Wetenschappelijke verantwoording Taalverzorging groep 6 tot en met 8

Ron Engelen, Pauline Roumans, Marie-Anne Keizer en Ineke Jongen

zeker weten

Wetenschappelijke verantwoording

Taalverzorging groep 6 tot en met 8

Ron Engelen
Pauline Roumans
Marie-Anne Keizer
Ineke Jongen

Cito
Arnhem, 2016

Inhoud

1 Inleiding 5

2 Uitgangspunten van de toetsconstructie 7
2.1 Meetpretentie 7
2.2 Doelgroep 7
2.3 Gebruiksdoel en functie 8
2.4 Theoretische inkadering 12
2.4.1 Inhoudelijk 12
2.4.2 Psychometrisch 17
2.4.2.1 Opgavenbanken 17
2.4.2.2 Het gehanteerde meetmodel 20

3 Beschrijving van de toets 25
3.1 Opbouw en structuur van de toets 25
3.2 Inhoudsverantwoording 28
3.2.1 Domeinbeschrijving en uitwerking in taalverzorgingscategorieën 28
3.2.2 Itemconstructie, onderzoeken en selectie van opgaven voor de toetsen Taalverzorging 33
3.3 Statistische beschrijving 36
3.3.1 Itemkenmerken: moeilijkheidsgraad en interne consistentie 36
3.3.2 Verdeling van de ruwe scores 39

4 Kalibratie en normering 43
4.1 Opzet en verloop van het kalibratie-, normerings- en referentie-onderzoek 43
4.2 Samenstellen van de normeringssteekproef en representativiteit 48
4.3 Kalibratie 52
4.3.1 De kalibratieprocedure 52
4.3.2 Resultaat van de kalibratieprocedure: modelfit 53
4.4 Normeringsresultaten 57
4.5 Referentie-onderzoek: het bepalen van de cesuur 61

5 Betrouwbaarheid en meetnauwkeurigheid 63
5.2 Betrouwbaarheid: bepalen resultaten 63
5.3 Lokale betrouwbaarheid en meetnauwkeurigheid 66

6 Validiteit 79
6.1 Inhoudsvaliditeit 79
6.2 Begripsvaliditeit 79
6.2.1 Unidimensionaliteit 80
6.2.2 Itemkwaliteit 80
6.2.3 Convergente en discriminante validiteit 82
6.2.3.1 Samenhangen met andere toetsen 82
6.2.3.2 Samenhangen tussen deelvaardigheden 85
6.3 Itembias 86
6.4 Verschillen tussen relevante subgroepen 87

7 Samenvatting 91

8 Literatuur 93

Bijlagen 97
1 Referentieniveaus Begrippenlijst en Taalverzorging 98
2 Moeilijkheid van opgaven per jaargroep en taak in de toetsen Taalverzorging voor groep 6, 7 en 8 101
3 Categorieënoverzichten 103
4 Voorbeeldopgaven 105
5 P50- en P80-kanspunten van de opgaven in de toetsen voor groep 6, 7 en 8 in relatie tot de gemiddelde vaardigheidsscore voor de afnamemomenten 110

1 Inleiding

Deze wetenschappelijke verantwoording heeft betrekking op de LVS-toetsen Taalverzorging voor de groepen 6 tot en met 8. De toetsen Taalverzorging maken deel uit van het Cito Volgsysteem primair en speciaal onderwijs en zijn bestemd voor leerlingen in de groepen 6 t/m 8 in het primair onderwijs. Het betreft papieren toetsen voor alle leerjaren.

Deze verantwoording biedt tezamen met de inhoud van het toetspakket Taalverzorging voor groep 6 tot en met 8 alle informatie die nodig is voor een snelle en efficiënte beoordeling van de kwaliteit van het betreffende meetinstrument. Het genoemde materiaal maakt een beoordeling van de toetsen Taalverzorging mogelijk op de volgende aspecten:

- Uitgangspunten van de toetsconstructie;
- De kwaliteit van het toetsmateriaal;
- De kwaliteit van de handleiding;
- Normen;
- Betrouwbaarheid;
- Validiteit.

Het laatstgenoemde aspect betreft alleen begripsvaliditeit en géén criteriumvaliditeit. Omdat de toetsen van het Cito Volgsysteem primair en speciaal onderwijs niet bedoeld zijn voor 'voorspellend gebruik' is criteriumvaliditeit niet van toepassing.

Het voorliggende document heeft met name betrekking op de uitgangspunten van de constructie (de hoofdstukken 2 en 3), de normen (hoofdstuk 4), de betrouwbaarheid en meetnauwkeurigheid (hoofdstuk 5) en de begripsvaliditeit (hoofdstuk 6) van de toetsen Taalverzorging voor de groepen 6 tot en met 8.
De kwaliteit van het toetsmateriaal en de handleiding is te bepalen door kennis te nemen van de inhoud van het toetspakket.

2 Uitgangspunten van de toetsconstructie

2.1 Meetpretentie

Taalverzorging (d.w.z., spelling en interpunctie) is van toepassing wanneer gesproken taal omgezet wordt in geschreven taal. Om woorden en zinnen correct te schrijven, moeten leerlingen spelling- en interpunctie-regels kunnen toepassen. Bij het schrijfproces maakt de leerling ook gebruik van grammaticale kennis; deze is echter ook van toepassing bij de mondelinge productie en dus niet exclusief voorbehouden aan het schrijfproces.
De toetsen in het toetspakket Taalverzorging van het Cito Volgsysteem primair en speciaal onderwijs zijn bedoeld om vast te stellen hoe goed een leerling de juiste spelling- en interpunctieregels kan toepassen en hoe de grammaticale kennis van de leerling zich in de loop van de jaren ontwikkelt.
Het vaststellen van de vaardigheid in de verschillende taalverzorgingscomponenten gebeurt door de leerling beslissingen te laten nemen over getoonde spellings- en interpunctiemogelijkheden. De spellingregels zelf worden niet expliciet bevraagd. De leerling laat indirect zien dat hij of zij de spellingregels beheerst door de correct geschreven woorden (spelling) en zinnen (interpunctie) te herkennen. Grammaticale kennis behelst het kennen van basale grammaticale begrippen. Zie voor een verdere beschrijving van de vier deelgebieden paragraaf 2.4.1. In de toetsen voor groep 6 wordt enkel de spellingvaardigheid van niet-werkwoorden getoetst. Werkwoordspelling komt in het onderwijs namelijk pas vanaf groep 7 uitgebreid aan bod.

