1. Algemene informatie

Algemeen en meetpretentie
De volgtoetsen Diaspel meten de spelvaardigheid bij leerlingen vanaf eind groep 5 tot en met midden groep 8 van het basisonderwijs, en de ontwikkeling daarvan in deze onderwijsperiode. Er zijn zes volgtoetsen, bedoeld om bij de leerlingen uit deze groepen het niveau en de groei van de spelvaardigheid te kunnen vaststellen. Voor de volgtoetsen is er een doorlopende leerlijn opgesteld die opbouwt naar de referentieniveaus Taalverzorging. De inhoudelijke opbouw is gebaseerd op de opbouw van de leerstof die in de meest gebruikte methodes voor spelling wordt gehanteerd, en op de tussendoelen en leerstoflijnen van het Expertisecentrum Nederlands (2010) en TULE (2008). Tevens is een doorlopende vaardigheidsschaal ontwikkeld. Deze vaardigheidsschaal is gerelateerd aan de referentieniveaus Taal van het Referentiekader (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009) en maakt het mogelijk om de ontwikkeling van leerlingen te volgen over de leerjaren heen.

Doelgroep
De doelgroep voor de toetsen die hier worden verantwoord bestaat uit de leerlingen in de bovenbouw van het reguliere en speciaal basisonderwijs vanaf eind groep 5 tot en met midden groep 8. De toetsen Diaspel zijn genormeerd bij leerlingen uit het reguliere basisonderwijs, voor het speciaal basisonderwijs is geen aparte normering beschikbaar.

Gebruiksdoel en functie
Diaspel bevat lineaire (niet-adaptieve) toetsen. Er zijn zes volgtoetsen. Er zijn drie toetsmomenten per schooljaar, waarbij de eindmeting van het vorige leerjaar inhoudelijk gelijk is aan de beginmeting van het volgende leerjaar. De volgtoetsen Diaspel zijn bedoeld om per meetmoment het niveau van de spelvaardigheid van leerlingen te bepalen.

Inhoudelijke theoretische inkadering:
Diaspel is geijkt aan de openbare referentiesets Taalverzorging (College voor Toetsen en Examens, 2016a).

Inhoud van het toetspakket
Het toetspakket Diaspel 678 bestaat uit de volgende documenten:

- Wetenschappelijke verantwoording, deze bevat informatie over:
  - Uitgangspunten (hoofdstuk 2);
  - Inhoudsverantwoording (hoofdstuk 3);
  - Kalibratie en normering (hoofdstuk 4);
  - Betrouwbaarheid (hoofdstuk 5);
  - 10 bijlagen, waaronder items en antwoorden.
- Handleiding Diaspel 678
- Toetsreglement Dia-LVS PO
- Wegwijs in Dia-groeiwijzer
- Inzage digitale toetsitems diaspel

2. Beoordeling van de kwaliteitsaspecten

De beoordeling vindt plaats volgens het 'Beoordelingskader voor instrumenten binnen leerlingvolgsystemen (LVS)', 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. Desiree Joosten-Ten Brinke (onderwijskundig expert) en Jennifer Roubiës MSc (secretaris).
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 kwaliteit van de dataverzameling

S1 Is de steekproef van leerlingen representatief?

Bevindingen:
Aan het kalibratieonderzoek deden in totaal 80 scholen mee met leerlingen uit de groepen 5, 6, 7 en 8. Voor het normeringsonderzoek zijn extra items ontwikkeld. Daarom zijn nieuwe scholen geworven, waarbij weer getracht is de steekproef zo representatief mogelijk te houden. In totaal hebben er 18.993 leerlingen meegedaan. Het geslacht van de leerling is door de scholen opgegeven. De achtergrondvariabelen regio, denominatie en urbanisatiegraad zijn aan de hand van het vestigingsnummer (BRIN-nummer plus vestigingscode) van de school nagezocht in de bestanden die beschikbaar zijn gesteld door DUO, peildatum 01-10-2018 (Dienst Uitvoering Onderwijs, 2019). De landelijke percentages zijn ook berekend met behulp van deze bestanden.

Om na te gaan of de steekproef representatief is, is getoetst op verschillen tussen de populatie en steekproef. Door de grote aantallen zijn de verschillen al snel significant, daarom is gekeken naar de effectgrootte, door de coëfficiënt φ te berekenen. Zoals in de laatste kolom van Tabel 5 (pag. 25, Wetenschappelijke verantwoording Diaspel 678) is af te lezen, wijken, behalve voor geslacht, voor alle achtergrondvariabelen de aantallen in de steekproef teveel af van die in de populatie en is de steekproef niet representatief te noemen.

