leerlingen. Uit Tabel 4.12 en Fig. 4.8 blijkt dat de gemiddelde vaardigheid van de leerlingen steeds toeneemt tussen de opeenvolgende metingen, dat de variantie in de fase van het aanvankelijk lezen kleiner is dan in latere fasen van de (technische) leesontwikkeling, en dat de afstanden tussen de grenswaarden voor niveau-indelingen A-E en I-V behoorlijk groot zijn. We kunnen dan ook concluderen dat de DMT gedurende de gehele basisschoolperiode goed in staat is om groepen leerlingen met een verschillend vaardigheidsniveau van elkaar te onderscheiden.

Conclusie
Op aspect N1.2.1 wordt de toets DMT als '**voldoende**' beoordeeld.

N1.2.2. Zijn de normgroepen representatief?

Bevindingen
De representativiteit van de steekproeven werd in S.1.1 besproken en daar werd geconstateerd dat de steekproeven representatief waren voor de achtergrondvariabelen regio, urbanisatiegraad, schooltype en sekse.

Conclusie
Op aspect N1.2.2 wordt de toets DMT als '**voldoende**' beoordeeld.

Betrouwbaarheid

B1.1. Zijn of worden de betrouwbaarheidsgegevens correct berekend?

Bevindingen
In hoofdstuk 5 van de wetenschappelijke verantwoording wordt beschreven hoe, gebruikmakend van het feit dat we beschikken over een zuivere schatter van de leessnelheid (het aantal correct gelezen woorden per tijdseenheid(minuut) op een standaard leeskaart), een betrouwbaarheidscoëfficiënt berekenend kan worden die qua interpretatie grote overeenkomst vertoont met de betrouwbaarheidscoëfficiënt uit de klassieke testtheorie. Hierbij geldt dat de betrouwbaarheid van de geobserveerde scores per definitie gelijk is aan de proportie geobserveerde variantie dat ware variantie is, waarbij de geobserveerde variantie gelijk is aan de variantie van de schattingen in een representatieve steekproef en de ware variantie gelijk is aan de geobserveerde variantie minus de gemiddelde meetfoutvariantie. De gemiddelde meetfoutvariantie wordt berekend door de variantie van de meetfout (formule (3) op pag. 52 van de wetenschappelijke verantwoording) te berekenen voor elke leerling en dan te middelen. De moeilijkheidsparameter σ van de leeskaart (taakparameter) in formule (3) op pag. 52 kan, onder gebruikmaking van het Poisson-model als meetmodel, consistent (d.w.z. zonder een aanname te hoeven maken over de verdeling van de technische leesvaardigheid in de populatie) worden geschat (Verhelst & Kamphuis, 2009). De aldus berekende

betrouwbaarheidscoëfficiënt van de schatter voor het aantal correct gelezen woorden per tijdseenheid(minuut) op een standaard leeskaart wordt in de psychometrische literatuur beschreven en als correct aangemerkt.

Conclusie
De betrouwbaarheidsgegevens worden correct berekend. Op aspect B1.1. wordt de toets DMT als '**voldoende**' beoordeeld.

B1.2. Zijn de betrouwbaarheidsgegevens voldoende gezien de beslissingen die met de toets genomen worden?

Bevindingen
De hierboven beschreven procedure om de betrouwbaarheid te schatten voor het aantal correct gelezen woorden per tijdseenheid(minuut) op een standaard leeskaart werd toegepast op alle toegestane combinaties van de DMT-leeskaarten bij alle afnamemomenten (zie Tabel 5.1 op pag. 52 van de wetenschappelijke verantwoording). Voor M3 is de betrouwbaarheid voor leeskaart 1 + 2 gelijk aan 0,966. Voor M5, E5, M6, E6, M7, E7 en M8 zijn de betrouwbaarheden voor leeskaart 2 + 3 gelijk aan resp. 0,921, 0,905, 0,905, 0,889, 0,883, 0,879 en 0,863. Voor E3, M4, E4, M5, E5, M6, E6, M7, E7 en M8 zijn de betrouwbaarheden voor leeskaart 1 + 2 + 3 gelijk aan resp. 0,976, 0,965, 0,966, 0,955, 0,946, 0,945, 0,936, 0,932, 0,927 en 0,917. Zoals uit deze resultaten blijkt, hebben de betrouwbaarheidsindices voor de toegestane combinaties van leeskaarten hoge waarden (0,863 of hoger). De auteurs van de wetenschappelijke verantwoording verwijzen naar het beoordelingssysteem van de COTAN waar voor tests die geen zware consequenties hebben voor leerlingen, zoals de DMT, een betrouwbaarheidscoëfficiënt van meer dan 0,80 als 'goed' aangemerkt moet worden.