2.2 Doelgroep

De toetsen Taalverzorging voor groep 6, 7 en 8 van het Cito Volgsysteem primair en speciaal onderwijs zijn bestemd voor en genormeerd bij leerlingen in groep 6, 7 en 8 van het Nederlandse basisonderwijs.
De populatieparameters voor de toets zijn zowel op het midden als op het einde van het schooljaar bepaald. Desgewenst kan de toets ook op een ander moment in het schooljaar worden afgenomen, maar dat maakt het moeilijker om uitspraken te doen over het niveau van de leerling ten opzichte van andere leerlingen in Nederland.
De toetsen zijn ook geschikt voor leerlingen in het speciaal basisonderwijs en het speciaal onderwijs cluster 2 en 4. Voor deze groepen speciale leerlingen zijn geen afzonderlijke normen vastgesteld. De toetsresultaten van deze leerlingen worden geïnterpreteerd met behulp van de gemiddelde vaardigheidsscores voor leerlingen uit het reguliere onderwijs. Voor deze leerlingen gelden namelijk dezelfde kerndoelen als voor leerlingen in het basisonderwijs, met dien verstande dat leerlingen in het speciaal (basis)onderwijs meer tijd krijgen om de kerndoelen te bereiken. Deze leerlingen kunnen én moeten dus langs dezelfde meetlat gehouden worden als de ‘reguliere’ leerlingen. De leerlingen in het regulier basisonderwijs waarop de normering gebaseerd is, vormen daarmee ook voor de leerlingen in het speciaal (basis)onderwijs een correcte referentiegroep.
De toetsen Taalverzorging voor de groepen 6, 7 en 8 kunnen ook gebruikt worden voor leerlingen in andere leerjaren die werken op een lager niveau dan de betreffende groep. Zo kan bijvoorbeeld een leerling van groep 7 of 8 die functioneert op het niveau van groep 6 de toets van die jaargroep maken. In de handleiding is toegelicht hoe dit toetsen op maat, met behulp van vaardigheidsscores, in zijn werk gaat. Voor leerlingen met een ontwikkelingsachterstand en/of extra onderwijsbehoeften zijn in de handleiding extra aanwijzingen opgenomen. Voor deze leerlingen zijn alternatieve rapportageformulieren ontwikkeld.

De toetsen kunnen worden afgenomen door de leerkracht of IB’er. We gaan daarbij uit van de professionaliteit van de leerkracht/IB’er. Deze wordt geacht in staat te zijn om aan de hand van de aanwijzingen in de handleiding een gestandaardiseerde en ongestoorde toetsafname te realiseren.

2.3 Gebruiksdoel en functie

De toetsen Taalverzorging uit het Cito Volgsysteem hebben twee doelen: niveaubepaling en progressie-bepaling.

Niveaubepaling
De toetsen Taalverzorging geven de leerkracht informatie over het niveau waarop leerlingen hun geschreven taal verzorgen, individueel en als groep. Iedere behaalde vaardigheidsscore kan normgericht geïnterpreteerd worden op basis van de vaardigheidsverdeling in een adequate referentiegroep (zie paragraaf 4.2 voor de beschrijving van de referentiegroep).
De referentiegroep is op basis van de scores van de leerlingen in deze groep in vijf niveaugroepen verdeeld. Er is sprake van twee indelingen. De eerste indeling, gebaseerd op de niveaugroepen I tot en met V, gaat uit van vijf groepen van ieder 20%. Bij deze indeling worden op de registratie-overzichten de laagste en de hoogste groep nog onderverdeeld in twee groepen die ieder 10% leerlingen bevatten. Deze groepen worden van elkaar gescheiden door een stippellijn. De tweede indeling levert de niveaugroepen A tot en met E op en is gebaseerd op een indeling in kwartielen. De niveaugroepen A, B en C bestrijken elk een kwart van de populatie. Het vierde kwartiel wordt opgesplitst in twee subgroepen: D (15%) en E (10%). Zie figuur 1 voor een beschrijving van de niveaugroepen.
Eerstgenoemde indeling is symmetrisch opgebouwd en heeft als voordeel – boven de indeling gebaseerd op kwartielen – dat er een gemiddelde[1] groep onderscheiden wordt, namelijk niveaugroep III. Deze indeling blijkt in de praktijk intuïtiever aan te voelen en minder gevoelig te zijn voor verkeerde interpretaties. Om die reden wordt deze indeling in de handleiding steeds als eerste genoemd.

Figuur 1 Onderscheiden niveaugroepen

Niveau	%	Interpretatie
I	**20**	Ver boven het gemiddelde
II	**20**	Boven het gemiddelde
III	**20**	De gemiddelde groep leerlingen
IV	**20**	Onder het gemiddelde
V	**20**	Ver onder het gemiddelde

Niveau	%	Interpretatie
A	**25**	De 25% hoogst scorende leerlingen
B	**25**	De 25% leerlingen die net boven tot ruim boven het landelijk gemiddelde scoren
C	**25**	De 25% leerlingen die net onder tot ruim onder het landelijk gemiddelde scoren
D	**15**	De 15% leerlingen die ruim onder het landelijk gemiddelde scoren
E	**10**	De 10% laagst scorende leerlingen

Referentieniveaubepaling
De toetsen Taalverzorging geven de leerkracht informatie over het referentieniveau waarop leerlingen hun geschreven taal verzorgen, individueel en als groep. Het referentieniveau geeft aan of een leerling op het

[1] Het betreft hier geen gemiddelde in de statistische betekenis van het woord.

totaal van Taalverzorging een drempelniveau heeft behaald in zijn schoolloopbaan. Dat drempelniveau of referentieniveau kan voor taalverzorging in het basisonderwijs <1F, 1F of 2F zijn. Niveau 1F is het fundamentele niveau aan het eind van het basisonderwijs. Oftewel: voor een goede overgang naar het voortgezet onderwijs is niveau 1F vereist. Niveau 2F (1S) is het streefniveau voor het basisonderwijs, voor leerlingen die meer aankunnen. In het Besluit referentieniveaus Nederlandse taal en rekenen is door de overheid bepaald welke basiskennis en -vaardigheden leerlingen moeten beheersen voor taal en rekenen.