Bij het kalibratieonderzoek is door middel van DIF-analyses (Differential Item Functioning) nagegaan of de itemparameterschattingen als gelijk kunnen worden beschouwd in de verschillende subgroepen die kunnen worden onderscheiden n.a.v. de achtergrondvariabelen. Wanneer dat het geval is, functioneren de items in de subgroepen gelijkwaardig als indicatoren voor de vaardigheid (Reise, 2015). Voor alle items van de volgtoetsen geld dat er geen sprak was van DIF. Omdat de steekproef niet representatief was zijn bij het vaststellen van de relatieve normen steekproefgewichten (4.2.2) toegepast.

Conclusie:
Alhoewel de steekproef op de aspecten regio, denominatie en urbanisatie niet representatief is, wordt hier op adequate wijze voor gecorrigeerd.

Op aspect S1 wordt aan de toetsen Diaspel 678 het volgende oordeel toegekend: '**voldoende'**

S2 In geval van een onvolledig dataverzamelingsdesign: is het design adequaat?

Bevindingen:
Het materiaal dat uitgezet is in het kalibratieonderzoek bevat, naast nieuw materiaal, een selectie vragen uit de openbare referentiesets, een selectie vragen uit de itembank Dia-eindtoets Taalverzorging en een selectie vragen met materiaal dat ook in het nieuw te ontwikkelen Dia-LVS voor het VO voorkomt. Het nieuwe materiaal is verdeeld over 76 itemblokken van 10 vragen. Het materiaal afkomstig uit de openbare referentiesets, Dia-eindtoets en Diaspel VO is verdeeld over 12 itemblokken met ieder 10 vragen (RVE blokken).

Leerlingen uit groep 5 maken opgaven uit 5 itemblokken bedoeld voor groep 5. Leerlingen uit groepen 6, 7 en 8 maken, naast enkele blokken bedoeld voor hun eigen niveau, ook 1 of 2 blok(ken) met items uit een leerjaar lager en 1 of 2 blok(ken) uit de zogenaamde RVE-blokken.

In groep 5 (Tabel 35, Bijlage 7) zijn er zo 18 verschillende toetsen samengesteld (A t/m R), waarbij er overlap bestaat tussen de verschillende toetsen. Voor de groepen 6, 7 en 8 zijn er steeds 24 verschillende toetsen samengesteld (A t/m X), ook hier weer met overlap tussen de verschillende toetsen. De toetsen bestonden uit twee delen, waarbij een evenwichtige verdeling per toetsdeel gemaakt is. Omdat de toetsen op verschillende dagen konden worden afgenomen, hebben niet alle leerlingen beide toetsdelen gemaakt, bijvoorbeeld vanwege afwezigheid door ziekte.

Voor de normeringsonderzoeken zijn items geselecteerd en daarnaast een aantal nieuwe items ontwikkeld. De data van deze normeringsonderzoeken is toegevoegd aan die van het kalibratieonderzoek, om een nieuwe kalibratie uit te kunnen voeren. Elk item in de kalibratie is gemaakt door 405 tot 3192 leerlingen (gemiddeld 1412). De items met de lagere leerlingaantallen betreffen voornamelijk items die niet geselecteerd zijn voor het normeringsonderzoek en dus alleen in het oorspronkelijke kalibratieonderzoek zijn afgenomen. De items die zijn geselecteerd voor de uiteindelijke toetsen zijn gemaakt door 431 tot 3192 leerlingen (gemiddeld 2338).

Opmerking: Omdat het 2-parameter logistisch model is gebruikt bij het kalibratieonderzoek, is dit aantal voldoende. Meestal wordt als vuistregel gehanteerd dat het aantal afnames per item groter moet zijn dan 400 a 500. Daar wordt aan voldaan. Uit de bijlagen kon niet worden achterhaald hoe de verdeling van het aantal afnames over de verschillende items was. Deze extra informatie had nog meer inzicht gegeven in het design.

Conclusie:
Het onvolledige maar 'verbonden' dataverzamelingsdesign is adequaat.

Op aspect S2 wordt aan de toetsen Diaspel 678 het volgende oordeel toegekend: '**voldoende'**

S3 In het geval van een observatie-instrument: is er sprake van een adequate steekproef van observatoren en randvoorwaarden waaronder de observatie wordt uitgevoerd?

