## 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 PI-dictee is een spellingtoets waarmee de vaardigheid in het schrijven van woorden kan worden onderzocht. Het dictee wordt ingezet als onderdeel van het leerlingvolgsysteem, maar ook voor de individuele diagnostiek van kinderen met spellingproblemen.
Het PI-dictee kent twee versies: A en B.

Meetpretentie

Het PI-dictee is een spellingtoets waarmee de vaardigheid in het schrijven van losse woorden kan onderzocht.

Doelgroep

Doelgroep van het PI-dictee is groep 3 tot en met 8.

Gebruiksdoel en functie

Het PI-dictee is ontwikkeld om antwoord te krijgen op de volgende vragen:

- Wat is het niveau van de spellingvaardigheid?
- Hoe is het beheersingsniveau van de spellingcategorieën?

Het PI-dictee kan op twee manieren worden ingezet:

- Als signaleringsinstrument (LVS): het PI-dictee wordt op vaste momenten afgenomen om de ontwikkeling van de leerling op langere termijn te volgen.
- Als diagnostisch instrument: wanneer het duidelijk is dat een leerling een achterstand in spellingvaardigheid heeft, geeft het PI-dictee informatie over de beheersing van deelvaardigheden en spellingcategorieën.

Inhoudelijke theoretische inkadering:

Het Referentiekader taal en rekenen (2009) en de uitwerking van het referentiekader Nederlandse taal voor het domein begrippenlijst en taalverzorging (Van der Beek & Paus, 2011) geven een concreet beeld van de spellingvaardigheid en spellingcategorieën die de leerling dient te beheersen op de verschillende niveaus aan het einde van de basisschoolperiode. In een leerlijn is vervolgens uitgewerkt in welke stapjes het leerstofaanbod kan worden aangeboden. De CED-groep heeft in 2018 een uitwerking uitgebracht van de leerlijn spelling (Machielsen, 2018).
In de handleiding van het PI-dictee (Geelhoed, Reitsma, Eenshuistra & Berends, 2019) wordt de relatie weergegeven tussen de uitgewerkte CED-leerlijn en de fouten- en spellingcategorieën die bij het PI-dictee gehanteerd worden (pagina 29 t/m 31). Het PI-dictee dekt zo goed als volledig de specifieke spellingcategorieën uit de leerlijn spelling

behorende bij het fundamentele niveau 1F en grotendeels de spellingcategorieën behorende bij het streefniveau 2F.

Inhoud van het toetspakket

Het toetspakket bestaat uit:

- Handleiding en Verantwoording (incl. losse pagina's 29 t/m 31, 116 en 189 t/m 196 in kleur vanwege leesbaarheid)
- Instructieboekje voor de leerkracht
- Toetsbladen voor de leerling
- Informatieblad voor ouders/verzorgers
- Voorbeeld van een Individueel rapport
- Voorbeeld van een Individueel overzicht
- USB-stick met drie Excelbestanden: 1. Normtabellen PI-dictee 2019, 2. Foutenanalyse Versie A en 3. Foutenanalyse Versie B.

## 2. Beoordeling van de kwaliteitsaspecten

*De beoordeling vindt plaats volgens het 'Beoordelingskader voor (reeksen van) toetsen uit 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 mevrouw Jennifer Roubiës MSc (secretaris).*

### *De kwaliteit van de dataverzameling*

S1 Is de steekproef representatief?

*Bevindingen:*
Het PI-dictee is in 2000 genormeerd voor het reguliere basisonderwijs (groep 3 t/m 8). In het schooljaar 2017/2018 is het (geactualiseerde) PI-dictee gehernormeerd. De hernormering van de toets PI-dictee vond plaats in de maanden januari en mei 2018. De scholen namen met alle groepen (i.e., groepen 3 t/m 8) deel aan het eerste normeringsonderzoek in januari 2018. Aan het tweede normeringsonderzoek in mei 2018 namen de scholen alleen met groep 8 deel. Op het eerste normeringsmoment maakten de leerlingen van iedere school versie A of versie B van het PI-dictee. Op het tweede normeringsmoment maakten de leerlingen beide versies. Het betreft dus in totaal drie verschillende versies (steekproef I en steekproef II op het eerste normeringsmoment en steekproef III op het tweede normeringsmoment). De versies A en B zijn parallelvormen en zijn ontwikkeld om de inzetbaarheid van het dictee in een doorlopend leerlingvolgsysteem te vergroten. Meerdere afnames van hetzelfde dictee zou de kans vergroten op onbedoelde leereffecten ('carry-over effecten'). De versies A en B zijn geheel vergelijkbaar en kunnen afwisselend worden afgenomen. Omdat de meeste scholen die zich aanmeldden voor het normeringsonderzoek het (oude) PI-dictee regelmatig afnamen, is om leereffecten te vermijden als voorwaarde gesteld dat het (oude) PI-dictee in het schooljaar van de normering niet bij de leerlingen was afgenomen en/of besproken.

Tabel 6.3 bevat de gegevens van de normeringssteekproef, d.w.z. het aantal scholen, klassen en leerlingen. Het aantal scholen van steekproef I en steekproef was 13 en het aantal scholen van steekproef III was 15 (totaal 41 scholen). Het aantal klassen van steekproef I, steekproef II en steekproef III was respectievelijk 122, 116 en 16 (totaal 249 klassen). Het aantal leerlingen van steekproef I, steekproef II en steekproef III was respectievelijk 2064, 2003 en 286 (totale normeringssteekproef bestond uit 4353 leerlingen). Omdat steekproef I en steekproef II hetzelfde afnameschema hebben en alleen zijn onderscheiden in de versie die bij de leerlingen is afgenomen (en versie A en versie B als equivalent worden beschouwd), worden in de beschrijving van de representativiteit van de normeringssteekproef steekproef I en steekproef II samengenomen. De leerlingen van steekproef III zijn in een gelijke verhouding random toegewezen om danwel met versie A danwel met versie B in de normeringssteekproef te worden vertegenwoordigd. De toetsresultaten en andere gegevens van voornoemde leerlingen vormen de kalibratiesteekproef (zie S1.2).

De representativiteit van de steekproef is onderzocht met betrekking tot de variabelen schoolgrootte, regio, stedelijkheid, status, denominatie en gemiddelde eindtoetsscore op school- en leerlingniveau en de variabelen sekse, leeftijd, migratieachtergrond en

leerlinggewicht alleen op leerlingniveau. Om de representativiteit van de steekproeven te beoordelen werd in de tabellen de effectmaat Cramers V opgenomen, welke een maat van samenhang tussen variabelen op nominaal niveau is. Voor de interpretatie van Cramers V werden tevens de richtlijnen van Cohen vermeld.

Wat betreft schoolgrootte werden de scholen onderverdeeld in kleine scholen (< 125 leerlingen), middelgrote scholen (< 250 leerlingen) en grote scholen (> 250 leerlingen). Tabel 6.4 laat zien dat in de steekproef de leerlingen afkomstig van kleine scholen oververtegenwoordigd zijn. De verdeling van de leerlingen van middelgrote en grote scholen ligt dicht bij de verdeling in de populatie.

Wat betreft regio werden de leerlingen onderverdeeld in de vier regio's (Noord, West, Oost en Zuid) die door het CBS gehanteerd worden. Tabel 6.5 laat zien dat voor steekproeven I en II met name regio Zuid oververtegenwoordigd is en voor steekproef III regio Oost. In Tabel 6.6 wordt ook de verdeling over de regio's van de leerlingen gepresenteerd, waarbij een opsplitsing naar DL (didactische leeftijd) is gemaakt. Tabel 6.6 laat zien dat de verdeling voor ieder normeringsmoment in grote lijnen in overeenstemming is met de verdeling in de totale steekproef. Voor zover er zich afwijkingen in de verdeling voordoen ten opzichte van de populatie, is hiervoor gecorrigeerd door middel van weging (zie paragraaf 6.6).