Progressiebepaling

Het volgen van leerlingen in hun groei, ook wel aangeduid als progressiebepaling, is een van de belangrijkste functies van het Cito Volgsysteem primair en speciaal onderwijs (LVS). De toetsen van het LVS geven de leerkracht (en ouders en leerlingen zelf) informatie over de ontwikkeling van de vaardigheden van de leerlingen, individueel en als groep, gedurende (vrijwel) de gehele basisschoolperiode. De toetsen geven antwoord op vragen als: is er sprake van vooruitgang, achteruitgang of van stabilisering? Is de vooruitgang – gelet op de gemiddelde vooruitgang in de populatie – volgens verwachting?

Om leerlingen te kunnen volgen wordt elke deelvaardigheid, in dit geval van ieder subdomein of deelgebied van taalverzorging, opgevat als een unidimensionale vaardigheid, of 'latente trek'. Het gehanteerde meetmodel (zie paragraaf 2.4.2) maakt het mogelijk om de scores van een leerling op verschillende toetsen, op verschillende momenten afgenomen, onderling te vergelijken. De ruwe scores op de toetsen (de ruwe score is het aantal opgaven goed) zijn daartoe te transformeren in scores op één vaardigheidsschaal. Deze unidimensionale vaardigheidsschalen die aan de afzonderlijke subdomeinen van de toetsen Taalverzorging ten grondslag liggen, zijn ontwikkeld met behulp van het *One Parameter Logistic Model* (Verhelst, 1993; Verhelst & Glas, 1995; Verhelst, Glas & Verstralen, 1995).

Geldigheid van de normen

De toetsen van het Cito Volgsysteem primair en speciaal onderwijs worden elke acht tot tien jaar vernieuwd. Niet alleen de inhoud wordt volledig vernieuwd en aangepast aan de ontwikkelingen in het onderwijs, ook worden de normen opnieuw vastgesteld. Omdat er enige tijd verloopt tussen de dataverzameling in het normeringsonderzoek en het moment waarop een vernieuwde toets wordt uitgebracht, kan men voor de toetsen Taalverzorging voor groep 6, 7 en 8 een geldigheid aanhouden tot en met 2024.

Daarnaast monitort Cito periodiek de normering: jaarlijks wordt aan de hand van representatieve afnamedata nagegaan of er systematisch verschuivingen in het prestatieniveau plaatsvinden. Indien nodig wordt de normering aangepast.

Het aantal afnamemomenten per jaar (en daaraan gekoppeld het aantal te construeren verschillende toetsen) wordt bepaald door het tempo waarin een deelvaardigheid gemiddeld gesproken binnen een leerjaar en over de gehele schoolperiode toeneemt. Meestal is er sprake van twee afnamemomenten per leerjaar ('medio' en 'einde' leerjaar, aangeduid als M en E) en twee - bij het betreffende afnamemoment passende - toetsen.

De toetsen Taalverzorging bieden voor de groep 6 en 7 één toets die genormeerd is voor de afname-momenten medio en einde. In groep 8 is alleen een toets voor het meetmoment M8 beschikbaar. Er zijn geen tussentoetsen ontwikkeld. Elk toetsonderdeel waarin een deelvaardigheid aan bod komt, wordt geconstrueerd op basis van een gekalibreerde itembank, waarbij een toets zo wordt samengesteld dat deze naar inhoud en moeilijkheidsgraad optimaal past bij het afnamemoment waarvoor deze bedoeld is.

Hoe kunnen we de LVS-toetsen Taalverzorging inzetten om leerlingen te volgen in de tijd?

Globaal kunnen we de toetsresultaten van leerlingen (of groepen leerlingen) op twee manieren interpreteren:

a. We kunnen het toetsresultaat van een leerling vergelijken met die van andere leerlingen op hetzelfde meettijdstip (afnamemoment).
b. We kunnen de toetsresultaten van dezelfde leerling vergelijken met diens eigen toetsresultaten op eerdere of latere meettijdstippen (afnamemomenten).

Bij beide vergelijkingen maken we gebruik van het feit dat de toetsresultaten door toepassing van het IRTmodel (OPLM) in de vorm van vaardigheidsscores afgebeeld kunnen worden op dezelfde vaardigheidsschaal. Dat geldt voor zowel individuele leerlingen als voor (gemiddelde) groepsresultaten.

Dit alles wordt in meer detail uitgelegd in de leerkrachthandleiding (zie het hoofdstuk 'Interpreteren en analyseren op leerling- en groepsniveau').

Ad a.
Bij de eerstgenoemde vergelijking worden de prestaties van een leerling vergeleken met de prestaties van de hele populatie op een gegeven afnamemoment. Hoe doet een leerling het, bijvoorbeeld, ten opzichte van de gemiddelde leerling? Voor dit doel is de populatie, op basis van de data die verzameld zijn in het kader van het normeringsonderzoek (zie hiervoor hoofdstuk 4 van deze wetenschappelijke verantwoording), ingedeeld in vaardigheidsniveaus (I-V, A-E). Vaardigheidsniveau I, bijvoorbeeld, bevat de 20% hoogst scorende leerlingen. Door de vaardigheidsscore van een leerling te vergelijken met deze vaardigheidsniveaus (die zijn afgebakend door percentielpunten die horen bij specifieke vaardigheidsscores), zijn uitspraken mogelijk zoals "Jelmer heeft op afnamemoment medio leerjaar 6 vaardigheidsniveau IV behaald voor de taak interpunctie". Voor de leerkracht (en voor Jelmer en zijn ouders) bevat deze uitspraak waardevolle informatie.
De leerkracht kan op basis hiervan bijvoorbeeld besluiten om Jelmer extra lesstof aan te bieden.

Ad b.
Voor het vergelijken ('volgen') van een leerling op twee verschillende tijdstippen komen twee methodes in aanmerking. Bij de eerste methode worden de vaardigheidsniveaus op de twee tijdstippen vergeleken: "op tijdstip M6 had Jelmer vaardigheidsniveau IV en op tijdstip M7 was het vaardigheidsniveau III". Bij de tweede methode worden de vaardigheidsscores op de twee verschillende momenten vergeleken: vaardigheidsscore 148, bijvoorbeeld, op tijdstip M6 en vaardigheidsscore 157 op tijdstip M7. Ook hier geldt, net als bij het vergelijken van prestaties met die van andere leerlingen, dat bij eventuele verdere acties van de leerkracht ook andere aspecten moeten worden betrokken.