Bevindingen:
n.v.t.

Conclusie:
n.v.t.

S4 Er is een handleiding met duidelijke instructies voor de leerkracht over het zo objectief mogelijk uitvoeren en weergeven van de observaties door de leerkracht.

Bevindingen:
n.v.t.

Conclusie:
n.v.t.

Normering

N1.1 Is de standaardbepalingsmethode gemotiveerd en op de juiste wijze uitgevoerd?

Bevindingen:
De toetsen van Diaspel 678 zijn geequivaleerd met de Dia-eindtoets. Omdat de Dia-eindtoets geequivaleerd is met de openbare en niet-openbare referentiesets, zijn de referentiecesuren uit de centrale eindtoets overgebracht op de Diaspel volgtoetsen.

Op basis van de uitkomsten uit het kalibratieonderzoek is de Spelniveau(SN)-schaal ontwikkeld. Hierbij is in eerste instantie gekozen om itemparameters uit de Dia-eindtoets, te fixeren. Hierdoor kwamen de itemparameters op de schaal van het gezamenlijk anker van alle eindtoetsaanbieders te liggen, dat gebruikt wordt bij de verschillende eindtoetsen om tot een gelijke normering voor de referentieniveau 1F en 2F te komen. Vervolgens is door middel van een lineaire transformatie de SN-schaal zo gemaakt dat het gemiddelde gelijk is aan 1000, met een standaarddeviatie van 100. De cesuren voor de referentieniveaus (1F en 2F) zijn vervolgens met behulp van de openbare referentiesets overgezet op de SN-schaal.

Na afloop van het normeringsjaar zijn de scores opnieuw geschat en is de transformatie van theta naar SN niet meer gelijk. Er is daarbij gekozen om de cesuren voor 1F en 2F op de SN-schaal gelijk te houden.

Aangezien de Dia-eindtoets, via het gezamenlijk anker van de eindtoetsaanbieders, is geankerd aan zowel de centrale eindtoets als aan de Diaspel volgtoetsen, kunnen de referentiecesuren uit de centrale eindtoets ook worden overgebracht op de Diaspel volgtoetsen.

Voor alle items uit de referentiesets heeft het College voor Toetsen en Examens itemparameters ter beschikking gesteld. Er is hier gekozen voor de schattingen met het OPLM, omdat het OPLM het meest lijkt op het 2PLM.

Tenslotte zijn de op de openbare referentiesets gebaseerde cesuren gemiddeld met de door de Expertgroep PO verstrekte cesuren die werden gebaseerd op de niet-openbare referentiesets.

Voor het relatief normeren is uitgegaan van de percentielscore die de leerlingen behalen. Daarmee kunnen leerlingen onderling vergeleken worden. Vanwege het niet representatief zijn van de steekproef, is hierbij met steekproefgewichten gewerkt.

Conclusie:
De equivaleringsmethode is gemotiveerd en op de juiste wijze uitgevoerd. Op aspect N1.1 wordt aan de toetsen Diaspel 678 het volgende oordeel toegekend: '**voldoende'.**

N1.2 Zijn de beoordelaars/vakdeskundigen/experts naar behoren geselecteerd en getraind?

Bevindingen:
n.v.t.

Conclusie:
n.v.t.

N1.3 Is er voldoende overeenstemming tussen de beoordelaars?

Bevindingen:
n.v.t.

Conclusie:
n.v.t.

N2.1 Zijn de normgroepen groot genoeg?

Bevindingen:
In Bijlage 6 is af te lezen dat de steekproef omvang voor het groep 6 normeringsonderzoek gelijk was aan (N = 1203), voor het groep 7 normeringsonderzoek gelijk was aan (N = 1271) en voor het groep 8 normeringsonderzoek gelijk was aan (N = 1273).

Conclusie:
De normgroepen zijn groot genoeg. Op aspect N2.1 wordt aan de toetsen Diaspel 678 het oordeel **'voldoende'** toegekend.

N2.2 Zijn de normgroepen representatief?