Wat betreft stedelijkheid werden de leerlingen onderverdeeld in de vijf gradaties van stedelijkheid die door het CBS onderscheiden worden (zeer sterk stedelijk, sterk stedelijk, matig stedelijk, weinig stedelijk, niet stedelijk), waarbij uitgegaan wordt van de bevolkingsdichtheid naar postcode. Tabel 6.8 laat zien dat op leerlingniveau de categorie 'weinig stedelijk' in de steekproeven I en II is oververtegenwoordigd, terwijl leerlingen in de categorieën 'zeer sterk stedelijk' en 'sterk stedelijk' oververtegenwoordigd waren met name in steekproef III en eveneens, maar in mindere mate, ook in steekproeven I en II.

Wat betreft status is de statusscore berekend welke aangeeft hoe de sociale status is van een postcodegebied in vergelijking met andere postcodegebieden in Nederland. De sociale status van een postcodegebied is afgeleid van een aantal kenmerken van het postcodegebied: het gemiddelde inkomen in een postcodegebied, het percentage mensen met een laag inkomen, het percentage laagopgeleiden en het percentage mensen dat niet werkt. Tabel 6.9 laat zien dat de statusscores van de postcodegebieden van de basisscholen in de normeringssteekproef enigszins boven het landelijke gemiddelde liggen, hetgeen met name geldt voor de steekproeven I en II op leerlingniveau. Mogelijk is dit te verklaren doordat de stedelijke scholen in de steekproeven relatief zijn oververtegenwoordigd.

Wat betreft denominatie werden de scholen onderverdeeld in openbaar, protestants-christelijk, rooms-katholiek, reformatorisch en overig. Tabel 6.10 laat zien dat de verdeling van de denominaties in de steekproeven I en II goed aansluit bij de landelijke verdeling van denominaties. In steekproef III zijn de reformatorische denominatie en de kleinere denominaties die in de categorie 'overige' vallen niet vertegenwoordigd.

Wat betreft gemiddelde eindtoetsscore is een complicatie dat sinds basisscholen in 2014 verplicht werden een eindtoets af te nemen, ze kunnen kiezen uit een aantal verschillende eindtoetsen. Ook in de normeringssteekproef hebben de scholen gekozen voor verschillende eindtoetsen. Deze eindtoetsen zijn verschillend geschaald, waardoor de

gemiddelde scores niet met elkaar vergelijkbaar zijn. Om dit op te lossen zijn de gemiddelde scores van de scholen voor elke eindtoets apart getransformeerd naar Z-scores, waarbij elk schoolgemiddelde gewogen is voor het leerlingaantal in groep 8. De aanname daarbij is dat het gemiddelde niveau van de leerlingen die een zogenoemde 'alternatieve' eindtoets hebben gemaakt, niet verschilt van het gemiddelde niveau van de leerlingen die de Centrale Eindtoets hebben gemaakt. In de analyse zijn alleen de leerlingen van groep 8 uit de normeringssteekproef betrokken, omdat de scores op de eindtoets op deze leerlingen betrekking hebben. Tabel 6.11 laat zien dat de gemiddelde Z-score op de eindtoets in de steekproeven I en II nagenoeg overeenkomt met het landelijk gemiddelde. Bij steekproef III is de gemiddelde score op de eindtoets in groep 8 relatief laag. De verschillen duiden op een klein effect. Hoewel het gemiddelde niveau van de leerlingen in de steekproeven I en II overeenkomt met het landelijke niveau, ligt het gemiddelde niveau in steekproef III vermoedelijk iets onder het landelijke niveau.

Wat betreft sekse (jongens en meisjes) is de verdeling in de steekproef nagenoeg gelijk aan die van de verdeling in de landelijke populatie. Op de verschillende normeringsmomenten treden slechts kleine (toevals)fluctuaties op (zie Tabel 6.12).

Wat betreft leeftijd zijn de gemiddelde leeftijden van de leerlingen (bepaald op de dag van de eerste testafname aan de hand van de geboortedatum van de leerling) op elk normeringsmoment zoals verwacht en nemen ook conform de verwachting toe met de opeenvolgende normeringsmomenten (zie Tabel 6.13).

Wat betreft achtergrond van de leerlingen is uitgegaan van de door het CBS gehanteerde definitie: leerlingen met een migratieachtergrond zijn leerlingen van wie tenminste één van de ouders niet in Nederland is geboren. Conform de CBS-definitie is een onderscheid gemaakt tussen leerlingen met een westerse migratieachtergrond en leerlingen met een niet-westerse migratieachtergrond. Tabel 6.14 laat zien dat het totale percentage in de steekproef van leerlingen met een migratieachtergrond 15,4% is, terwijl landelijk het totale percentage 25,9% bedraagt. Verder blijkt uit Tabel 6.14 dat zowel de groep leerlingen met een westerse als met een niet-westerse migratieachtergrond in de steekproef ondervertegenwoordigd is. Ook blijken de afwijkingen in de normeringssteekproef ten opzichte van de populatie voor iedere DL (didactische leeftijd) vrij constant te zijn wanneer we een uitsplitsing maken naar DL (zie Tabel 6.15). De leerlingen in de steekproef met een migratieachtergrond zijn duidelijk ondervertegenwoordigd. Waarschijnlijk heeft dit te maken met het feit dat scholen in sterk stedelijk gebied (waar de meeste leerlingen met een migratieachtergrond wonen) ondervertegenwoordigd waren (met name in steekproef III en eveneens, maar in mindere mate, ook in steekproeven I en II), zoals reeds eerder geconstateerd.

Wat betreft leerlinggewichten is uitgegaan van de opleiding van de ouders. Het gewicht 0.3 wordt toegekend aan leerlingen van wie beide ouders (of de verzorgende ouder) onderwijs op maximaal vmbo basis- of kaderniveau heeft gevolgd, terwijl het gewicht 1.2 wordt toegekend aan leerlingen van wie een van de ouders maximaal basisonderwijs heeft gevolgd, en de andere ouder maximaal vmbo basis of kader. Tabel 6.16 laat zien dat het totaalpercentage leerlingen met een leerlinggewicht in de steekproef 7.0% bedraagt en iets lager is dan het percentage van 8.4% in de populatie. Dit komt met name doordat de 1.2 leerlingen in de steekproef licht zijn ondervertegenwoordigd. Ook als dezelfde

gegevens zijn uitgesplitst naar DL, zien we op de verschillende normeringsmomenten in de steekproef een vergelijkbare verdeling van de leerlinggewichten (zie Tabel 6.17).

Om de normeringssteekproef in overeenstemming te brengen met de verdeling in de populatie is binnen de normeringssteekproeven gewogen voor regio en achtergrond als zijn de belangrijke beschrijvende variabelen waarop de representativiteit van de steekproef kan worden beoordeeld. Er is niet gewogen voor leerlinggewicht, omdat hierbij een belemmering is dat voor 7.0% van de leerlingen in de steekproef het leerlinggewicht onbekend is (zie Tabel 6.16). Er wordt echter van uitgegaan dat wanneer door middel van weging de verdeling van de achtergrond van de leerlingen in de steekproef wordt gecorrigeerd, daarmee impliciet ook de verdeling van de leerlinggewichten in de steekproef wordt gecorrigeerd. Voor de weging is volgens de richtlijnen van de COTAN een maximum aan de gewichten gesteld van 2.00. Tabel 6.18 en Tabel 6.19 laten zien dat door deze beperkingen de verdelingen naar regio en achtergrond van de leerlingen in de steekproef na weging niet helemaal gelijk zijn aan de verdelingen in de populatie, maar er wel heel dicht bij liggen (zowel voor de totale steekproef als voor de verschillende normeringsmomenten). Alleen de leerlingen met een niet-westerse migratieachtergrond blijven licht ondervertegenwoordigd in de steekproef. Uit Tabel 6.20 blijkt dat ook de verdeling van de leerlinggewichten na weging in de steekproef de verdeling van de leerlinggewichten in de populatie zeer dicht nadert, zowel voor de totale steekproef als voor de verschillende normeringsmomenten. Tenslotte wordt op pag. 102 van de 'Handleiding en Verantwoording' nog vastgesteld dat weging nauwelijks invloed had op de overige variabelen waarmee de representativiteit van de steekproef is onderzocht.