Bij alle vergelijkingen die mogelijk zijn, zowel die ad a. als die ad b., moeten uitspraken over leerlingen worden gerelativeerd. In de handleiding is meer informatie te vinden over de wijze waarop de gebruiker dit kan doen. Hieronder gaan we vooral in op het belang van de (on)betrouwbaarheid van de afgenomen toetsen hierbij.
Voor elke toets geldt dat de vaardigheidsscore die bij een toetsresultaat van een leerling hoort, behept is met een meetfout. Als we rekening houden met die meetfout dan zou het best zo kunnen zijn dat Jelmer vaardigheidsniveau III heeft behaald op het eerste tijdstip en niet vaardigheidsniveau IV. Of dat we moeten concluderen dat het verschil in de prestaties op de twee tijdstippen statistisch niet significant is: Jelmer is voor- noch achteruitgegaan. In alle gevallen speelt het betrouwbaarheidsinterval (BI) rondom de vaardigheidsscore een belangrijke rol. Dat geldt ook voor de indeling in vaardigheidsniveaus. Hoe het BI doorwerkt in de indeling in vaardigheidsniveaus en de verdere gevolgen daarvan wordt beschreven in hoofdstuk 5. Op deze plaats beperken we ons tot de vraag of we de uitspraak kunnen doen dat een leerling of groep (werkelijk) 'gegroeid' is. De eenvoudigste manier is om te kijken of de BI's voor de twee tijdstippen overlappen. Als deze twee BI's niet overlappen dan is er sprake van een significant verschil in vaardigheid tussen beide tijdstippen. Overlappen ze wel dan is er geen verschil in vaardigheid. Deze eenvoudige manier van vergelijken kan de leerkracht zelf uitvoeren en wordt ook in de handleiding beschreven. We geven hier een voorbeeld. Bij de afname Taalverzorging M6 Interpunctie behaalde Keyla een vaardigheidsscore van 107 (IV) met een 67 % betrouwbaarheidsinterval van 98-112. Bij de afname M7 behaalde Keyla een vaardigheidsscore van 120 (III); het bijbehorende betrouwbaarheidsinterval daarbij is 112 -127. Aangezien de betrouwbaarheidsintervallen niet overlappen kunnen we zeggen dat Keyla's vaardigheid is toegenomen.

Conclusie
De vaardigheidsgroei voor de subdomeinen van Taalverzorging voltrekt zich langzaam in de tijd.
De verschillen tussen vaardigheidsscores op achtereenvolgende meettijdstippen zijn betrekkelijk klein. Bovendien is er sprake van meetfouten. De verschillen in vaardigheidsgroei moeten tegen de achtergrond van die meetfouten worden geïnterpreteerd. Dit betekent dat men weliswaar uitspraken kan doen over de vaardigheidsgroei van een leerling, maar dat deze uitspraken met voorzichtigheid dienen te worden gehanteerd. Dit geldt ook wanneer men de progressie van een leerling volgt in termen van vaardigheids-niveaus of een vergelijking maakt met andere leerlingen in termen van vaardigheidsniveaus. Want ook bij de indeling in vaardigheidsniveaus speelt de nauwkeurigheid van de toets een rol. Hoe de leerkracht hier in de praktijk mee om moet gaan wordt toegelicht in de handleiding voor de leerkracht.

Vaardigheidsscore

Voor de deelvaardigheden taalverzorging die in deze wetenschappelijke verantwoording aan de orde zijn, hebben we in proeftoetsingen vastgesteld dat er in de leerjaren 6 tot en met 8 sprake is van een relatief bescheiden gemiddelde vaardigheidsgroei. Met name is er sprake van een minimale groei tussen de E- en de M-afname. Dat betekent dat naar onze mening in deze leerjaren zou kunnen worden volstaan met één toetsafname per leerjaar; hetzij op het M-moment, hetzij op het E-moment. We hebben ons daarom beperkt tot de constructie van één toets per leerjaar die voor beide afnamemomenten geschikt is. Dat deze keuze correct is geweest, blijkt uit onderstaande gegevens over gemiddelde vaardigheid en vaardigheidsgroei.
We nemen als voorbeeld de deelvaardigheid Spelling niet-werkwoorden. De gemiddelde toename is steeds aanmerkelijk kleiner dan de spreiding in vaardigheid binnen de groep op enig afnamemoment. Soms is de toename niet veel meer dan een kwart van de standaarddeviatie. Bovendien lijkt de gemiddelde toename over een vol jaar gezien (dat wil zeggen M6-M7, E6-E7 et cetera) steeds kleiner te worden (achtereenvolgens 8,5 - 5,2 en 6,6).

Tabel 2.1 Gemiddelde vaardigheden en vaardigheidstoename

Onderdeel en afnamemoment	**Vaardigheidsscores**			**Jaarlijkse toename**		
Interpunctie	**Gemiddelde**	**SD**	**Toename**		**Medio**	**Eind**
M6	113,0	16,1	--	-----		
E6	120,1	19,1	7,1	-----		
M7	120,5	16,9	0,4	M6 – M7	7,5	
E7	123,2	17,7	2,7	E6 – E7		3,1
M8	128,6	19,4	5,4	M7 – M8	8,1	
Spelling niet-werkwoorden						
M6	107,4	12,7	--	-----		
E6	112,7	13,8	5,3	-----		
M7	115,9	13,4	3,2	M6 – M7	8,5	
E7	119,3	14,7	3,4	E6 – E7		6,6
M8	121,1	13,6	1,8	M7 – M8	5,2	
Grammatica						
M6	100,1	12,2	--	-----		
E6	108,3	15,2	8,2	-----		
M7	110,1	14,1	1,8	M6 – M7	10,0	
E7	115,7	16,7	5,6	E6 – E7		7,4
M8	116,3	15,7	0,6	M7 – M8	6,2	
Spelling werkwoorden						
M7	99,4	8,0	--	-----		
E7	102,9	9,1	3,5	-----		
M8	104,7	9,4	1,8	M7 – M8	5,3	

2.4 Theoretische inkadering

2.4.1 Inhoudelijk

In deze paragraaf wordt toegelicht wat het concept 'Taalverzorging' inhoudt. Ook komen de deelgebieden spelling (werkwoorden en niet-werkwoorden), interpunctie en grammatica aan bod, evenals de context waarin taalverzorging plaatsvindt. Daarnaast bespreken we de relatie tussen de toetsen Taalverzorging van het Cito Volgsysteem primair en speciaal onderwijs en de onderwijsdoelen, de leerstoflijnen, de kern- en tussendoelen primair onderwijs en het Referentiekader Taal en Rekenen (Expertgroep Doorlopende Leerlijnen Taal en rekenen, 2009a).