Er zijn geen test-hertest betrouwbaarheden uitgerekend, omdat de afnamecontext van de DMT-kaarten hier zich niet goed voor leent. Het feit dat alle leeskaarten echter gekalibreerd zijn (d.w.z. hun moeilijkheidsparameters consistent geschat zijn), maakt het mogelijk een hertest te simuleren. Er is dan ook een dubbele afname gesimuleerd van 100.000 leerlingen, waarbij enerzijds de vaardigheidsverdeling van alle leerlingen en anderzijds alle kaartparameters als uitgangspunt is genomen. Hierbij is steeds een bepaalde vaardigheid aselect uit de verdeling genomen en zijn twee bij deze vaardigheid horende afnames gesimuleerd. De correlatie tussen deze 100.000 dubbele (virtuele) afnames is berekend en kan worden beschouwd als de test-hertest betrouwbaarheid onder ideale condities. De tweede toetsafname is immers volledig onafhankelijk van de eerste toetsafname en wordt dus niet beïnvloed door de kennis die de leerling mogelijk verworven heeft in de eerste toetsafname, m.a.w. er is geen sprake van een leer/herinneringseffect ('carry-over effect'). Daarnaast is er geen sprake van invloed van een test-hertest interval: beide afnames worden gesimuleerd alsof zij op hetzelfde moment plaats zouden vinden.

Voor toegestane combinaties van de DMT-leeskaarten bij alle afnamemomenten zijn de gesimuleerde test-hertest betrouwbaarheden als volgt berekend (zie Tabel 5.1 op pag. 52 van de wetenschappelijke verantwoording): Voor M3 is de betrouwbaarheid voor leeskaart 1 + 2 gelijk aan 0,951. Voor M5, E5, M6, E6, M7, E7 en M8 zijn de betrouwbaarheden voor leeskaart 2 + 3 gelijk aan resp. 0,925, 0,910, 0,909, 0,894, 0,885, 0,881 en 0,866. Voor E3, M4, E4, M5, E5, M6, E6, M7, E7 en M8 zijn de betrouwbaarheden voor leeskaart 1 + 2 + 3 gelijk aan resp. 0,969, 0,955, 0,957, 0,943, 0,931, 0,934, 0,924, 0,922, 0,919 en

0,910. De uitkomsten komen goed overeen met de eerder berekende betrouwbaarheidscoëfficiënten van de schatter voor het aantal correct gelezen woorden per tijdseenheid(minuut) op een standaard leeskaart en leiden dan ook tot dezelfde conclusies m.b.t. de betrouwbaarheid van de DMT, nl. dat deze goed zijn te noemen in relatie tot het beoogde gebruik van de toets.

De globale betrouwbaarheidscoëfficiënten zijn adequaat gekozen en de globale meetnauwkeurigheid van de toegestane combinaties van leeskaarten blijkt ruim voldoende te zijn. Naast globale betrouwbaarheidscoëfficiënten zijn ook indices berekend (zowel grafisch als tabellarisch), die een beeld geven van de lokale meetnauwkeurigheid. In Figuur 5.1 op pag. 53 van de wetenschappelijke verantwoording wordt voor toegestane combinaties van DMT-leeskaarten (Kaart 1 + 2, Kaart 2 + 3 en Kaart 1 + 2 + 3) grafisch de grootte van de (standaard)meetfout afgebeeld, alsmede de cumulatieve vaardigheidsverdelingen voor de afnamemomenten M3, E3, M4, E4, M5, E5, M6, E6, M7, E7 en M8. Figuur 5.1 maakt duidelijk dat voor combinaties van leeskaarten de meetfout toeneemt met toenemende vaardigheid, maar alleszins binnen de perken blijft. Dit geldt zeker voor de aanbevolen combinatie van leeskaart 1 + 2 + 3 (m.u.v. afnamemoment M3).