In paragraaf 6.8.3 wordt terecht aandacht besteed aan het feit dat de steekproef gestratificeerd is (in paragraaf 6.4 werd de normeringssteekproef immers op schoolniveau beschreven), waardoor er in feite sprake is van een multilevel probleem in de vorm van schoolafhankelijkheid en er mogelijke substantiële design effecten kunnen ontstaan. Er zal dus gecontroleerd moeten worden op schoolafhankelijkheid. Om deze reden werden er ICC's (Intra Klasse Correlaties) berekend per toetsvorm en DL (Didactische Leeftijd) in Tabel 6.25. Over de totale normeringssteekproef bleek er geen sprake van schoolafhankelijkheid (ICC < .05) te zijn. Voor sommige normeringsmomenten bleken de ICC's weliswaar hoger, maar nog steeds laag. Een uitzondering vormde de ICC voor groep 3 (DL = 5), waarvoor een sterk verband werd gevonden (respectievelijk ICC = 0.497 en ICC = 0.432). De ICC's dalen met de leerjaren; in de onderbouw is nog sprake van een lichte schoolafhankelijkheid terwijl deze in de bovenbouw nagenoeg verdwijnt. Met andere woorden: er treedt schoolafhankelijkheid op voor groep 3 (DL = 5), maar dit effect dooft snel uit. Waarschijnlijk is de gevonden schoolafhankelijkheid voor groep 3 (DL = 5) te verklaren doordat scholen verschillen in de onderwijsmaand waarin ze in groep 3 het schrijf- en spellingsonderwijs gaan intensiveren (sommige scholen beginnen daarmee vanaf het begin van groep 3 terwijl andere basisscholen later in het schooljaar beginnen). Voor steekproef III, waarvoor de gegevens aan het einde van het schooljaar zijn verzameld, werd geen schoolafhankelijkheid gevonden (ICC = 0.042).

Bij het opstellen van het continue normeringsmodel (kan worden gebruikt om voor elke maand in het schooljaar via interpolatie te normeren) werd gecontroleerd voor een eventueel effect van de school en de groep. Daartoe werd op de gegevens van de normeringssteekproef het normeringsmodel van Tellegen (gebruikmakend van non-parametrische regressieanalyse) toegepast, zoals beschreven op pag. 103-104, maar is

de school als variabele in de regressieanalyse voor de continue normering opgenomen om te bekijken of deze variabele een toegevoegde waarde heeft bij de voorspelling van de voorspelde normscore. Tabel 6.26 laat de β-gewichten van de schoolvariabele zien in de normeringsmodellen. Deze zijn over de hele breedte zeer laag en in geen geval significant ($p < 0.05$), waaruit de conclusie kan worden getrokken dat schoolafhankelijkheid geen relevante factor is bij het opstellen van het normeringsmodel.

*Conclusie:*
De steekproeven zijn representatief, zijn adequaat gestratificeerd naar schoolgrootte, regio, stedelijkheid, status, denominatie, gemiddelde eindtoetsscore, sekse, leeftijd, migratieachtergrond en leerlinggewicht en geven informatie over hoe de steekproeven zich verhouden tot de populatiewaarden. De procedure voor het samenstellen van de steekproeven is onderbouwd en de omstandigheden waaronder data is verzameld, is redelijk vergelijkbaar met de omstandigheden waaronder de toetsen worden afgenomen. Op aspect S1 wordt aan de toets PI-dictee groep 3 t/m 8 het volgende oordeel toegekend: '**voldoende'**.

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

*Bevindingen:*
De woorden van het PI-dictee zijn gegroepeerd in 9 blokken van elk 15 woorden, die oplopen in moeilijkheid. De blokken zijn oplopend genummerd, waarbij de nummers van de blokken (5, 10, 15 etc.) corresponderen met de didactische leeftijd waarop verwacht wordt dat de betreffende woorden en de daarin voorkomende spellingscategorieën door de gemiddelde leerling beheerst worden. De didactische leeftijd (DL) is een standaardmaat voor het aantal maanden genoten onderwijs vanaf groep 3. Een heel schooljaar telt daarbij als 10 didactische maanden.

In iedere groep is een vast aantal blokken afgenomen en deze blokken zijn zodanig gekozen dat alle vaardigheidsverschillen aan bod konden komen. In Tabel 6.1 van de Handleiding en Verantwoording wordt het normeringsschema weergegeven, d.w.z. de blokken die per groep zijn aangeboden.

Het normeringsschema voor PI-dictee versie A ziet er als volgt uit voor steekproef I:

- In januari 2018 maakt groep 3 blok 5, 10 en 15;
- In januari 2018 maakt groep 4 blok 5, 10, 15, 20 en 25;
- In januari 2018 maakt groep 5 blok 10, 15, 20, 25, 30 en 40;
- In januari 2018 maakt groep 6 blok 15, 20, 25, 30, 40 en 50;
- In januari 2018 maakt groep 7 blok 20, 25, 30, 40, 50 en 60;
- In januari 2018 maakt groep 8 blok 20, 25, 30, 40, 50 en 60;

Het normeringsschema voor PI-dictee versie B ziet er voor steekproef II hetzelfde uit als het normeringsschema voor PI-dictee versie A voor steekproef I:

- In januari 2018 maakt groep 3 blok 5, 10 en 15;
- In januari 2018 maakt groep 4 blok 5, 10, 15, 20 en 25;

- In januari 2018 maakt groep 5 blok 10, 15, 20, 25, 30 en 40;
- In januari 2018 maakt groep 6 blok 15, 20, 25, 30, 40 en 50;
- In januari 2018 maakt groep 7 blok 20, 25, 30, 40, 50 en 60;
- In januari 2018 maakt groep 8 blok 20, 25, 30, 40, 50 en 60;

Daarnaast hebben de leerlingen in steekproef II, die versie B van het PI-dictee in januari 2018 maakten, ten behoeve van de ankering en de berekening van de parallelle betrouwbaarheid, ook nog een aantal blokken van versie A gemaakt en wel de volgende:

- In januari 2018 maakt groep 3 blok 5;
- In januari 2018 maakt groep 4 blok 10 en 15;
- In januari 2018 maakt groep 5 blok 20 en 25;
- In januari 2018 maakt groep 6 blok 20 en 25;
- In januari 2018 maakt groep 7 blok 30 en 40;
- In januari 2018 maakt groep 8 blok 30 en 40;

Het normeringsschema voor PI-dictee versie A en versie B ziet er als volgt uit voor steekproef III:

- In mei 2018 maakt groep 8 blok, 30, 40, 50 en 60 voor zowel versie A als versie B.

De leerlingen van groep 7 en 8 in de steekproef van januari 2018, die versie A van het PI-dictee maakten, hebben daarnaast de referentieset Taalverzorging gemaakt (zie paragraaf 8.5 ('Relatie tussen PI-dictee en referentieniveau').

Het normeringsschema of kalibratiedesign laat zien dat er sprake is van voldoende linking.

Het meetmodel dat voor de kalibratie gebruikt is, is het Rasch-model. Voor de parameterschatting van dit model (i.e., de moeilijkheidsparameter) dient er zowel verticale linking (tussen de blokken voor de verschillende leerjaren) als horizontale linking (tussen versie A en versie B van het PI-dictee) plaats te vinden. Omdat er in de normering een relatief groot aantal blokken is afgenomen (zie Tabel 6.1) en er dus over relatief veel gegevens wordt beschikt voor de verschillende leerjaren, is ervoor gekozen de linking te realiseren door parameters in de schattingsprocedure te fixeren. Deze methode van linking (concurrent calibration with fixed parameters) blijkt uit de literatuur vrij robuust te zijn (Hanson & Béguin, 2002).