Schriftelijke vaardigheden zijn van groot belang om goed te functioneren en te communiceren in de samenleving. Taalverzorging is een fase van het schrijfproces waarbij de schrijver alle vaardigheden en kennis inzet om tot een verzorgde schriftelijke taalproductie te komen. Als een aspect van schriftelijke taalvaardigheid wordt Taalverzorging in de taalkunde beschouwd als een productieve vaardigheid, net als spreken. Voor de productieve taalvaardigheid zijn een aantal fases door Levelt (1989) onderscheiden:

a) het bedenken en ordenen van de inhoud (*conceptualizer*),
b) het omzetten van de inhoud in woorden en zinnen (*formulator*, gebruik makend van het *lexicon*),
c) het vinden en plaatsen van de juiste schrifttekens (de *articulator*).

In het schrijfproces speelt taalverzorging met name een rol in de laatste fase, *de articulator*, als de inhoud en de structuur van een tekst al ver gevorderd zijn en de tekst zijn voltooiing nadert. 'Spelling, interpunctie en het gebruik van hoofdletters zijn aspecten van de verzorging van het uiteindelijke product' (Hoogeveen & Kouwenberg, 2011).

Uit onderzoek is gebleken dat het produceren van een correct geschreven tekst een behoorlijke cognitieve krachtsinspanning vraagt van zowel beginnende als gevorderde schrijvers. Kellog (1996) heeft de cognitieve processen die van belang zijn bij schriftelijke vaardigheid onderzocht en stelt dat het werkgeheugen minder belast wordt door automatisering van taalverzorgingsvaardigheden. Dit komt de vertaalslag van concept naar tekst ten goede. Als een schrijver onder andere beschikt over automatismen op het gebied van spelling, interpunctie en grammaticale vaardigheid, kan hij zijn aandacht met name richten op de inhoud van de tekst. Goede taalverzorgers maken dan ook gebruik van taalconventies zoals spellingregels en/of -strategieën, grammaticale en interpunctieregels om tot een verzorgd schrijfproduct te komen.
Die laatste fase van het schrijfproces behelst ook de revisie. Voor een taalverzorger is het noodzakelijk om zijn schrijfproduct te controleren. In de beschrijving van Levelt (1989) wordt dit *self-monitoring* genoemd. Volgens Pullens (2012) beslaat schrijfproces naast het plannen, formuleren ook het reviseren van een tekst. Tekstrevisie kan beschouwd worden als een schrijfstrategie. In het onderwijs wordt veel aandacht besteed aan leerstrategieën. 'Een strategie helpt een taalgebruiker om een taaltaak gerichter en efficiënter uit te voeren" (Over drempels met taal en rekenen (2008b).

De verschillende aspecten van het domein Taalverzorging zijn verkaveld over bestaande deelgebieden van het taalaanbod in het basisonderwijs. Deze deelgebieden zijn didactisch gezien met elkaar verbonden: spelling niet-werkwoorden, spelling werkwoorden, interpunctie en grammatica. In de volgende paragrafen beschrijven we de invulling van deze deelgebieden.

Spelling
Spelling omvat de notatiesystematiek voor zowel niet-werkwoorden als werkwoorden. De Schrijver & Neijt (2002) omschrijven spelling als een systeem van regels met behulp waarvan we een bepaalde gesproken taal schriftelijk weergeven. De laatste uitgave van de spelling van het Nederlands is in 2005 vastgelegd in de Woordenlijst Nederlandse Taal (het Groene Boekje). Het gebruik van deze spelling is verplicht binnen het onderwijs en het openbaar bestuur. Bij spellen gaat het erom dat de leerling van gesproken woorden 'schriftbeelden' (dus geschreven woorden) kan maken. Bij spelling maken we onderscheid tussen klankzuivere en niet- klankzuivere woorden. De eerste fase van het onderwijs richt zich op het correct leren schrijven van de klankzuivere woorden; de leerling schrijft op wat hij hoort (in groep 3). Al snel daarna komen

de niet-klankzuivere woorden. Dat zijn de woorden waarbij er geen eenduidige relatie is tussen klank en letter. Het gaat dan om woorden zoals bomen, trein, begin. Om die goed te kunnen schrijven moet de leerling de regels kunnen toepassen of weten dat hij de letters in een woord net zo moet schrijven als in een ander woord dat hij al kent (vanaf eind groep 3).

Over het algemeen wordt de moeilijkheid bij spelling bepaald door een aantal factoren waaronder woordlengte, woordfrequentie en de van toepassing zijnde spellingcategorie. In het referentiekader taal en rekenen (Expertgroep doorlopende leerlijnen Taal en rekenen, 2009a) wordt voor de ordening van spellingproblemen een indeling in vijf klassen vermeld die is gebaseerd op een foutenanalyse door Schijf (2009).

1 alfabetisch, d.w.z. fouten in de alfabetische spelling, de koppeling tussen fonemen en grafemen, bijvoorbeeld schuirdeur of schuurduer;
2 orthografisch, d.w.z. fouten waarbij de autonome spellingregels niet goed zijn toegepast, bijvoorbeeld: programa, kooningin;
3 lexicaal-morfologisch, d.w.z. fouten in de morfologische spelling die alleen lexicaal bepaald zijn, bijvoorbeeld taloze, handoek, voordurend;
4 grammaticaal-morfologisch, d.w.z. fouten in de morfologische spelling die ook grammaticaal bepaald zijn, bijvoorbeeld vermelde/vermeldde, wordt/word, veranderd/verandert;
5 logografisch, d.w.z. fouten in de woordspecifieke schrijfwijzen, bijvoorbeeld: apoteek, milieu.

De vijf klassen, gebaseerd op een analyse van het Nederlandse taalsysteem, worden hierna uitvoerig beschreven. De indeling is zeer globaal en is niet afgebakend per leerjaar.