Bevindingen:
De steekproeven niet representatief voor de populatie. De aantallen in de steekproef wijken te veel af van die in de populatie voor de achtergrondvariabelen regio, urbanisatie en denominatie. Er is door middel van een regressieanalyse nagegaan of achtergrondvariabelen effecten hebben op de vaardigheidsscores van de leerlingen. Voor alle achtergrondvariabelen geldt dat de effecten significant zijn (dit betreffen kleine effecten, waarbij de absolute waarden van Cohen's d liggen tussen de 0,01 en 0,28). Vanwege de gevonden effecten wordt statistische weging toegepast, om vertekening in de relatieve normen te voorkomen,. Met behulp van het R-package anesrake (Pasek, 2018) wordt aan de leerlingen een steekproefgewicht toegekend. Hierbij wordt een maximum van 2 aangehouden, wat volgens het COTAN-beoordelingssysteem (Evers et al., 2010) de maximale acceptabele factor is

Conclusie:
De normgroepen zijn niet representatief. Maar daar wordt op een adequate wijze voor gecorrigeerd. Op aspect N2.2 wordt aan de toetsen Diaspel 678 het oordeel '**voldoende**' toegekend.

N2.3 Zijn de normen correct bepaald?

Bevindingen:
De equivaleringsprocedure is correct beschreven en toegepast.

Conclusie:
De normen zijn correct bepaald. Op aspect N2.3 wordt aan de toetsen Diaspel 678 het oordeel '**voldoende**' toegekend

Betrouwbaarheid

B1 Zijn of worden de betrouwbaarheidsgegevens correct berekend?

Bevindingen:
Bij het berekenen van de betrouwbaarheidsmaten wordt gebruikt gemaakt van de empirische data uit het normeringsonderzoek en van simulaties. Bij het gebruik van een IRT model, wordt onderscheid gemaakt tussen lokale betrouwbaarheid en globale betrouwbaarheid. De lokale betrouwbaarheid is niet gelijk voor iedere respondent, maar hangt af van de relatie tussen de vaardigheidsschatting en de itemparameters van de toets. De globale betrouwbaarheid is de proportie door de vaardigheidsparameters verklaarde variantie. Beide maten zijn uitgerekend met het softwarepakket Lexter. Voor de verschillende meetmomenten worden zowel de gevonden betrouwbaarheden tijdens de normeringsonderzoeken weergegeven, als de betrouwbaarheden voor gesimuleerde toetsdata.

Om te testen of het IRT model bij de data past, zijn er DIF statistieken en de First Order Statistics uitgerekend met het software programma Lexter. In de kalibratieset was sprake van een relevante effectgrootte bij 56 van de 1312 uitgerekende statistieken, hetgeen

duidt op uitstekende modelpassing. Ook bij de voor de volgtoetsen geselecteerde items vielen alle waardes binnen de norm. Op basis hiervan kan geconcludeerd worden dat het IRT model fit bij de data.

Conclusie:
De betrouwbaarheidsgegevens worden correct berekend. Op aspect B1 wordt aan de toetsen Diaspel 678 het oordeel **'voldoende'** toegekend.

B2 Zijn de betrouwbaarheidsgegevens voldoende gezien de conclusies en eventuele beslissingen die met het instrument genomen worden?

Bevindingen:
In Tabel 15 (pag. 36, Wetenschappelijke verantwoording Diaspel 678) worden voor de verschillende meetmomenten zowel de gevonden betrouwbaarheden tijdens de normeringsonderzoeken weergegeven, als de betrouwbaarheden voor gesimuleerde toetsdata. Deze globale betrouwbaarheden liggen allen boven de 0,9 en kunnen daarmee als goed gekwalificeerd worden.

In een simulatiestudie zijn met behulp van de 'ware vaardigheid' en de geschatte vaardigheid de marginal classification accuracy en de accuracy plus/minus 1 berekend. De accuracy plus/minus 1 die werd voorgesteld door Pilliner (geciteerd in Cresswell, 1986; Tomesen, M., Engelen, R. & Hiddink, L., 2019) geeft aan welk percentage leerlingen met hun 'ware vaardigheid' in dezelfde of een direct ernaast liggende percentielgroep valt als de met de toets bepaalde percentielgroep. Voor de accuracy plus/minus 1 stelde Pilliner een minimum streefwaarde voor van 95%. De marginal classification accuracy geeft aan welk percentage leerlingen met zowel hun 'ware vaardigheid' als met hun op de toets geschatte vaardigheid in dezelfde percentielgroep valt. De marginal classification accuracy zou in het ideale geval moeten liggen tussen 75% en 80%, in de praktijk worden meestal waarden tussen de 60% en 70% gevonden (Tomesen, M., Engelen, R. & Hiddink, L., 2019). In de uitgevoerde simulatiestudie varieert de marginal classification accuracy tussen de 71% en de 77%. De accuracy plus/minus 1 varieert tussen 99% en 100%.

