## 1. Algemene informatie

Algemeen en meetpretentie
De volgtoetsen Diawoord meten de woordenschat 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. 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 Diawoord zijn bedoeld om per meetmoment het niveau van de woordenschat van leerlingen te bepalen. Doordat de vaardigheidsscore (WSN, woordenschatniveau) door middel van IRT-analyses op één schaal is gebracht, is het bovendien mogelijk om groei van leerlingen te volgen, dit is naast de niveaubepaling het tweede primaire gebruiksdoel van de toetsen Diawoord.

Leerlingen krijgen als uitslag een WSN-score. De WSN-score wordt ook weergegeven ten opzichte van de streefscore, die het gebied duidt waarin de score van de leerling idealiter ligt op weg naar het uiteindelijk te behalen woordenschatniveau. Daarnaast wordt per toetsmoment een percentielscore gegeven, die de leerlingen op basis van een landelijke steekproef vergelijkt met leerlingen op hetzelfde moment in de schoolloopbaan.

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 Diawoord zijn genormeerd bij leerlingen uit het reguliere basisonderwijs, voor het speciaal basisonderwijs is geen aparte normering beschikbaar.

Inhoudelijke theoretische inkadering:
Aangezien er voor woordenschat geen Referentiekader is, zijn de toetsen Diawoord hier niet op gebaseerd. Diawoord meet de receptieve kennis van algemene schooltaalwoorden. Bij de toetsmatrijzen is rekening gehouden met de indeling in sluizen volgens de streefwoordenlijst, de verdeling van de vraagsoorten (betekenisvragen en antoniemen) en de verdeling van verschillende woordsoorten (zelfstandige naamwoorden, werkwoorden, bijvoeglijke naamwoorden en overige).

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

- Wetenschappelijke verantwoording Diawoord groep 678, deze bevat informatie over:
  - Uitgangspunten (hoofdstuk 2)
  - Inhoudsverantwoording (hoofdstuk 3)
  - Kalibratie en normering (hoofdstuk 4)
  - Betrouwbaarheid, volgaspect en validiteit (hoofdstuk 5)
  - 7 bijlagen, waaronder items en antwoorden
- Inzage digitale toetsitems (instructie)
- Handleiding Diawoord 678
- Toetsreglement Dia-LVS PO
- Wegwijs in Dia-groeiwijzer

## 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 Liza Kozlowska MA (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 81 scholen mee met leerlingen uit de groepen 5, 6, 7 en 8. In totaal hebben er 20.458 leerlingen meegedaan aan een of meer toetsen. Het geslacht van de leerling is door de scholen opgegeven. Eén school heeft hierover geen gegevens aangeleverd. 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. 22, Wetenschappelijke verantwoording Diawoord 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 Diawoord 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 zowel de Dia-eindtoets 2017 als 2018 en een selectie vragen met materiaal dat ook in het Dia-LVS voor het VO voorkomt. Het nieuwe materiaal is verdeeld over 72 itemblokken van 7 vragen. Het materiaal afkomstig uit de Dia-eindtoets en Diawoord VO is verdeeld over 7 itemblokken met 7 vragen (W-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 1,2 of 3 blok(ken) met items uit een leerjaar lager en 1 blok uit de zogenaamde W-blokken.

In groep 5 (Tabel 26, Bijlage 4) 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.

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 367 tot 4987 leerlingen (gemiddeld 1761). De items met de lagere leerlingaantallen betreffen voornamelijk items die niet geselecteerd zijn voor het normeringsonderzoek en dus alleen in het oorspronkelijke kalibratieonderzoek zijn afgenomen5 . De items die zijn geselecteerd voor de uiteindelijke toetsen zijn gemaakt door 1906 tot 4987 leerlingen (gemiddeld 2827).

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

Op aspect S2 wordt aan de toetsen Diawoord 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:*
Op basis van de uitkomsten uit het kalibratieonderzoek is de WSN-schaal ontwikkeld. Hierbij is in eerste instantie gekozen om itemparameters uit de Dia-eindtoets te fixeren. Vervolgens is de WSN-schaal gemaakt door middel van een lineaire transformatie van theta naar WSN met voor de gebruiker goed hanteerbare scores zonder decimalen. Uiteindelijk is de koppeling met de eindtoets losgelaten en zijn de parameters na afloop van het normeringsjaar opnieuw vrij geschat. De transformatie van theta naar WSN werd vervolgens opnieuw bepaald, zodat de scores op de definitieve WSN schaal voor de scholen die al deelnamen aan het normeringsonderzoek zoveel mogelijk gelijk blijven aan de scores die werden gebruikt tijdens het normeringsonderzoek.