Op basis van de geschatte parameters is vervolgens voor iedere combinatie van blokken een correspondentietabel met ruwe scores en vaardigheidsscores (theta's) opgesteld. Omdat in het Rasch-model een lineaire transformatie van de parameters is toegestaan, zijn de parameters zodanig getransformeerd dat de laagst mogelijke vaardigheidsscore rond de nul ligt. De reden hiervoor is uitsluitend dat een vaardigheidsschaal met nul als startpunt voor de gebruiker intuïtief aantrekkelijker is.

Een eerste belangrijke aanname van het gehanteerde Rasch-model is een-dimensionaliteit. Deze aanname veronderstelt voor het PI-dictee dat alle opgaven van de toets een beroep doen op de spellingvaardigheid en niet op een andere, onbedoelde,

vaardigheid. Een onderzoek naar de een-dimensionaliteit van een meetinstrument kan daarom worden opgevat als een onderdeel van het validiteitsonderzoek (zie paragraaf 8.1).

De een-dimensionaliteit van de toets PI-dictee is onderzocht via confirmatieve factoranalyse (SEM) en wel door na te gaan of een een-factormodel past. Zoals eerder geconstateerd wordt in de praktijk het PI-dictee niet in zijn geheel afgenomen, maar volstaat een combinatie van een aantal blokken. In aansluiting op deze afnamepraktijk is ervoor gekozen de een-dimensionaliteit van het PI-dictee te onderzoeken in een combinatie van relevante blokken (zie Tabel 5.1). In Tabel 5.2 worden de passingsmaten van het een-factormodel gepresenteerd. Op basis van de veel gebruikte RMSEA als passingsmaat (welke rekening houdt met het aantal vrijheidsgraden en de steekproefgrootte), kan worden geconcludeerd dat het een-dimensionaliteitsmodel i.h.a. een goede passing blijkt op te leveren voor het PI-dictee. Hoewel voor de blokken 10-15-20 (afnamemoment DL 15-25) en de blokken 25-30-40-50 (afnamemoment DL 35-40) een matige passing geldt, maar niet zodanig dat de aanname van een-dimensionaliteit moet worden verworpen. De passing van een een-factormodel is in lijn met eerdere bevindingen bij een vergelijkbare toets als SVT Spelling, waarbij eveneens een sterk een-dimensionale structuur aan de data ten grondslag bleek te liggen (Braams & De Vos, 2015, zie paragraaf 4.3.1).

Een tweede belangrijke aanname van het gehanteerde Rasch-model is lokale onafhankelijkheid. Wanneer in twee opgaven hetzelfde spellingsprobleem aan bod komt, is het waarschijnlijk dat deze opgaven, behalve dat ze in een factoranalyse laden op een algemene factor, ook specifieke variantie delen. Een dergelijke samenhang tussen opgaven wordt lokale afhankelijkheid genoemd. Deze lokale afhankelijkheid kan men eveneens als een schending van de een-dimensionaliteit beschouwen.

In het geval van blok 5-10-15 versie A en de blokken 25-30-40-50 (afnamemoment DL 35-45) laat Tabel 5.2 een evidente overschrijding van de 0.05-cesuur zien wanneer we naar het 90%-betrouwbaarheidsinterval van de RMSEA kijken. Inspectie van de modificatie-indices laat zien dat de passing van het een-factormodel voor deze combinaties wordt verslechterd doordat de residuele varianties van een aantal opgaven sterk met elkaar samenhangen. Dit betreft slechts enkele opgaven, waaruit geconcludeerd kan worden dat er niet zozeer sprake is van een duidelijke tweede factor, maar eerder schending van de een-dimensionaliteit door het optreden van lokale afhankelijkheid.

Betrouwbaarheid en separatie-index zijn gebruikt om de passing van het Rasch-model te beoordelen. Deze beide passingsmaten kunnen in het Rasch-model zowel voor items als personen worden berekend. De betrouwbaarheid voor personen geeft de betrouwbaarheid van de rangschikking van personen aan op de latente trek, d.w.z. op de vaardigheidsschaal. We kunnen als richtlijn aanhouden dat een betrouwbaarheid van > 0.80 wenselijk is, vergelijkbaar met de klassieke testtheorie en de COTAN-richtlijnen. Vergelijkbaar met de betrouwbaarheid is de separatie-index een maat voor hoe nauwkeurig de Rasch-schaal personen kan rangschikken. De separatie-index is bij voorkeur > 2.00.

Op dezelfde wijze kunnen de betrouwbaarheid en de separatie-index voor items worden berekend. Hierbij is vooral de separatie-index van belang om na te gaan of de items

voldoende spreiding in itemmoeilijkheid hebben en daarmee de latente trek voldoende representeren. De betrouwbaarheid van de items wordt mede bepaald door de steekproefgrootte.

Tabel 5.5 laat zien dat voor alle combinaties van blokken de Rasch-betrouwbaarheid voor zowel personen als items voldoende hoog is (≥ 0.81). Alleen voor de combinatie van blok 15-20-25 geldt zowel voor versie A als voor versie B een lagere betrouwbaarheid voor personen (≥ 0.74). De separatie-indices liggen i.h.a. iets boven de grenswaarde van 2.00, met name voor de combinatie blok 5-10-15. Dit veronderstelt een minder groot vermogen van deze subtoets om personen nauwkeurig te kunnen onderscheiden (vgl. in dit opzicht de accuratesse van de classificatie, paragraaf 7.4). Op basis van de waarden van de Rasch-betrouwbaarheid en separatie-index kan dus geconcludeerd worden dat het Rasch-model past.

Ter onderbouwing van bovenstaande conclusie is ook gekeken naar de passing van de items en wel via de infit- en outfit-statistieken. Infit en outfit zijn beide gebaseerd op het kwadraat van het gestandaardiseerde residu tussen de geobserveerde en de op het model gebaseerde verwachte waarden. Het verschil tussen beide statistieken is dat infit is gewogen naar de iteminformatie en daarom minder gevoelig is voor extreme itemresponsen (outliers) dan outfit. Als vuistregel geldt dat items met een infit of een outfit > 2.00 en < 0.50 niet bijdragen aan de Rasch-schaal. Bijlage 5 van de Handleiding en Verantwoording laat zien dat schendingen van het Rasch-model met name optreden bij de outfit-maat, terwijl voor de infit-maat opgaven met afwijkingen gering zijn.

Klassieke Testtheorie (KTT) en Rasch-modellering kunnen ook worden gecombineerd om de psychometrische kwaliteit van items te beoordelen, waarbij de item-restcorrelaties kunnen worden opgevat als een maat voor het discriminerend vermogen van het item. De COTAN stelt als ondergrens voor item-restcorrelaties > 0.20. Bijlage 5 van de Handleiding en Verantwoording laat zien dat de item-restcorrelaties voor alle opgaven hoog is. Slechts in enkele gevallen vallen de waarden onder de ondergrens = 0.20, maar in geen enkel geval in sterke mate (item-restcorrelatie < 0.10).

Wat betreft de nauwkeurigheid van de parameterschattingen in het Rasch-model kan ook nog gekeken worden naar de standaardfout van de moeilijkheidsparameter, welke niet groter moet zijn dan de standaarddeviatie van de vaardigheidsparameter maal een constante *c*. De COTAN stelt dat $c \leq 0.2$ als goed moet worden aangemerkt en een waarde $c \geq 0.5$ als onvoldoende geldt. Bijlage 5 van de Handleiding en Verantwoording laat zien dat de nauwkeurigheid van de parameterschattingen van de opgaven bevredigend is. Wanneer we de constante *c* daarvoor als criterium nemen, geldt voor de parameterschatting van bijna alle opgaven dat deze als goed kunnen worden beoordeeld.