Conclusie:
De betrouwbaarheidsgegevens zijn voldoende gezien de conclusies en eventuele beslissingen die met het instrument genomen worden. Op aspect B2 wordt aan de toetsen Diaspel 678 het volgende oordeel toegekend: **'voldoende'.**

Validiteit

V1 Inhoudsvaliditeit: Dragen de items in het instrument bij aan de validiteit van het instrument (hierbij gaat het om aspecten als relevantie, objectiviteit en efficiëntie van de items)?

Bevindingen:
Bij de indeling in domeinen en toetsmatrijzen is de toetsaanbieder uitgegaan van het Referentiekader (Meijerink, 2009) en heeft daarbij gebruik gemaakt van de *Leerstoflijnen begrippenlijst en taalverzorging beschreven*, ontwikkeld door SLO (van der Beek & Paus, 2011). In deze uitwerking van SLO is aangegeven hoe de opbouw van de leerstoflijnen

vormgegeven kan worden over de leerjaren heen. Ook is door de toetsaanbieder rekening gehouden met het moment van aanbod van bepaalde spellingsregels in diverse veelgebruikte taal-/spellingsmethodes. Door goed gebruik te maken van deze bronnen is door de toetsaanbieder gezorgd voor een evenwichtige spreiding over alle domeinen en spellingsregels passend bij de verschillende leerjaren. De verschillende vraagtypen zijn passend bij de domeinen. Het is goed om te lezen dat foute woordbeelden worden voorkomen door geen meerkeuzevragen te stellen voor woordspelling en werkwoordspelling. De items zijn een goede afspiegeling van de te meten doelen. Zie bijlage voor enkele opmerkingen op item-niveau.

Conclusie:
'**voldoende**'.

V2 Constructvaliditeit: Meet het instrument in zijn geheel datgene wat het beoogt te meten?

Bevindingen:
Soortgenootvaliditeit is onderzocht door de correlatie te berekenen tussen scores op de toetsen van de Cito LVS 3.0 spelling en de toetsen van de Diaspel 678. De correlatie is berekend voor 1246 leerlingen en is gelijk aan r=0.82.

Opmerking: De validiteit zou sterker onderbouwd kunnen worden door niet alleen naar de convergente validiteit, maar ook naar de divergente validiteit te kijken.

Opmerking: Om de construct validiteit te onderbouwen zou ook gekeken kunnen worden naar de dimensionaliteit van het construct en de bijbehorende correlatiematrix. Met behulp van IRT of met factoranalyse zou deze dimensionaliteit in kaart gebracht kunnen worden. Als een unidimensioneel IRT model beter fit dan een multidimensioneel model, of als een factoranalyse resulteert in een model met één dominante factor, dan wordt daarmee onderbouwd dat de toetsen één construct meten. Als het instrument niet unidimensioneel zou blijken te zijn, zou een confirmatorische factor- of IRT analyse me een factorstructuur gebaseerd op theoretische overwegingen en de toetsmatrijs de validiteit kunnen onderbouwen.

Conclusie:
De onderbouwing van de constructvaliditeit is erg minimaal. De hierboven gemaakte Opmerkingen verdienen in de toekomst nadrukkelijk aandacht. Ondanks dat wordt op aspect V2 aan de toetsen Diaspel 678 het volgende oordeel toegekend: '**voldoende'.**

Het volg-aspect

Va1 Is er een voldoende empirische onderbouwing van de schaal waarop de groei van een leerling wordt uitgedrukt? Wordt groei op een correcte manier gemeten?

Bevindingen:
Met behulp van de vaardigheid van een leerling op twee verschillende meetmomenten en de bijbehorende schattingsfouten kan er worden bepaald of een leerling significant is gegroeid. Hiervoor kan de Reliable Change Index voor (RCI) gebruikt worden Jacobsen en Truax (1991). De RCI kan dan in combinatie met een inschatting van de onderwijskundige relevantie gebruikt worden om vast te stellen of individuele verandering zowel statistisch als onderwijskundig significant is. De RCI kan gebruikt worden binnen de itemresponstheorie. Omdat de items zijn gekalibreerd op dezelfde schaal heeft dat het voordeel dat het niet noodzakelijk is dat metingen zijn verricht met dezelfde items (Jabrayilov, Emons, & Sijtsma, 2016).