Alfabetische spelling is gebaseerd op een een-op-een-relatie tussen klank en teken. Het schrijven van klankzuivere woorden behoort tot de elementaire spelhandeling. In groep 3 leren leerlingen vooral klankzuivere woorden te schrijven. Woorden als ‘tak’, ‘krant’ en ‘ziek’ worden dan voor het eerst aangeboden. Aan de basis van **orthografische** spelling liggen afspraken over de schrijfwijzen van (groepen) woorden, waarbij er geen automatische klank-letteromzetting plaatsvindt. Deze woorden volgen geen regels maar moeten door leerlingen ingeprent of geleerd worden, zoals woorden met -ieuw of een verdubbeling bij de meervoudsvorm.

De **lexicaal-morfologische** spelling is spelling op basis van de opbouw van het woord, los van de grammaticale context. De speller moet inzicht hebben in de subdelen van het woord, ook wel verwante woorddelen genoemd. Doordat deze delen hetzelfde worden geschreven komt de leerling tot een correcte spelling. Het volstaat om naar het woord zelf te kijken. Woorden als ondiep (voor-achtervoegsel), boompje (verkleinwoord) en tuindeur (samenstelling) komen vanaf groep 4 aan bod.

Binnen het onderwijs komen kinderen ook in aanraking met de grammaticale context om een woord goed te kunnen spellen. Dit noemen we **morfologische spelling op syntactische basis**. Het betreft vooral werkwoordvormen als ‘ik word/hij wordt’, maar ook woorden als ‘alle(n)’ en ‘enkele(n)’. Dit vraagt om een hogere spellingvaardigheid. Leerlingen komen hier pas mee in aanraking vanaf groep 6.

Zowel in de hoge als in de lage groepen krijgen leerlingen **logografische** spelling aangeboden. Logografische spelling is gebaseerd op vaststaande combinaties zonder regelvorming, ofwel woorden met een specifieke schrijfwijze. Hier gaat het om relatief eenvoudige woorden als ‘trein’ en ‘lijst’, maar ook om leenwoorden, zoals bijvoorbeeld ‘trottoir’ en ‘team’.

Klasse 4 betreft de spelling van werkwoorden, de overige klassen betreffen de spelling van niet werkwoorden. Bij werkwoordspelling gaat het om het spellen van veranderlijke woorden. Dat wil zeggen dat onder invloed van het zinsverband de schrijfwijze kan veranderen. Dit betekent dat naast kennis van de algemene beginselen van de Nederlandse spelling, het voor werkwoordspelling noodzakelijk is om inzicht te hebben in grammaticale principes. Aarnoutse & Verhoeven (2003) geven als voorbeeld de zin *De jongens*

verwachtten gisteren een reactie. Om het werkwoord *verwachtten* correct te spellen, moeten leerlingen zowel de relatie tussen onderwerp (de jongen-s) en de persoonsvorm (verwacht-te-n) herkennen als kunnen afleiden dat het gaat om een gebeurtenis in het verleden.

Voor elke bovengenoemde klasse geldt dat deze bestaat uit een aantal spellingcategorieën die we hebben ondergebracht in een apart categorieënoverzicht voor spelling niet-werkwoorden en spelling werkwoorden. Deze zijn omschreven in bijlage 3 en ontleend aan de Leerstoflijnen begrippenlijst en taalverzorging (van der Beek & Paus, 2011).

Naast de actieve kant van spelling (zelf foutloos woorden schrijven) kent spelling ook een passieve kant, namelijk het herkennen en verbeteren van fouten in geschreven tekst (tekstrevisie). Het zelf kunnen schrijven van begrijpelijke en correct gespelde teksten is een heel belangrijke communicatieve vaardigheid. Een tekst zonder spelfouten maakt, op welk niveau dan ook, een heel andere indruk dan een tekst met spelfouten. Om die reden is ook het passief spellen van belang: als leerlingen een tekst geschreven hebben, staan daar meestal nog wel wat fouten in. Het is dan zaak dat zij geleerd hebben om die fouten op te sporen en te verbeteren. Zowel in de kern- en tussendoelen primair onderwijs en het Referentiekader Taal en Rekenen (Expertgroep Doorlopende leerlijnen Taal en rekenen, 2009a) wordt het belang van revisie onderstreept.

Interpunctie
Onder interpunctie verstaan we het plaatsen van leestekens in geschreven taal, gebruikmakend van regels en conventies. ‘Leestekens zijn het hang- en sluitwerk van de taal. Ze zijn bedoeld om de structuur van een tekst, in het bijzonder van zinnen, te verduidelijken’ (Genootschap Onze Taal, 2009). Het primaire doel van het plaatsen van leestekens is het bevorderen van de leesbaarheid en het tekstbegrip voor de lezer. De lezer krijgt meer grip op een tekst omdat leestekens helpen bij het doorgronden van de structuur. Bij gesproken taal geven intonatie en pauzes aan waar de accenten liggen en ondersteunen zo de luisteraar bij het interpreteren. In geschreven taal ontbreekt deze ondersteuning en heeft de lezer een hulpmiddel nodig in de vorm van leestekens.

Leestekens kunnen geplaatst worden tussen zinnen om zinsgrenzen aan te geven. Denk hierbij aan punten, hoofdletters en vraagtekens. Ook zijn er leestekens die in een zin aangebracht worden om het verband tussen woordgroepen of delen van een zin aan te geven, bijvoorbeeld komma’s en dubbele punten. Verder gebruiken schrijvers markeerders van citaten of bijzondere woorden, de aanhalingstekens. Om leestekens correct te kunnen plaatsen en zo de lezer te kunnen ondersteunen is het van belang dat de schrijver inzicht in de opbouw van een zin en/of tekst heeft.

Leestekens kunnen in drie groepen worden verdeeld (Renkema, 2002):
- Zinsgelederen. Dit zijn leestekens die grenzen binnen zinnen aangeven. Het gaat hierbij om de komma, de puntkomma, de dubbele punt, de liggende streepjes of haakjes.
- Zinseindetekens. Dit zijn leestekens die het einde van een zin markeren. Het gaat om de punt, het vraagteken en het uitroepteken.
- Markeerders van een citaat of van bijzondere woorden. Hierbij gaat het om het gebruik van dubbele of enkele aanhalingstekens.