Samenvattend kan geconcludeerd worden dat er maar enkele slecht passende opgaven zijn in termen van het Rasch-model. Gezien de lengte van de combinatie van blokken (30 tot 60 opgaven) is dit een goed resultaat. Zoals reeds eerder geconstateerd treden de schendingen van het model met name bij de outfit-maat op.

De Rasch-parameters van versie A en versie B zijn op één schaal gezet. Daarmee wordt geïmpliceerd dat beide versies als parallelle toetsen kunnen worden beschouwd. De parallelliteit van beide versies is gebaseerd op inhoudelijke argumenten; de toetsversies zijn zodanig geconstrueerd dat in ieder dicteewoord precies dezelfde spellingsproblemen

worden getoetst (i.e., inhoudsvaliditeit). Dit impliceert dat de opgaven hetzelfde moeilijkheidsniveau hebben, hoewel kleine afwijkingen hierin nooit helemaal zijn uit te sluiten.

Toetsen worden in de itemresponstheorie (IRT) als parallel beschouwd wanneer de informatiefuncties niet afwijken. Er worden dus geen afzonderlijke items vergeleken, maar de meetefficiëntie van de toetsen in hun geheel. In Bijlage 6 van de Handleiding en Verantwoording worden de informatiefuncties van de verschillende subtoetsen gepresenteerd. Daarnaast wordt in Bijlage 6 ook de verhouding van de informatiefuncties afgebeeld, dat wil zeggen: de relatieve efficiëntie. De grafieken laten zien dat, zoals verwacht, de relatieve efficiëntie steeds gelijk aan 1 is of daarbij dicht in de buurt ligt. Hiermee wordt evidentie voor de parallelliteit van de toetsen gegeven.

Deze conclusie is ook af te leiden uit de verschillen tussen de schatting van de Rasch-parameters. In Tabel 5.6 wordt per blok van het PI-dictee het aantal significante verschillen weergegeven. Van de in totaal 135 opgaven van het PI-dictee zijn er 5 opgaven significant voor $p < 0.01$ en 12 opgaven significant voor $p < 0.05$. Tabel 5.6 laat zien dat de parameterinvariantie met name wordt geschonden voor blok 30 en blok 40. Hoewel er een aantal opgaven dus wel verschilt in moeilijkheidsgraad, is dat aantal gegeven het aantal opgaven in de combinaties van blokken aanvaardbaar. Dit wordt ook bevestigd door Figuur 5.1 waarin de Rasch-parameters van beide versies zijn afgebeeld en kan worden geconcludeerd dat de beide lijnen opmerkelijk samenvallen en alleen in enkele gevallen duidelijke verschillen laten zien. Samenvattend kan worden gesteld dat de combinaties van blokken in moeilijkheid en spellingsproblematiek zeer vergelijkbaar zijn en de beide versies van PI-dictee dus als parallel kunnen worden beschouwd.

*Conclusie:*
Het dataverzamelingsdesign dat hier is toegepast, is adequaat. Op aspect S2 wordt aan de toets PI-dictee groep 3 t/m 8 derhalve het volgende oordeel toegekend: '**voldoende'.**

***Normering***

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

*Bevindingen:*
Dit criterium heeft betrekking op absoluut normeren en is dus niet van toepassing (n.v.t.) op de toets PI-dictee groep 3 t/m 8, omdat het hier gaat om relatief normeren.

*Conclusie:*
**n.v.t.**

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

*Bevindingen:*
Dit criterium heeft betrekking op absoluut normeren en is dus niet van toepassing (n.v.t.) op de toets PI-dictee groep 3 t/m 8, omdat het hier gaat om relatief normeren.

*Conclusie:*
**n.v.t.**

N1.3 Is er voldoende overeenstemming tussen de beoordelaars?

*Bevindingen:*
Dit criterium heeft betrekking op absoluut normeren en is dus niet van toepassing (n.v.t.) op de toets PI-dictee groep 3 t/m 8, omdat het hier gaat om relatief normeren.

*Conclusie:*
**n.v.t.**

N2.1 Zijn de normgroepen groot genoeg?

*Bevindingen:*
In Tabel 6.2 van de Handleiding en Verantwoording worden de aantallen leerlingen vermeld in de normeringssteekproef voor steekproef I (versie A), steekproef II (versie B) en steekproef III (versie A en versie B). Voor elk normeringsmoment en elk combinatie blok (blok 5-10-15, blok 5-10-15-20-25, etc.) zijn de toetsgegevens van 327 tot 372 leerlingen gebruikt.

*Conclusie:*
Uitgaande van de COTAN beoordelingscriteria wordt op aspect N1.2.1 aan de toets PI-dictee groep 3 t/m 8 het oordeel '**goed**' toegekend.

N2.2 Zijn de normgroepen representatief?

*Bevindingen:*
De representativiteit van de steekproeven werd in S1 en S2 besproken en daar werd geconstateerd dat de steekproeven representatief waren voor schoolgrootte, regio, stedelijkheid, status, denominatie, gemiddelde eindtoetsscore, sekse, leeftijd, migratieachtergrond en leerlinggewicht.

*Conclusie:*
Op aspect N2.2 wordt aan de toets PI-dictee groep 3 t/m 8 het volgende oordeel toegekend: '**voldoende'.**

***Betrouwbaarheid***

B1 Zijn of worden de betrouwbaarheidsgegevens correct berekend?

*Bevindingen:*
In het Rasch-model als gehanteerd meetmodel geeft de betrouwbaarheid voor personen de betrouwbaarheid van de rangschikking aan van personen op de latente trek, d.w.z. op de vaardigheidsschaal. De Rasch-betrouwbaarheid wordt berekend als de verhouding van de ware score variantie op de latente trek en de geobserveerde variantie op de latente

trek. Deze coëfficiënt wordt in de literatuur onder andere beschreven in het boek 'Measurement Essentials' van Wright en Stone (1999) en vertoont qua interpretatie grote overeenkomst met de betrouwbaarheidscoëfficiënt uit de klassieke testtheorie. Daarnaast is Cronbach's alpha en de parallelle betrouwbaarheid berekend. De Rasch-betrouwbaarheid is een conservatievere manier om de betrouwbaarheid te berekenen dan Cronbachs alpha.

*Conclusie:*
De betrouwbaarheidsgegevens worden correct berekend en op aspect B1 wordt aan de toets PI-dictee groep 3 t/m 8 het oordeel '**voldoende**' toegekend.

B2 Zijn de betrouwbaarheidsgegevens voldoende gezien de beslissingen die met de toets genomen worden?

*Bevindingen:*
Tabel 7.1 en Tabel 7.2 laat voor respectievelijk versie A en versie B de betrouwbaarheden zien voor zowel de Rasch-betrouwbaarheid als Cronbachs alpha (interne consistentie), zoals die in het afnameschema zijn opgenomen. Cronbachs alpha blijkt in alle groepen en voor de verschillende toetsvormen hoog, namelijk > 0.87 voor versie A en > 0.89 voor versie B. Dit duidt volgens de COTAN richtlijnen op een hoge betrouwbaarheid. Daarbij is de toetsvorm blok 5-10 voor DL = 15 niet meegenomen. Deze combinatie van blokken heeft, waarschijnlijk door een plafondeffect van de toets, een relatief lage betrouwbaarheid (versie A = 0.84 en versie B = 0.79). Aangezien voor blok 5-10 slechts een normering is opgesteld voor de periode DL = 5 t/m DL = 8 en de resultaten van blok 5-10 voor DL = 5 wel hoog betrouwbaar zijn (versie A = 0.87 en versie B = 0.89), is deze relatief lage betrouwbaarheid minder bezwaarlijk. Overigens vallen de Rasch-betrouwbaarheden, zoals verwacht, systematisch iets lager uit.