Ook wordt de betekenis van de meetnauwkeurigheid voor de beslissingen die met de toetsen genomen worden, gepresenteerd in classificatie-/misclassificatietabellen (Tabel 5.2 op pag. 54 voor de combinatie van leeskaarten 1 + 2 + 3 eind jaargroep 4 (E4) en Bijlage 1 voor de toegestane afnamecombinaties voor alle afnamemomenten). Uit Tabel 5.2 blijkt bijv. dat 83,6% en 82,8% van de leerlingen die eind jaargroep 4 op basis van hun geschatte vaardigheidsscore aan resp. niveaugroep A en E worden toegekend, ook met hun werkelijke vaardigheidsscore in deze groepen vallen. Voor de middelste niveaugroepen B, C en D liggen deze percentages iets lager, nl. 66,6%, 75,2% en 71,8%. Verder laat Tabel 5.2 het volgende zien:

- 16,3% en 0,2% van de leerlingen in niveaugroep A heeft een werkelijke vaardigheidsscore die in resp. niveaugroep B en niveaugroep C valt en er zijn geen leerlingen in niveaugroep A die een werkelijke vaardigheidsscore hebben die in niveaugroep D of E valt.
- 17,2% van de leerlingen in niveaugroep E heeft een werkelijke vaardigheidsscore die in niveaugroep D valt en er zijn geen leerlingen in niveaugroep E die een werkelijke vaardigheidsscore hebben die in niveaugroep C, B of A valt.
- 10,4%, 22,9% en 0,1% van de leerlingen in niveaugroep D heeft een werkelijke vaardigheidsscore die in resp. niveau E, C en B valt en er zijn geen leerlingen in niveaugroep D die een werkelijke vaardigheidsscore hebben die in niveaugroep A valt.
- 10,6% en 14,1% van de leerlingen in niveaugroep C heeft een werkelijke vaardigheidsscore die in resp. niveaugroep D en B valt en er zijn geen leerlingen in niveaugroep C die een werkelijke vaardigheidsscore hebben die in niveaugroep E of A valt.
- 0,1%, 17,2% en 11% van de leerlingen in niveaugroep B heeft een werkelijke vaardigheidsscore die in resp. niveaugroep D, C en A valt en er zijn geen leerlingen in niveaugroep B die een werkelijke vaardigheidsscore hebben die in niveaugroep E valt.

Bijlage 1 toont ook aan dat de lokale meetnauwkeurigheid groter is bij de afname van drie kaarten dan bij de afname van twee kaarten. Daarom wordt in de handleiding bij de

toetsen ook aanbevolen om drie leeskaarten af te nemen, m.u.v. het afnamemoment medio groep 3 (M3), wanneer de woorden van leeskaart 3 nog te hoog gegrepen zijn voor de meeste leerlingen. Verder bevestigt Bijlage 1 het beeld dat we reeds zagen voor eind jaargroep 4 (E4) in Tabel 5.2 op pag. 54 dat de leerlingen met de hoogste en de laagste niveau-indicatie in het algemeen voldoende betrouwbaar geclassificeerd worden, maar de andere niveau-indicaties (middelste niveau-indicaties) iets minder betrouwbaar geclassificeerd worden (bijv. 61,3% en 63,0% voor resp. niveaugroep D en niveaugroep B voor de combinatie van leeskaarten 1 + 2 + 3 eind groep 6 (E6); zie Tabel 12 uit Bijlage 1). Evenals bij Tabel 5.2 op pag. 54 voor de leerlingen eind jaargroep 4 geldt echter wel dat alle niveau-indicaties in Bijlage 1 (dus zowel de extreme als de middelste niveau-indicaties) er maximaal één niveau naast kunnen zitten.

In dit verband zou nog kunnen worden opgemerkt dat het feit dat de lokale meetnauwkeurigheid groter is bij de afname van drie kaarten dan bij de afname van twee kaarten ook te maken heeft met het gegeven dat de betrouwbaarheden bij de afname van drie leeskaarten groter is dan bij de afname van twee leeskaarten. Immers: hoe betrouwbaarder de toets, des te minder classificatiefouten dankzij een grotere meetnauwkeurigheid.