Grammatica
Grammatica wordt gerekend tot het domein van de taalbeschouwing en kan heel breed gezien worden als het leren reflecteren op het taalsysteem en op de betekenis en functies van taal. Taalbeschouwingsonderwijs wordt door Van Gelderen (1988) als volgt omschreven: “In taalbeschouwingsonderwijs wordt leerlingen geleerd op een systematische wijze talige verschijnselen – formeel, semantisch of pragmatisch – te onderzoeken; ze leren hierbij met taal over taal te spreken en met behulp hiervan conclusies te trekken over het eigen of andermans taalgebruik.” Bonset (2011) onderscheidt vier vormen van taalbeschouwingsonderwijs: traditioneel grammaticaonderwijs in de zin van woordbenoemen en zinsontleden, alternatieve vormen van grammaticaonderwijs (bijvoorbeeld transformationeel-generatief grammaticaonderwijs of direct taalvaardigheidsonderwijs), taalbeschouwingsonderwijs geïntegreerd in taalvaardigheidsonderwijs en

taalkundeonderwijs. Vanwege het reflecteren op taal zijn de grammaticale begrippen in het sinds 2010 wettelijk verplicht gestelde Referentiekader taal en rekenen (Expertgroep Doorlopende Leerlijnen taal en rekenenen, 2009a) opgenomen in de begrippenlijst. Deze lijst omvat begrippen die nodig zijn om te spreken over taal en taalverschijnselen. Er wordt een indeling gemaakt in acht categorieën:

- leestekens;
- woordsoorten;
- grammaticale kennis;
- tekstkennis;
- stilistiek en semantiek;
- morfologie;
- opmaak;
- klanken.

Van de bovenstaande categorieën hebben alleen categorie twee en drie: Woordsoorten en Grammaticale kennis (zinsontleden) betrekking op traditioneel grammaticaonderwijs, de andere categorieën vallen daarbuiten.

In het basisonderwijs zien we dat grammatica naast bovengenoemde reflecterende kant met name gericht is op redekundig en taalkundig ontleden. Het grammaticaonderwijs richt zich op het verschaffen van kennis van grammaticale regels en relaties. Van belang is het kunnen toepassen van kennis van grammaticale relaties binnen woorden, tussen woorden, binnen zinnen en tussen zinnen. Het toepassen van die kennis in de vorm van het correct formuleren van woorden, zinsdelen en zinnen is een voorwaarde om goed te kunnen schrijven.
Grammaticaal inzicht en kennis van de begrippen beschouwen we niet alleen als een aparte vaardigheid binnen taalverzorging, maar ook als een onderliggende vaardigheid. Bij het aanleren van de werkwoordspelling zijn de grammaticale begrippen van wezenlijk belang. Om een werkwoordsvorm goed te kunnen spellen, moet de leerling de functie ervan in de zin kunnen vaststellen. Zo stellen Aarnoutse & Verhoeven (2003) dat het noodzakelijk is dat de leerling onderscheid kan maken tussen persoonsvorm en onderwerp om werkwoorden juist te kunnen spellen. Ook bij interpunctie is inzicht in zinsbouw en taalstructuur een vereiste, een schrijver moet zijn grammaticale kennis immers inzetten om leestekens in en tussen zinnen te plaatsen.

Taalverzorging in het basisonderwijs
Vanaf het moment dat een kind op school leert lezen en schrijven, wordt er aandacht besteed aan taalverzorging en taalverzorgingsstrategieën. In de eerste jaren van het taalverzorgingsonderwijs ligt de nadruk op spelling. Bij spelling wordt onderscheid gemaakt tussen aanvankelijk spellen (groep 3) en voortgezet spellen (vanaf groep 4). In de fase van het aanvankelijk spellen ligt de nadruk of alfabetische spelling en leert een leerling klankzuivere eenlettergrepige woorden schrijven. Vervolgens komen de orthografische en logografische spelling aan bod waarbij de leerling verschillende strategieën aangeboden krijgt om de woorden correct te leren spellen. In de fase van het voortgezet spellen, vanaf groep 4, 5 komt de nadruk te liggen op morfologische spelling en vindt een uitbreiding plaats met meer ingewikkelde woordvormen van de lexicaal-morfologische spelling en komen niet-klankzuivere meerlettergrepige woorden aan bod (Gijsel, Scheltinga, Van Druenen & Verhoeven, 2011a, 2011b; Bonset & Hoogeveen, 2009). In groep 7 en 8 is veel van de kennis van de spelling van niet-werkwoorden geautomatiseerd en is er ruimte om aandacht te besteden aan de minder gemakkelijke categorieën zoals complexe leenwoorden, woorden met moeilijke meervoudsvorming met ‘s en woorden met trema.
Vanaf groep 6 komen de leerlingen in aanraking met de basisprincipes van de werkwoordspelling, de morfologische spelling. Dit betekent tevens dat ter ondersteuning van het spellen van werkwoorden aandacht wordt besteed aan basale grammaticale principes zoals de persoonsvorm, het onderwerp en de verschillende tijden van het werkwoord. Als hulpmiddel leren de leerlingen te werken met algoritmes. In groep 7 en 8 komen ook de moeilijkere vormen van de werkwoordspelling aan bod zoals de spelling van de voltooid deelwoordvormen van de homofone werkwoorden (verhuist-verhuisd). De leerling wordt geacht veelvuldig gebruik te maken van het verworven inzicht in de grammaticale regels om werkwoorden correct te spellen.

Al vanaf het moment dat leerlingen eigen teksten gaan schrijven, leren ze interpunctie toe te passen. Aanvankelijk in groep 3, 4 alleen hoofdletters en punten, maar gaandeweg komen daar steeds meer leestekens bij. In de hogere jaargroepen wordt het gebruik van leestekens en hoofdletters steeds meer verfijnd en geautomatiseerd. Uitgangspunt voor het taalverzorgingsonderwijs is dat de kennis in alle vakken toegepast dient te worden.