Omdat in het normeringsonderzoek Steekproef III de leerlingen in groep 8 blok 30-40-50-60 van beide versies hebben gemaakt, kan op basis van de correlatie tussen de scores op beide versies een inschatting worden gekregen van de parallelle betrouwbaarheid. Tabel 7.3 laat zien dat de correlatie tussen de vaardigheidsscores van de beide versies gelijk is aan 0.83, gebaseerd op een onderzoeksgroep van 286 leerlingen. Dit duidt op een parallelle betrouwbaarheid die iets lager ligt dan de betrouwbaarheid gebaseerd op Cronbachs alpha, maar is volgens de COTAN richtlijnen wel voldoende hoog. Tabel 7.3 laat ook een klein effect van de groepsgemiddelden zien, waarbij versie A (gemiddelde = 12.09) iets moeilijker is dan versie B (gemiddelde = 12.39). De T-toets voor gepaarde waarnemingen blijkt significant voor $p < 0.01$.

Naast voornoemde betrouwbaarheidscoëfficiënten wordt ook de classificatieaccuratesse besproken als maat voor de betrouwbaarheid van de classificatie op basis van een toetsscore. Een veel gehanteerde niveau-indeling voor schoolvaardigheidstoetsen zijn de niveau-indelingen A t/m E en I t/m V ontleend aan het Cito (zie Tabellen 4.1 en 4.2). De classificatieaccuratesse (CA) wordt dan gedefinieerd als de mate waarin de 'ware' classificaties van de leerlingen overeenkomen met de geobserveerde niveau-indelingen (Lee, 2010). Omdat de toets PI-dictee groep 3 t/m 8 ook bedoeld is om het

voortgangsniveau van de leerling te bepalen, is het belangrijk om te weten hoe betrouwbaar de classificaties van de niveau-indelingen zijn.

Om de classificatieaccuratesse te bepalen wordt in dit geval de methode van Rudner (Rudner, 2001, 2005) gehanteerd, welke relatief simpel is. Deze methode gaat ervan uit dat voor de berekening van de accuratesse de cesuren zijn bepaald voor een op IRT gebaseerde vaardigheidsschaal, zoals het geval is bij het PI-dictee. De benadering veronderstelt dat voor iedere 'ware' θ (= vaardigheidsscore), de geobserveerde θ^ normaal verdeeld is, met een gemiddelde θ en een standaarddeviatie van SE (θ^). Figuur 7.1 van de Handleiding en Verantwoording laat aan de hand van een voorbeeld voor een enkele cesuur ('geslaagd' versus 'gezakt') zien hoe de accuratesse vervolgens wordt berekend uit de grootte van het niet-gearceerde gebied aan de rechterzijde van de cesuur als het gebied dat de waarschijnlijkheid weergeeft dat een leerling correct wordt geclassificeerd als zijnde 'geslaagd'.

In bijlage 8 van de Handleiding en Verantwoording wordt de accuratesse van de niveau-indeling A t/m E en I t/m V voor de verschillende normeringsmomenten gegeven. De accuratesse van de niveau-indelingen varieert tussen 0.57-0.73. De accuratesse is relatief laag, maar deze waarden zijn volgens de auteurs wel gebruikelijk voor toetsen met vier cesuren zoals het PI-dictee (vgl. Deng, 2011, voor vergelijkbare accuratessewaarden voor toetsen met vier cesuren). Uit inspectie van de tabellen in Bijlage 8 valt ten eerste op dat de niveau-indeling A t/m E iets accurater is dan de niveau-indeling I t/m V. Ten tweede valt op dat de zwakke en excellente leerlingen (de randcategorieën) zich over het algemeen accurater laten classificeren in de bij hen passende niveaus. In het middengebied (de middencategorieën) kennen de niveau-indelingen een grotere onzekerheid van veelal maximaal 1 niveau. In sommige gevallen zijn er zelfs substantiële kansen dat een classificatie er twee niveaus naast zit. Zie bijv. Tabel 2 uit Bijlage 8 voor blok 10-15-20 waar een leerling een kans heeft van 8.6% om in niveau A te worden geclassificeerd, terwijl hij/zij in werkelijkheid in niveau C zou moeten worden geclassificeerd.

*Conclusie:*
Met verwijzing naar het beoordelingssysteem van de COTAN kan worden vastgesteld dat de betrouwbaarheid van de toets PI-dictee groep 3 t/m 8 'voldoende' is als mag worden aangenomen dat de toets geen zware consequenties voor de leerlingen heeft en er rekening mee gehouden wordt dat er altijd sprake zal zijn van misclassificaties ter grootte van veelal maximaal 1 niveau (maar in sommige gevallen zelfs substantiële kansen van misclassificaties ter grootte van twee niveaus). Op basis van het voorgaande wordt op aspect B2 aan de toets PI-dictee groep 3 t/m 8 het oordeel '**voldoende**' toegekend.

***Validiteit***

V1 Inhoudsvaliditeit: 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 toetsontwikkelaars hebben een geheel nieuwe toets ontwikkeld om de vaardigheid in spelling van niet-werkwoorden te meten bij leerlingen in groep 3 tot en met 8 van het PO. Zij koppelen afnamemomenten aan de DL (Didactische Leeftijd) van leerlingen. De toets kan als onderdeel van een Leerlingvolgsysteem afgenomen worden, om de vorderingen op het gebied van spellingvaardigheid onafhankelijk van methode en aanbod te meten, maar ook als diagnostisch instrument voor het vaststellen van spellingproblemen. Voor beide instrument-opties worden aanwijzingen voor gebruik gegeven.

De toets bestaat uit 9 blokken, die verdeeld over de schooljaren afgenomen worden. Leerlingen in groep 3 kunnen volgens planning in een LVS setting een toets maken die uit twee blokken bestaat, in groep 4 tot en met 5 worden drie blokken getoetst, en in groep 6 tot en met 8 vier blokken. Deze vorm met opgaven in twee versies geeft voldoende mogelijkheid tot planning van afname. De belasting van leerkrachten en leerlingen door het jaar is hiermee in verhouding tot de opbrengst.

Uit de inhoudelijke verantwoording wordt duidelijk dat de ontwikkelaars zich richten op de in het Referentiekader Taal vastgelegde uitwerkingen van het domein Spelling, subdomein niet-werkwoorden, met gebruikmaking van een aantal passende aanvullende bronnen om de doorgaande lijn in ontwikkeling van leerlingen op dit vlak te volgen.

In de inhoudelijke verantwoording is geen aandacht gegeven aan onderbouwing van de keuze voor enkel actieve spelling na een mondelinge prompt (zin met context); het gaat in alle gevallen om een (klassikaal) zinsdictee en niet bijvoorbeeld (ook) met gebruikmaking van (digitale) meerkeuzeopgaven. Hierbij zou wellicht ook nog een onderbouwing hebben gepast van de keuze om het dictee door de leerkracht zelf te laten voorlezen – een ingesproken cd met de dictees zal de betrouwbaarheid van de afname verhogen.

De opgaven en de vraagstelling zijn doorgaans helder, passend en eenduidig. Er is echter een aantal opgaven met gedateerd woordgebruik en een aantal opgaven dat sterker geformuleerd zou kunnen worden.

Versie A, blok 20:

item 9: ‘*geeuw* is een teken van slaap’ > nee, een teken van slaperigheid, moeheid. Suggestie: ‘slaap hebben’.

item 12: ‘gelijk’ in betekenis ‘meteen, direct’ is regionaal meer/minder gebruikelijk (al is het niet officieel ‘fout’, het kan regionaal gekleurd zijn.)

Versie A blok 25:

item 3: aardig > ‘het is toch nog aardig gelukt’ > dit gebruik als bijwoord is redelijk idiomatisch. Suggestie: ‘gewoon’ als bijvoeglijk naamwoord inzetten.

Versie A, blok 30:

item 1 'vind je dit een moeilijk dictee?' > Beoordelaar is geen voorstander van deze zin, omdat hij (als enige, dus opvallend anders,) over de taak zelf gaat.