Aangezien er voor woordenschat geen landelijke referentieniveaus beschikbaar zijn, is een absolute normering voor deze toets niet van toepassing en kan er alleen een relatieve normering worden gegeven.

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 Diawoord 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 = 1384), voor het groep 7 normeringsonderzoek gelijk was aan (N = 1414) en voor het groep 8 normeringsonderzoek gelijk was aan (N = 1453).

*Conclusie:*
De normgroepen zijn groot genoeg. Op aspect N2.1 wordt aan de toetsen Diawoord 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 anova's nagegaan of achtergrondvariabelen effecten hebben op de leesvaardigheidsscores van de leerlingen in de verschillende normgroepen. Hoewel er in de verschillende groepen soms significante effecten optraden bleken deze altijd klein. Vanwege de gevonden effecten, wordt om vertekening in de relatieve normen te voorkomen, statistische weging toegepast. 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 Diawoord 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 Diawoord 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 simulaties. Als je gebruik maakt van IRT is de betrouwbaarheid niet gelijk voor iedere respondent. met IRT kan wel een globale betrouwbaarheid geschat worden. De globale

betrouwbaarheid is de proportie door scoring verklaarde variantie en wordt gegeven in de output van het softwarepakket Lexter. Voor de verschillende meetmomenten worden in Tabel 12 zowel de gevonden betrouwbaarheden tijdens de normeringsonderzoeken weergegeven, als de betrouwbaarheden voor gesimuleerde toetsdata.

*Conclusie:*
De betrouwbaarheidsgegevens worden correct berekend. Op aspect B1 wordt aan de toetsen Diawoord 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 12 (pag. 30, Wetenschappelijke verantwoording Diawoord 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 0,82 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 59,7% en de 71,5%. De accuracy plus/minus 1 varieert tussen 97% 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 Diawoord 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:*
Diawoord meet de receptieve kennis van algemene schooltaalwoorden. Deze woorden zijn geselecteerd uit de Streefwoordenlijst (Hacquebord, Albers & Andringa, 2007): een lijst met algemene schooltaalwoorden die voorkomen bij verschillende vakken op school. De

lijst is op basis van frequentiematen en expertbeoordelingen ingedeeld in 'sluizen' van moeilijkheid: van sluis 1 (relatief makkelijk) tot sluis 5 (relatief moeilijk). Aan de hand van deze sluizen is gezorgd voor een opbouw in moeilijkheid tussen de meetmomenten van Diawoord. Uit de tabellen in de wetenschappelijke verantwoording (paragraaf 3.3) is af te lezen dat er vanaf toetsmoment eind 5/midden 6 een duidelijke afbouw te zien in vragen ingedeeld in de lage sluizen en een opbouw in vragen in de hoge sluizen richting toetsmoment midden 8. Behalve met de moeilijkheidsgraad wordt er rekening gehouden met de verdeling van het type woorden over de meetmomenten. In Diawoord zijn verschillende woordsoorten opgenomen: zelfstandige naamwoorden, werkwoorden, bijvoeglijke naamwoorden en overige (bijwoorden, functiewoorden en vaste uitdrukkingen). In de wetenschappelijke verantwoording wordt de verdeling en opbouw van de woordsoorten over de toetsmomenten in tabellen weergegeven, ook ten opzichte van de vooraf gestelde streefpercentages. Tot slot wordt er een overzicht gegeven in de opbouw en verdeling van het type vraag. Diawoord meet woordenschat middels driekeuze-items. Deze bestaan uit betekenisvragen (grotendeels) en vragen naar antoniemen. De betekenisvragen worden aan de hand van passende en minimale contextzinnen gesteld.

De items zijn daarmee een goede afspiegeling van het te meten doel in de verschillende meetmomenten in de verschillende jaargroepen.

*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 woordenschat en de toetsen van de Diawoord 678. Helaas bleken de deelnemende scholen de LVS-toets woordenschat van Cito niet of nauwelijks af te nemen. Aangezien de vaardigheid begrijpend lezen mede afhankelijk is van de woordenschat van leerlingen werd er ook daar een positieve samenhang verwacht. Voor een groot aantal leerlingen (N = 11.289) waren zowel Diawoord gegevens als de gegevens van Diatekst beschikbaar. De correlatie tussen de vaardigheidsscore van beide toetsen blijkt hoog (r = 0,76). Aangezien er voor een aantal leerlingen (N = 169) ook gegevens van de Cito LVS 3.0 begrijpend lezen toets beschikbaar waren, kon ook hiervoor de correlatie met Diawoord worden bepaald (r = 0,75).