Omdat er in de onderzoeksliteratuur weinig is geschreven over de beoordeling van betrouwbaarheidstabellen (wanneer kan een betrouwbaarheidstabel als goed of voldoende worden beschouwd en wat mag onder ideale omstandigheden verwacht worden), wordt de informatie uit de betrouwbaarheidstabellen vaak samengevat in één of meerdere indices. De auteurs gebruiken twee indices: de *plus/minus niveau-index* en de *Marginal Classification Accuracy* (MAC). De eerste maat is bedacht door Pilliner (1969) en stelt als ambitieniveau dat 95% van de leerlingen in een scoregroep in werkelijkheid ook in die scoregroep moet scoren, óf een scoregroep daarboven óf een scoregroep daaronder. De rationale voor dit ambitieniveau is dat geen enkele toets perfect meet en dat er dus altijd sprake is van misclassificaties (zelfs bij heel hoge betrouwbaarheden). Vanuit de plus/minus niveau-index geredeneerd is de maximale accuraatheid die op het individuele niveau bereikt kan worden dus plus of minus één scoregroep. De tweede maat laat zien hoe vaak de classificatie op basis van de geschatte vaardigheidsscore gemiddeld gezien overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. Bij een ideale, maar in de praktijk niet te realiseren, toetsafname lijkt de MAC rond 0,75 – 0,80 uit te komen. In de praktijk liggen de waarden vaak tussen 0,60 en 0,70.

Deze twee samenvattende indices voor afnamemoment E4 voor de afnamecombinatie van leeskaart 1 + 2 + 3 zijn te vinden in Tabel 5.3 op pag. 55. Uit deze tabel blijkt dat de uitkomsten van de plus/minus 1 niveau-index in lijn liggen met het ambitieniveau zoals dat door Pilliner (1969) geformuleerd is. Gemiddeld gezien scoort nl. 99,9% en 99,7% van de leerlingen in een niveaugroep ook in werkelijkheid in die niveaugroep, óf een niveaugroep daarboven óf daaronder voor resp. vaardigheidsniveaus E t/m A en vaardigheidsniveaus I t/m V, d.w.z. dat gebruikers rekening dienen te houden met maximaal 1 niveauverschil.

Uit Tabel 5.3 op pag. 55 blijkt verder dat de MAC voor de afnamecombinatie van leeskaarten 1 + 2 + 3 eind jaargroep 4 op 76% en 73,5% ligt voor resp. vaardigheidsniveaus E t/m A en vaardigheidsniveaus I t/m V. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore bij afnamemoment eind groep 4

op basis van drie leeskaarten gemiddeld gezien in ongeveer driekwart van de gevallen overeenkomt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. De MAC-waarden van 76% en 73,5% voor de afnamecombinatie van leeskaarten 1 + 2 + 3 eind jaargroep 4 voor resp. vaardigheidsniveaus E t/m A en vaardigheidsniveaus I t/m V zijn dus in overeenstemming met het ambitieniveau (idealiter rond 0,75 – 0,80, maar in de praktijk vaak tussen 0,60 en 0,70).

De samenvattende indices plus/minus 1 niveau-index en de MAC voor de andere afnamemomenten op basis van alle toegestane afnamecombinaties zijn te vinden in Bijlage 2 van de wetenschappelijke verantwoording. Deze Bijlage toont aan dat de twee samenvattende indices bij afname van drie leeskaarten min of meer hetzelfde beeld laten zien als bij de waarden voor de twee samenvattende indices in Tabel 5.3 op pag. 55 voor de afnamecombinatie van leeskaarten 1 + 2 + 3 eind jaargroep 4. Alleen de MAC daalt in de hogere groepen naar ongeveer 70%, d.w.z. dat de geschatte vaardigheidsscore in de hogere groepen iets minder correct wordt geclassificeerd dan in de lagere groepen.

Omdat er rekening moet worden gehouden met het gegeven dat men er bij de classificering van leerlingen in vaardigheidsniveaus altijd één niveau naast kan zitten, verdient dit de aanbeveling om dit expliciet in de Handleiding te vermelden.

Conclusie
De betrouwbaarheid van de toets DMT is 'voldoende' als aangenomen mag worden dat de toets geen zware consequenties voor de leerlingen heeft en ingestemd wordt met de beoordelingscriteria voor de betrouwbaarheid van de COTAN en er bij de classificatie van leerlingen in vaardigheidsniveaus rekening wordt gehouden met het gegeven dat men er altijd één niveau naast kan zitten.