Versie B, blok 20:

item 13 'Wat een schrik! Mijn hond steekt zomaar de weg over.' > waarom 'schrik' als zelfstandig naamwoord bevraagd, terwijl het als een werkwoord veel frequenter voorkomt? Het oogt heel gedateerd zo, de opgave. Voor de bevraagde spellingcategorie (sch+medeklinker) zijn ook doelwoorden, zelfstandig naamwoorden, te vinden die frequenter als zodanig gebruikt worden.

Versie B, blok 30:

item 1: formeel, gedateerd taalgebruik 'in het huwelijk treden'. Suggestie: ze vierden hun zilveren huwelijk, want ze zijn 25 jaar getrouwd.

Versie B, blok 50:

3. 'ontzaglijk kwaad worden' > er is een wat herkenbaarder woord in deze categorie te vinden dan het infrequente 'ontzaglijk'. Dat woord is onbedoeld moeilijk.

*Conclusie:*
Op aspect V1 wordt aan de toets PI-dictee groep 3 t/m 8 het oordeel '**voldoende**' toegekend.

V2 Constructvaliditeit: Meet de toets in zijn geheel datgene wat hij beoogt te meten?

*Bevindingen:*
In Hoofdstuk 8 van de Handleiding en Verantwoording, getiteld 'Begripsvaliditeit', wordt een psychometrische onderbouwing gegeven van de begripsvaliditeit van de toets PI-dictee groep 3 t/m 8. Het onderzoek heeft betrekking op de volgende aspecten, welke kunnen worden opgevat als enkele argumenten die pleiten voor de begripsvaliditeit van het PI-dictee groep 3 t/m 8:

- Een-dimensionaliteit en psychometrische kwaliteit van de items. Voor dit onderzoek wordt verwezen naar kwaliteitsaspect S2, waar de passing van het Rasch-model als gehanteerd meetmodel besproken werd en de een-dimensionaliteit (een belangrijke aanname in het Rasch-model) werd nagegaan via confirmatieve factoranalyse (SEM) door de factoriële structuur van het PI-dictee groep 3 t/m 8 te onderzoeken. Geconstateerd werd dat het Rasch-model een bevredigende passing liet zien en ook werd er voldoende empirische evidentie gevonden voor een een-factormodel. Er kan dan ook geconcludeerd worden dat aan de noodzakelijke voorwaarde voor begripsvaliditeit wordt voldaan, omdat bij verschillende combinaties van toetsblokken het een-factormodel goed bleek te passen. Alle opgaven van de toets PI-dictee groep 3 t/m 8 doen dus een beroep op de spellingvaardigheid en niet op een andere, onbedoelde, vaardigheid. De psychometrische kwaliteit van de items werd ook nagegaan in kwaliteitsaspect S2

(o.a. via de item-restcorrelaties van de klassieke testtheorie) en deze voldeden i.h.a. aan de COTAN richtlijnen.

- Ontwikkeling van de spellingvaardigheid. Grafiek 8.1 van de Handleiding en Verantwoording laat zien dat de spellingvaardigheid zich op een nagenoeg lineaire wijze ontwikkelt wanneer gekeken wordt naar de gemiddelde vaardigheidsscores op de onderscheiden normeringsmomenten. Wel is er sprake van een lichte afvlakking van de lineariteit na het omslagpunt halverwege groep 7 (DL = 45). Grafiek 8.1 geeft aan dat de spellingvaardigheid zich gedurende de gehele basisschool blijft ontwikkelen en dat er dus maar in beperkte mate een afvlakking van de leergroei optreedt, zoals soms geconstateerd kan worden bij toetsen voor technisch lezen. Een uitzondering hierop vormt de laatste periode in groep 8 (vanaf DL = 55). In deze periode blijft de spellingvaardigheid nagenoeg constant. Vermoedelijk speelt hierbij een rol dat nadat de eindtoets in groep 8 is afgenomen, het onderwijs in de spellingvaardigheid in de regel minder intensief is.

- Groepsverschillen. Conform de verwachting blijken meisjes op de toets PI-dictee groep 3 t/m 8 het over het algemeen iets beter te doen dan jongens in alle groepen van het basisonderwijs. Dit verschil is echter niet significant en het gaat hierbij om een klein effect, met uitzondering van groep 5, waar een relatief groot effect optreedt (8.29 versus 8.11 en Cohens d = -0.31; zie Tabel 8.1). Ook wordt geconstateerd dat de gevonden effectgroottes bij het PI-dictee in overeenstemming zijn met de gevonden effectgroottes in het normeringsonderzoek van SVT Spelling, een vergelijkbare spellingstoets (Braams & De Vos, 2015, par. 7.5.1).

  Zoals verwacht doen leerlingen met een Nederlandse achtergrond het beter op het PI-dictee dan leerlingen met een niet-westerse migratieachtergrond. Dit verschil treedt eveneens in alle groepen van het basisonderwijs op, maar betreft ook een klein effect en de verschillen zijn wederom niet significant (zie Tabel 8.2). Tabel 8.3 laat zien dat er geen eenduidige verschillen worden gevonden tussen de leerlingen met een westerse migratieachtergrond en leerlingen met een Nederlandse achtergrond. Over de totale onderzoeksgroep doen de leerlingen met een westerse migratieachtergrond het zelfs iets beter dan leerlingen met een Nederlandse achtergrond (9.36 versus 8.96 en Cohens d = -0.13; zie Tabel 8.3), maar dit verschil is wederom niet significant.

  Tabel 8.4 laat tenslotte zien dat de verschillen tussen leerlingen met een dyslexieverklaring en leerlingen zonder dyslexieverklaring, zoals verwacht, zeer geprononceerd en significant zijn. Over de totale steekproef in de bovenbouw (in de onderbouw zijn om begrijpelijke redenen nog relatief weinig leerlingen met een dyslexieverklaring) presteren de leerlingen met een dyslexieverklaring gemiddeld een sd lager dan de leerlingen zonder een dyslexieverklaring. Wanneer we uitsplitsen naar leerjaar (groep 6, groep 7 en groep 8) zijn de verschillen zelfs nog groter: de leerlingen met een dyslexieverklaring presteren zo'n anderhalf sd onder het gemiddelde niveau van leerlingen zonder een dyslexieverklaring.

- Vraagonzuiverheid. Onderzocht is vraagonzuiverheid (of kortweg DIF = Differential Item Functioning) ten opzichte van sekse en ten opzichte van leerlingen met een Nederlandse en een niet-westerse migratieachtergrond. Daarbij werd gebruikgemaakt van de Mantel-Haenszel-procedure en de ETS-classificatie (A: 'geen tot verwaarloosbaar effect'; B: 'licht tot matig effect'; C: 'matig tot ernstig effect'). Zie Zwick, 2012, voor een bespreking en evaluatie van deze categorisatie. Uit Tabel 8.5 t/m Tabel 8.8 blijkt dat, gegeven het grote aantal opgaven waarvoor op vraagonzuiverheid is getoetst, het aantal opgaven met een verdenking van vraagonzuiverheid gering is. De auteurs stellen daarom dat, wanneer het PI-dictee als LVS wordt ingezet, het effect op de totaalscore voor zowel sekse als migratieachtergrond dan ook minimaal zal zijn. De inhoudelijke interpretatie van de vraagonzuiverheid blijft echter van belang, omdat door kanskapitalisatie de kans wordt vergroot (door de vele uitgevoerde toetsen) dat opgaven worden gevonden waarbij de statistische analyse vraagonzuiverheid laat zien. Wanneer er indicaties van vraagonzuiverheid zijn, is het dus van belang dit ook te kunnen duiden. Tabel 8.7 laat zien dat in versie B van het PI-dictee een opgave duidelijk vraagonzuiver te zijn ten opzichte van sekse en wel vraagpartijdig ten opzichte van meisjes ('De baby had een *schattig* jurkje aan'). Verder laat Tabel 8.8 zien dat er in versie A drie opgaven zijn gevonden die relatief moeilijker zijn voor leerlingen met een niet-westerse achtergrond ('*Gelukkig* was de regenbui heel snel voorbij', 'Deze wol is van zeer goede *kwaliteit*' en 'De *aanwezigheid* van de minister werd erg op prijs gesteld'), maar de interpretatie ervan is lastig.