Opmerking: De soortgenootvaliditeit 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. 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.

*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 Diawoord 678 het volgende oordeel toegekend: '**voldoende'.**

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

*Bevindingen:*
Bij de toetesn van Diawoord 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. 33) 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 Diawoord het oordeel '**voldoende'** toegekend.

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

*Bevindingen:*
Diawoord is bedoeld om het niveau en de groei van de woordenschat van leerlingen te kunnen vaststellen. Per meetmoment wordt het niveau in woordenschat bepaald met de vaardigheidsscore. Door vaardigheidsscores van de verschillende meetmomenten vast te leggen in een doorlopende vaardigheidsschaal is het mogelijk om scores tussen

meetmomenten met elkaar te vergelijken en daarmee de groei van een leerling over de leerjaren heen te volgen. De vaardigheidsschaal loopt van het basisonderwijs door tot en met het voortgezet onderwijs. Naast de vaardigheidsscore wordt per meetmoment de streefscore weergegeven. Die streefscore duidt het gebied aan waarin de score van de leerling idealiter ligt op weg naar het uiteindelijk te behalen woordenschatniveau. Ook wordt er bij elk meetmoment een percentielscore gegeven, die de leerling (op basis van een landelijke steekproef) vergelijkt met leerlingen die dezelfde toets op hetzelfde moment hebben gemaakt. Hiermee wordt de leerling ingedeeld in een niveaugroep (1-5 of A-E). De vaardigheidsscore, streefscore en percentielscore worden per leerling weergegeven in een tabel. In een grafiek wordt de groei van de leerling ten opzichte van zichzelf en ten opzichte van de streefscore inzichtelijk gemaakt. In de handleiding voor leerkrachten wordt duidelijk gemaakt hoe de verschillende scores en percentielverdelingen geïnterpreteerd moeten worden.

*Conclusie:*
Er worden voldoende gegevens verstrekt (aan de gebruiker) over hoe groei geïnterpreteerd dient te worden. Op aspect V31 wordt aan de toetsen Diawoord 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 een algemene informatiefolder voor ouders beschikbaar (bijlage 2 uit de handleiding), waarin kort wordt weergegeven welke Diatoetsen er zijn, hoe de resultaten zichtbaar worden gemaakt en wat ze betekenen. De indeling van vaardigheidsscores in kleurgebieden, die corresponderen met de kleurenliniaal in de grafiek, kan voor de leerkracht helpend zijn om ouders inzicht te geven in de vorderingen van hun 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 Diawoord 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 diawoord basisonderwijs groep 678' wordt toegelicht hoe de docent vanuit de leerlinggrafiek Diawoord waarin de groei van de leerling wordt weergegeven door kan klikken 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.

Voor Diawoord is er geen subdomein-analyse mogelijk. Er zijn dus geen subscores op onderdelen van woordenschat beschikbaar, bijvoorbeeld subscores op de verschillende

woordsoorten. Dit is ook niet wenselijk; het zou de leerkracht geen bruikbare informatie opleveren waarmee hij of zij aanpassingen kan maken in het onderwijsaanbod. Wel kan de leerkracht op itemniveau inzicht krijgen in de (foute) antwoorden die de leerling heeft gegeven en weet de leerkracht daarmee welke schooltaalwoorden onvoldoende bekend zijn bij de leerling.

*Conclusie:*
Er is enkel op itemniveau inzicht mogelijk in de antwoorden; vervolgstappen worden daarom ook niet geformuleerd. Inzicht in subscores op de verschillende woordsoorten of vraagtype zou de leerkracht geen bruikbare informatie opleveren voor aanpassingen in het onderwijsaanbod. Op aspect I2 wordt aan de toetsen Diawoord 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:*
Aangezien woordenschat niet als apart onderdeel opgenomen is in het Referentiekader, maar enkel als kenmerk van de taakuitvoering bij verschillende taaldomeinen, is Diawoord hier niet op gebaseerd.

*Conclusie:*
**N.v.t.**

## 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* | **n.v.t.** |

## 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.

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.