Conclusie:
Er is voldoende empirische onderbouwing van de schaal waarop de groei van een leerling wordt uitgedrukt. Op aspect Va1 wordt aan de toetsen Diaspel 678 het volgende oordeel toegekend: '**voldoende'.**

Va2 Wordt de betrouwbaarheid van de groei op die schaal correct weergegeven?

Bevindingen:
Bij de toetesn van Diaspel 678 wordt gebruik gemaakt van IRT. Daarmee kan de betrouwbaarheid van de groei op de schaal correct weergegeven worden. In de wetenschappelijke verantwoording wordt in Figuur 2 (pag. 39) geïllustreerd wat de schattingsfouten zijn voor de verschillende scores en de verschillende leerjaren.

Conclusie:
De betrouwbaarheid van de groei wordt correct weergegeven. Op aspect Va2 wordt aan de toetsen Diaspel het oordeel '**voldoende'** toegekend.

Va3 Worden er voldoende gegevens verstrekt (aan de gebruiker) over hoe groei geïnterpreteerd dient te worden?

Bevindingen:
Op dit moment kunnen op school-, groeps- en leerlingniveau de resultaten uit Diaspel gehaald worden.

Op leerlingniveau wordt een vaardigheidsscore weergegeven in een tabel en in een grafiek. De vaardigheidsscore geeft het algehele spellingsniveau van de leerling aan. De doorlopende vaardigheidsschaal maakt het mogelijk om de ontwikkeling van de leerlingen over de leerjaren heen te volgen en groei in kaart te brengen. De vaardigheidsscore is gekoppeld aan een kleur op een kleurenliniaal. De kleuren kunnen volgens de toetsaanbieder gezien worden als 'ontwikkelingszones' en geven naast de grafiek een beeld van de ontwikkeling van de leerling. Ook wordt aangegeven wat de streefscore van de leerling is. Deze streefscores duiden het gebied aan waarin de score van de leerling

idealiter ligt op weg naar het uiteindelijk te behalen spellingsniveau. Standaard worden de streefscoregebieden weergegeven rondom de gemiddelde prestaties van de leeftijdsgenoten uit de landelijke steekproef. De streefscore is instelbaar per leerling, eventueel in overleg met ouders. Daarnaast wordt per toetsmoment een percentielscore gegeven, waardoor het mogelijk is om de leerling te vergelijken met andere leerlingen in het land die dezelfde toets op hetzelfde toetsmoment gemaakt hebben. Bovendien wordt aangegeven zodra een leerling referentieniveau 1F of 2F heeft behaald.

Door deze verschillende scores (vaardigheidsscore, streefscore en percentielscore), en het wel of niet behalen van een referentieniveau, over de leerjaren heen op verschillende manieren weer te geven (in een tabel, grafiek met kleurenliniaal) wordt groei van de leerling -ten opzichte van zichzelf en vergeleken met de landelijke normgroep- inzichtelijk gemaakt.

Conclusie:
Er worden voldoende gegevens verstrekt (aan de gebruiker) over hoe groei geïnterpreteerd dient te worden. Op aspect V31 wordt aan de toetsen Diaspel 678 het volgende oordeel toegekend: '**voldoende'.**

Inzicht in leervorderingen

I1 Levert de aanbieder een geschreven toelichting bij de leervorderingen van de leerling die (ook) voor ouders /verzorgers/voogden/docenten begrijpelijk is?

Bevindingen:
Er is informatie voor ouders beschikbaar (bijlage 6) waarin kort wordt uitgelegd wat de verschillende scores op de Diatoetsen betekenen. Door de vaardigheidsscores te koppelen aan kleuren (ontwikkelingszones) in een kleurenliniaal –die loopt van het basisonderwijs tot en met het voortgezet onderwijs- maakt de toetsaanbieder de ontwikkeling van leerlingen nog inzichtelijker voor ouders (en ook voor leerlingen). Samen met de grafiek kan de kleurenliniaal zeker ondersteunend zijn bij het gesprek met de ouders over de ontwikkeling van het kind.

Conclusie:
De aanbieder levert een geschreven toelichting bij de leervorderingen van de leerling die voor ouders/verzorgers/voogden/docenten begrijpelijk is. Op aspect I1 wordt aan de toetsen Diaspel 678 het oordeel '**voldoende**' toegekend.