- Relatie tussen PI-dictee en referentieniveau. Voor het PI-dictee zijn ook de cesuren bepaald voor de referentieniveaus 1F en 2F (zie ook aspect R1). Tabel 8.11 laat zien dat in groep 8 ongeveer 85% van de leerlingen in groep 8 het fundamentele niveau 1F (het niveau dat een leerling aan het eind van de basisschool minimaal moet halen) hebben behaald en ongeveer 50% het streefniveau 2F (een niveau hoger dan het fundamentele niveau voor leerlingen die meer kunnen) hebben gehaald. Deze percentages komen overeen met de verwachtingen.

*Conclusie:*
Op aspect V2 wordt aan de toets PI-dictee groep 3 t/m 8 het oordeel '**voldoende**' toegekend.

***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 adequate manier gemeten?

*Bevindingen:*
In paragraaf 5.2 van de Handleiding en Verantwoording wordt de kalibratiestudie beschreven. De kalibratiestudie bestaat uit twee stappen. In de eerste stap werd met confirmatieve factoranalyse (SEM) de aanname van een-dimensionaliteit onderzocht, waarvoor voldoende empirische steun werd gevonden. In de tweede stap werd de passing van het Rasch-model als gehanteerd meetmodel onderzocht en werd op basis van

gepresenteerde statistieken over de passing geconcludeerd dat de aanname van passing van het eendimensionale Rasch-model gerechtvaardigd is. Het voorgaande impliceert dat men beschikt over een (vaardigheids)schaal (vertaling van ruwe scores naar een vaardigheidsscore) voor PI-dictee groep 3 t/m 8 waarop de items (en leerlingen) van de beide parallelle toetsversies A en B kunnen worden afgebeeld, welke het mogelijk maken om zowel het niveau als de groei van de leerlingen vast te stellen gedurende het gehele basisonderwijs.

Voor de vraag of groei op een adequate manier gemeten wordt, wordt verwezen naar het 'Individueel rapport' en het 'Individueel overzicht' van Hoofdstuk 4 van de Handleiding en Verantwoording. In het 'Individueel rapport' worden onder de normaalverdeling de normscores grafisch weergegeven. In het rapport op pag. 46 van de Handleiding en Verantwoording wordt de toets vermeld die een fictieve leerling gemaakt heeft, in dit geval blok 30-40-50-60 van versie A, op normeringsmoment DL = 49, de behaalde ruwe score van 49, de percentielscore van 65, de twee niveau-indelingen (Niveau I-V: II; Niveau A-E: B), de DLE van 51 (didactisch leeftijdsequivalent welke aangeeft bij hoeveel maanden onderwijs een gemiddelde leerling een bepaalde prestatie levert), de vaardigheidsscore van 12 met bijbehorend 90% betrouwbaarheidsinterval en de T-score van 53.8 (een normscore met een gemiddelde van 50 en een standaarddeviatie van 10, welke vaak wordt gehanteerd in het dyslexieonderwijs). Door de vaardigheidsscore van de leerling af te zetten tegen eerdere toetsresultaten kan men constateren of er sprake is van vooruitgang. Als de betrouwbaarheidsintervallen van de laatste en de voorgaande toetsafname elkaar niet overlappen, kan men bijna zeker concluderen dat er sprake is van werkelijke vooruitgang. Daarnaast kan men ook percentielscores, de gebruikelijke twee niveau-indelingen en DLE scores van opeenvolgende toetsafnames vergelijken om groei vast te stellen (zie ook aspect Va2 en aspect Va3). Tenslotte kunnen we uit Grafiek 8.1 op pag. 120 van de Handleiding en Verantwoording de ontwikkeling aflezen van de gemiddelde vaardigheidsscore van het PI-dictee voor alle normeringsmomenten, welke zich gedurende de gehele basisschool nagenoeg lineair blijft ontwikkelen.

*Conclusie:*
Er is voldoende empirische onderbouwing van de schaal waarop de groei van een leerling wordt uitgedrukt en de groei wordt op een adequate manier gemeten. Op aspect Va1 wordt aan de toets PI-dictee groep 3 t/m 8 het oordeel '**voldoende**' toegekend.

<u>Va2 Wordt de betrouwbaarheid van de groei op die schaal adequaat weergegeven?</u>

*Bevindingen:*
De (ruwe) toetsscores van de leerling zijn op een (gekalibreerde) vaardigheidsschaal (Rasch-schaal) geplaatst aan de hand waarvan de ontwikkeling gevolgd kan worden. Voor een interpretatie van de vaardigheidsscore kan de gemiddelde groei van de vaardigheidsscore behulpzaam zijn. Met behulp van de 90% betrouwbaarheidsintervallen kan worden vastgesteld of er sprake is van werkelijke groei. De betrouwbaarheid van de groei op die schaal wordt adequaat weergegeven.

*Conclusie:*
Op aspect Va2 wordt aan de toets PI-dictee groep 3 t/m 8 het oordeel '**voldoende**' toegekend.

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

*Bevindingen:*
In hoofdstuk 4 van de Handleiding en Verantwoording wordt onderscheid gemaakt tussen 'niveau' en 'groei', hetgeen wordt onderbouwd met diverse rapportagemogelijkheden (individueel rapport, individueel overzicht, individuele foutenanalyse en foutenanalyse groep). Ook wordt in hoofdstuk 4 van de Handleiding en Verantwoording aan de hand van een individueel leerlingrapport de interpretatie van groei op een duidelijke manier beschreven. In Bijlage 4 wordt de verwachte gemiddelde groei in vaardigheidsscore na een maand en de cumulatie van de gemiddelde groei na meerdere maanden in tabelvorm gerapporteerd. Op pag. 52-54 wordt aan de hand van voorbeelden toegelicht hoe groei geïnterpreteerd dient te worden in deze tabellen.

*Conclusie:*
Op aspect Va3 wordt aan de toets PI-dictee groep 3 t/m 8 het oordeel '**voldoende**' toegekend.

***Inzicht in leervorderingen***

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

*Bevindingen:*
In een bijgeleverd 'Informatieblad voor ouders/verzorgers over het PI-dictee' wordt informatie gegeven over:

- Wat toetst het PI-dictee?
- Wat gebeurt er na de toets?
- Wat betekenen de verschillende normscores? (vergezeld met een stukje uit een Individueel rapport voor een fictieve leerling met verschillende normscores).
- Referentieniveau.
- De ontwikkeling van het niveau, geïllustreerd aan de hand van de ontwikkeling van een fictieve leerling (groeicurve PI-dictee) met wie het de goede kant uit gaat.

*Conclusie:*
De toelichting bij de leervorderingen van de leerling in de vorm van het bijgeleverde Informatieblad is zodanig geschreven dat het (ook) begrijpelijk is voor de ouders/voogden/verzorgers. Op aspect I1 wordt aan de toets PI-dictee groep 3 t/m 8 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 toetsontwikkelaars maken duidelijk dat ze zich richten op de in het Referentiekader Taal vastgelegde uitwerkingen van het domein Spelling, subdomein niet-werkwoorden, met gebruikmaking van een aantal passende aanvullende bronnen om de doorgaande lijn in ontwikkeling van leerlingen vanaf groep 3 op dit vlak te volgen en duiden.

*Conclusie:*
Op aspect R1 wordt aan de toets PI-dictee groep 3 t/m 8 het oordeel '**voldoende**' toegekend.

## 3. Verzamelstaat

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

## 4. Literatuurlijst

Geelhoed, J., Reitsma, P., Eenshuistra, R. & Berends, I. (2019). *PI-dictee, handleiding en verantwoording.* Amsterdam: PI Research & Boom uitgevers.