Op basis van het voorgaande wordt op aspect B1.2 aan de toets DMT groep 3 t/m 8 het oordeel '**voldoende**' toegekend.

Validiteit

V1. Dragen de items in de toets bij aan de validiteit van de toets (hierbij gaat het om aspecten als relevantie, objectiviteit en efficiëntie van de items)

Bevindingen
De toets bestaat uit het hardop lezen van losse woorden van uiteenlopende moeilijkheidsgraad. De woorden bestaan uit schooltaalwoorden (bijvoorbeeld het woord 'zelfs') en hoogfrequente leenwoorden (als het woord mobiel uit kaart 3).
De woorden op de kaarten komen uit het BasiLex-corpus (verzameling geschreven teksten die Nederlandse kinderen vanaf groep 3 krijgen aangeboden en zijn gelabeld op frequentie, woordsoort, lengte en leeftijd). Dit heeft de ontwikkelaars goede handvatten geboden om te komen tot een juiste keuze van woorden op de kaarten.
Er is een goede afgewogen keuze gemaakt om per kaart ongeveer 75% zelfstandig naamwaarden op te nemen, ongeveer 20% bijvoeglijk naamwoorden en ongeveer 7% overige woordsoorten. Met als reden dat er anders onbedoeld halve zinnen konden ontstaan op de leeskaart bij het hardop lezen van de woorden en dat overige woordsoorten (lidwoorden, voornaamwoorden etc.) te frequent zouden voorkomen.

Een goede verbetering t.o.v. de DMT (Cito, 2009) is te werken met een kaartstructuur bestaande uit 15 units van steeds 10 woorden met ook nog aandacht voor verdeling van de leesmoeilijkheden (bijv. Woorden met sch-, aai-,-x etc.). Binnen iedere unit is er een evenwichtige verdeling tussen verschillende woordstructuren. Dus niet de moeilijke woorden achteraan (denk aan vierlettergrepige woorden), waardoor nu langzame en snellere lezers onderling goed te vergelijken zijn.

Merk op dat de beoordeling van dit aspect zich hieronder beperkt tot het statistisch/psychometrisch onderzoek dat verricht is.

In de wetenschappelijke verantwoording wordt verslag gedaan van verschillende aspecten van begripsvaliditeit, betrekking hebbende op de vraag in hoeverre de toets het door de constructeur beoogde kenmerk van de leerling (onderliggende trek, vaardigheid) meet: unidimensionaliteit, soortgenotenvaliditeit (convergente en divergente validiteit) en verschillen tussen relevante subgroepen.

Evidentie voor unidimensionaliteit wordt aangedragen door te verwijzen naar het kalibratieonderzoek van de DMT in paragraaf 4.4 van de wetenschappelijke verantwoording, waarvan de conclusie is dat het aannemelijk is dat er sprake is van unidimensionaliteit en dat de items één latente trek of construct meten dat we gezien ook de uitkomsten van het hierna genoemde onderzoek kunnen aanduiden als 'technisch lezen'. Hiermee is dus voldaan aan de noodzakelijke voorwaarde voor begripsvaliditeit.

Tabel 6.1 t/m 6.6 op pag. 58 t/m 61 van de wetenschappelijke verantwoording laten hoge correlaties zien zoals verwacht van de DMT met de soortgenotentoetsen EMT (Een-Minuut-Test) en AVI-leestijd (Analyse van Individualiseringsvormen), terwijl de correlaties van de DMT met de niet-soortgenotentoetsen Begrijpend Lezen en Woordenschat matig tot laag zijn, wat ook naar verwachting is omdat bij deze toetsen andere vaardigheden dan alleen technisch lezen een rol spelen. Samenvattend kan dan ook worden geconcludeerd dat de hoge correlaties van de DMT met de soortgenotentoets en de matige tot lage correlaties met de niet-soortgenotentoets kan worden beschouwd als een aanwijzing voor resp. convergente validiteit en divergente validiteit.