I2 Is er een evaluatie van de leervorderingen en worden op basis van deze evaluatie vervolgstappen geformuleerd?

Bevindingen:
In de 'Handleiding diaspel basisonderwijs groep 678' wordt toegelicht hoe de docent vanuit de grafiek waarin de groei van de leerling wordt weergegeven door kan clicken naar de toetsvragen die de leerling heeft gemaakt. Per opgave is aangegeven of de leerling de vraag goed of fout heeft beantwoord, en het gegeven antwoord van de leerling is zichtbaar.

Via het bestand (sub)domeinanalyse is te vinden bij welke domein of subdomein de opgave hoort. Op deze manier krijgt de leerkracht zicht op de leervorderingen en kunnen vervolgstappen worden geformuleerd.

Conclusie:
De toetsaanbieder biedt de mogelijkheid om op subdomeinniveau de resultaten van de leerlingen te analyseren en zo een gedetailleerd beeld te krijgen van het niveau van de leerling. Het is vervolgens aan de leerkracht om passende vervolgstappen te formuleren.
Op aspect I2 wordt aan de toetsen Diaspel 678 het oordeel '**voldoende**' toegekend.

Referentieniveaus

R1 Sluit de inhoud van de toets aan op de kennis en vaardigheden zoals omschreven in de referentieniveaus van het betreffende domein (voor toetsen vanaf groep 6)?

Bevindingen:
De volgtoetsen Diaspel zijn genormeerd aan de referentiesets Taalverzorging (College voor Toetsen en Examens, 2016a). Behalve vaardigheidsscores op de kleurenliniaal worden ook de referentieniveaus 1F en 2F gegeven. Vanaf toetsmoment eind groep 6 kan het referentieniveau 1F gegeven worden en vanaf eind groep 7 komt daar referentieniveau 2F bij. Pas vanaf eind 7 bevatten de toetsen voldoende 2F-materiaal om een uitspraak te kunnen doen of het referentieniveau 2F ook echt behaald is door de leerling. In 3.3.1 van de Wetenschappelijke verantwoording Diaspel 678 wordt een toelichting gegeven op wat de precieze verdeling en opbouw van de referentieniveaus over de verschillende toetsmomenten is.

Conclusie:
Diaspel sluit aan bij de referentieniveaus voor Taalverzorging. Op aspect R1 wordt aan de toetsen Diaspel 678 het oordeel '**voldoende**' toegekend.

3. Verzamelstaat

Kwaliteitsaspect	Code	Oordeel
De kwaliteit van de steekproef	*S1*	**Voldoende**
	S2	**Voldoende**
	S3	**n.v.t.**
	S4	**n.v.t.**
Normering	*N1.1*	**Voldoende**
	N1.2	**n.v.t.**
	N1.3	**n.v.t.**
	N2.1	**Voldoende**
	N2.2	**Voldoende**
	N2.3	**Voldoende**
Betrouwbaarheid	*B1*	**Voldoende**
	B2	**Voldoende**
Validiteit	*V1*	**n.v.t.**
	V2	**Voldoende**
Volg-aspect	*Va1*	**Voldoende**
	Va2	**Voldoende**
	Va3	**Voldoende**
Inzicht in leervorderingen	*I1*	**Voldoende**
	I2	**Voldoende**
Referentieniveaus	*R1*	**Voldoende**

4. Literatuurlijst

Cresswell, M. J. (1986). Examination Grades: How many should there be? *British Educational Research Journal*, Vol. 12, No. 1, 37-54.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests.* Utrecht: Nederlands Instituut voor Psychologen.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen. (2009). *Referentiekader taal en rekenen. De referentieniveaus.* Enschede.

Expertisecentrum Nederlands. (2010). *Doorlopende leerlijnen taal basisonderwijs*. Opgehaald van Leerlijnen Taal: http://www.leerlijnentaal.nl/

Jabrayilov, R., Emons, W. H., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, 40(8),559-572.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1), 12–19.

Pasek, J. (2018). anesrake: ANES Raking Implementation. R package version 0.80. https://CRAN.R-Project.org/package=anesrake.

Tomesen, M., Engelen, R., & Hiddink, L. (2019). *Wetenschappelijke verantwoording Begrijpend lezen 3.0 voor groep 8.* Arnhem: Cito.

TULE. (2008). *TULE - Nederlands.* Enschede: SLO.