Tenslotte wordt in Tabel 6.7 op pag. 61 van de wetenschappelijke verantwoording getoond dat er geen betekenisvolle verschillen (Cohen's D < 0.1) worden gevonden voor de variabelen sekse en leerlinggewicht, terwijl voor de dichotome variabele dyslexie (wel of geen dyslexie) grote positieve effecten (Cohen's D van 0,32 en 0,48 voor resp. DMT medio en DMT eind) worden gevonden ten gunste van leerlingen bij wie geen dyslexie is vastgesteld (bij 10% van de ongeveer 2000 leerlingen in de steekproef was dyslexie vastgesteld). Deze prestaties van relevante subgroepen op de DMT sluiten aan bij theoretische verwachtingen ten aanzien van de prestaties van deze subgroepen (Schijf, 2009).

Conclusie
Op aspect V1. wordt de toets DMT als '**voldoende**' beoordeeld.

Het volg-aspect

VA1.1. Is er een voldoende empirische onderbouwing van de schaal waarop de groei van een leerling wordt uitgedrukt? Wordt groei op een adequate manier gemeten?

Bevindingen
De resultaten van het kalibratieonderzoek (zie paragraaf 4.4) laten zien dat de items van de DMT groep 3 t/m 8 op een eendimensionale vaardigheidsschaal afgebeeld kunnen worden en dat aan de hand van de door de leerling behaalde vaardigheidsscores (geschat via formule 6 op pag. 43 van de wetenschappelijke verantwoording) op de onderscheiden toetsen diens groei adequaat gemeten kan worden.

Conclusie
Op aspect VA1.1 wordt de toets DMT als '**voldoende**' beoordeeld.

VA1.2. Worden er gegevens verstrekt over hoe groei geïnterpreteerd dient te worden? Wordt de betrouwbaarheid van de groei op die schaal adequaat weergegeven?

Bevindingen
In Hoofdstuk 3 van de Handleiding wordt aan de hand van een leerlingrapport de interpretatie van groei op een duidelijke manier beschreven. De 67% betrouwbaarheidsintervallen van de vaardigheidsscores worden zowel op het leerlingrapport als in afzonderlijke tabellen vermeld. Groei wordt zowel met vergelijkingen tussen verschillende leerlingen als met vergelijkingen op andere tijdstippen van dezelfde leerlingen weergegeven. Zowel in de wetenschappelijke verantwoording en de Handleiding wordt het volgen van groei van leerlingen adequaat toegelicht.

Conclusie
Op aspect VA1.2 wordt de toets DMT als '**voldoende**' beoordeeld.

Inzicht in leervorderingen

I1.1 Levert de toetsaanbieder een format voor een geschreven toelichting bij de leervorderingen van de leerling die (ook) voor ouders/voogden/verzorgers begrijpelijk is?

Bevindingen
De toetsen worden 2 keer per jaar afgenomen (behalve in groep 8 waar alleen de middentoetsing plaatsvindt). In een afnameschema wordt aangegeven wanneer een keuze kan worden gemaakt om 1 van de 3 kaarten niet af te nemen wat wel invloed heeft op de nauwkeurigheid van de vaardigheidsscores. Dit wordt duidelijk vermeld.
Bij de keuze van de afname van de DMT-toets worden de doelen van de DMT naast die van de leeskaarten van de AVI-toets gezet om het verschil duidelijk te maken en leraren te ondersteunen bij de afnameopties.
De afname van de toets is helder beschreven en wordt ondersteund door een aparte kaart afname-instructies en een kaart score-instructies. Daarnaast is er een uitwerking gemaakt van FAQ's voor ondersteuning en advies bij het toetsproces.

Resultaten kunnen verwerkt worden op leerlingniveau (mogelijkheden: signaleren, interpreteren, analyseren, plannen en op volgend toetsmoment weer evalueren),

groepsniveau (resultaten van de groep als geheel met mogelijkheden voor analyse en aanpak) en schoolniveau (resultaten van alle groepen en onderwijsaanbod).
Er wordt helder aangegeven hoe de totaalscore van een toets via omzettingstabellen (opgenomen in de handleiding) omgezet kunnen worden in een vaardigheidsscore en vaardigheidsniveau per leerling. Digitaal kan ook nog het functioneringsniveau (met welk leerlingniveau in het basisonderwijs is de behaalde toetsscore van de leerling te vergelijken) bepaald worden en een categoriënanalyse worden gemaakt. Ook kan er per kaart nog een kaartenanalyse worden gemaakt (afzonderlijke score per kaart en de totaalscore over 2 of 3 kaarten weergegeven).
Bij de interpretatie en analyse, uitgewerkt in hoofdstuk 4, wordt nadrukkelijk zoals ook bij de handleidingen van andere toetsen van Cito de relatie gelegd tussen de ontwikkeling van de leerling zelf als naar de scores en ontwikkeling van de normgroep. Dit bestaat uit een opsomming van een aantal helder geformuleerde vragen en bij de analysevragen ook vragen specifiek toegespitst op de DMT. Regelmatig wordt er extra aandacht besteed aan hoe te handelen bij dyslectische kinderen. Dit biedt handvatten voor het eigen handelen maar biedt ook handvatten om het gesprek met ouders inhoudelijk te voeden.

Via de portal van Cito B.V. kan gebruik worden gemaakt van rapportageformulieren voor een leerlingrapport, groepsrapport, groepsoverzicht (overzicht van één groep leerlingen tijdens hun schoolperiode) en een alternatief leerlingrapport (voor leerlingen die op een eigen niveau werken).

In de Handleiding wordt in hoofdstuk 7 aandacht besteed aan hoe met ouders over de toetsresultaten gecommuniceerd kan/moet worden. In het bijzonder wordt daarbij gewezen op het leerlingrapport waarin zowel het niveau van de leerling als de progressie van de leerling numeriek en grafisch gepresenteerd worden. Hiervoor is ook een folder ouderinformatie beschikbaar die men via de website van het Cito kan downloaden. Daarnaast wordt de docent gewezen op misverstanden die zich bij de interpretatie van de niveau-indelingen bij de ouders kunnen voordoen. Ook moeten zij aan ouders het verschil tussen methode-onafhankelijke en methodegebonden toetsen duidelijk maken en erop wijzen dat deze toetsen leerlingen anders (kunnen) beoordelen.

In hoofdstuk 8 van de Handleiding worden ook veel gestelde vragen behandeld die weliswaar voor de docenten bestemd zijn maar zij kunnen met die informatie bijvoorbeeld via tienminutengesprekken ook de ouders beter voorlichten.

De wijze waarop de registratieformulieren zijn vormgegeven en de uitleg die voor docenten beschikbaar is geven het niveau en de groei van de individuele leerling weer.

Extra aandacht is er ook voor de communicatie tussen scholen en zorgaanbieders wat cruciaal is. De school levert namelijk aan de ouders alle benodigde gegevens over hun kind in de vorm van een leerlingdossier. Het leerlingdossier wordt bij aanmelding bij een dyslexiebehandelaar ingebracht.

Conclusie
Op aspect I1.2 wordt de toets DMT als '**voldoende**' beoordeeld.

3. Verzamelstaat

Kwaliteitsaspect	Code	Oordeel
De kwaliteit van de steekproef	S1.1	**Voldoende**
	S1.2	**Voldoende**
Normering	N1.1	**Voldoende**
	N1.2	**Voldoende**
Betrouwbaarheid	B1.1	**Voldoende**
	B1.2	**Voldoende**
Validiteit	V1.1	**Voldoende**
Volg-aspect	VA1.1	**Voldoende**
	VA1.2	**Voldoende**
Inzicht in leervorderingen	I1.1	**Voldoende**

4. Literatuurlijst

- Alma van Til, Frans Kamphuis, Jos Keuning, Martine Gijsel, Judith Vloedgraven, Anja de Wijs (2018). *Wetenschappelijke verantwoording LVS-toetsen DMT.* Arnhem: Cito.
- Cito (2017). *DMT.* Arnhem: Cito.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Erlbaum.
- Pilliner, A. (1969). *Estimation of number of grades to be awarded in an examination by considering of its reliability coefficient*. Edinburgh: The Godfrey Thomson Unit for Educational Research.
- Schijf, G.M. (2009). *Lees- en spellingvaardigheden van brugklassers* (proefschrift). Amsterdam: SCO-Kohnstamm Instituut, Universiteit van Amsterdam.
- Verhelst, N.D., and Kamphuis, F.H. (2009). *A Poisson-Gamma model for speed tests*. Measurement and Research Department Reports 2009-2. Arnhem: Cito.