Wettelijke basis voor het taalverzorgingsonderwijs
De wettelijke basis voor het onderwijs in taalverzorging is vastgelegd in het 'Referentiekader taal en rekenen' (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a). Hierin staat beschreven wat kinderen op verschillende momenten in hun schoolloopbaan op het gebied van taal en rekenen moeten kennen en kunnen. Het referentiekader onderscheidt voor taal vier domeinen: Mondelinge taalvaardigheid, Lezen, Schrijven en Begrippenlijst en taalverzorging. Er zijn voor deze domeinen vier niveaus onderscheiden.
Die niveaus zijn de fundamentele niveaus (F) genoemd. Het fundamenteel niveau 1 (niveau 1F) geldt voor het eind van het primair en speciaal onderwijs en het praktijkonderwijs, niveau 2F voor mbo 1, 2, 3 en vmbo, niveau 3F voor mbo 4 en eind havo en ten slotte niveau 4F voor eind vwo. Leerlingen die een fundamenteel niveau hebben behaald, krijgen meer aangeboden: ze gaan op weg naar het volgende niveau, het zogenoemde streefniveau. Het streefniveau 1S voor het primair en speciaal onderwijs staat gelijk aan niveau 2F. Voor de leerlingen die het fundamenteel niveau 1F op het eind van de basisschool niet halen, biedt de leerkracht adequate leerstof aan, aansluitend op de mogelijkheden van de leerlingen. Het geheel aan beschrijvingen wordt aangeduid met 'het referentiekader' en is vastgelegd in de Wet referentieniveaus Nederlandse taal en rekenen die op 1 augustus 2010 van kracht is geworden. Voor de beschrijving van taalverzorging in 'het referentiekader' kunt u bijlage 1 raadplegen.
De niveaus 1F en 2F geven een eindpunt aan. In de publicatie 'Leerstoflijnen begrippenlijst en taalverzorging beschreven' (van der Beek & Paus, 2011) is aangegeven langs welke weg de eindniveaus 1F en 1S/2F te bereiken zijn. Deze publicatie geeft een antwoord op de vraag hoe de opbouw van de leerstoflijnen voor Begrippenlijst en taalverzorging eruit kan zien. Deze leerstoflijnen kunnen worden gebruikt voor de planning en opbouw van het onderwijsaanbod. Voor de inhoud van de toetsen Taalverzorging zijn deze leerstoflijnen bepalend geweest, zowel als theoretische basis als voor de toetsmatrijs en de indeling van de categorieënoverzichten (zie paragraaf 3.2).

Voor taalverzorging is het domein 'Begrippenlijst en taalverzorging' van belang. De inhoud van het domein taalverzorging sluit aan bij drie kerndoelen Nederlandse taal voor het basisonderwijs (8, 11 en 12) (Ministerie van Onderwijs, Cultuur en Wetenschappen, 2006) en de tussendoelen en leerstoflijnen van TULE (3.7 tot en met 3.12) (TULE,2008):

Kerndoel 8
De leerlingen leren informatie en meningen te ordenen bij het schrijven van een brief, een verslag, een formulier of een werkstuk. Zij besteden daarbij aandacht aan zinsbouw, correcte spelling, een leesbaar handschrift, bladspiegel, eventueel beeldende elementen en kleur.

Toelichting en verantwoording:
Tijdens de revisie werken de leerlingen toe naar een uiteindelijke versie die publiceerbaar is. Ze herformuleren en herstructureren. Ze leren hierbij niet alleen inhoudelijk te reflecteren op hun teksten, maar letten ook op de verzorging ervan.
In dit kerndoel is spelling (zie ook kerndoel 11) een aspect van de verzorging van teksten, naast interpunctie, vormgeving (en een leesbaar handschrift).

Kerndoel 11
De leerlingen leren een aantal taalkundige principes en regels. Zij kunnen in een zin het onderwerp, het werkwoordelijke gezegde en delen van dat gezegde onderscheiden. De leerlingen kennen:
– regels voor het spellen van werkwoorden;
– regels voor het spellen van andere woorden dan werkwoorden;
– regels voor het gebruik van leestekens.

Toelichting en verantwoording:
Bij dit kerndoel gaat het erom dat leerlingen:
– de spellingsregels kennen en toepassen;
– de regels voor het gebruik van interpunctie kennen en toepassen;
– grammaticaal inzicht verwerven en zinvol gebruikmaken van dit inzicht bij het toepassen van de spellingregels en het gebruik van interpunctie.

Voor het leren van taalkundige principes en regels is het verwerven van grammaticaal inzicht noodzakelijk. Om bijvoorbeeld de werkwoorden zuiver te kunnen spellen, moeten leerlingen het onderwerp en het gezegde (en dan vooral de persoonsvorm) kunnen onderscheiden. Ook om bijvoorbeeld aanhalingstekens of een komma te kunnen plaatsen, is inzicht in de taalstructuur vereist.

Kerndoel 12
De leerlingen verwerven een adequate woordenschat en strategieën voor het begrijpen van voor hen onbekende woorden. Onder ‘woordenschat’ vallen ook begrippen die het leerlingen mogelijk maken over taal te denken en te spreken.

Toelichting en verantwoording:
De grammatica, waar dit kerndoel over gaat, wordt gerekend tot het domein van de taalbeschouwing. Maar taalbeschouwing is meer dan reflecteren op het systeem van taal. Het gaat ook om reflecteren op:
– de betekenis van taal (woord- en zinsbetekenis);
– de functies van taal (communicatief, expressief en conceptualiserend);
– de taalcultuur (taalvariatie).

Tussendoelen spelling en interpunctie bovenbouw
3.7 Kinderen zijn in staat lange, gelede woorden en woordsamenstellingen te spellen (geleidelijk, ademhaling, voetbalwedstrijd).
3.8 Ze beheersen de regels van de werkwoordspelling (hij verwachtte, de verwachte brief).
3.9 Ze zijn redelijk in staat leenwoorden correct te spellen (politie, liter, computer).
3.10 Ze kunnen complexe interpunctie duiden en toepassen: komma, puntkomma, dubbele punt, aanhalingstekens en haakjes.
3.11 Ze zijn in staat om zelfstandig hun spelling- en interpunctiefouten te onderkennen en te corrigeren.
3.12 Ze ontwikkelen een attitude voor correct schriftelijk taalgebruik.

2.4.2 Psychometrisch

2.4.2.1 Opgavenbanken
Voor het samenstellen van toetsen voor het primair en speciaal onderwijs beschikt Cito over opgavenbanken. Die liggen ten grondslag aan onder meer de toetsen in het Cito Volgsysteem primair en speciaal onderwijs (LVS-toetsen). Voor de constructie van de toetsen Taalverzorging is gebruik gemaakt van vier opgavenbanken Taalverzorging waarin de deelgebieden spelling niet-werkwoorden, spelling werkwoorden, interpunctie en grammatica zijn opgenomen. We hebben voor elke deelgebied een aparte itembank ingericht. Binnen het leerlingvolgsysteem bestonden er reeds aparte itembanken voor deze deelvaardigheden.
We hebben ervoor gekozen om deze indeling te handhaven, aangezien deze het meest recht doet aan de dimensionele structuur in de data die we in de loop der tijd over de deelgebieden verzameld hebben.

Voor andere vakgebieden in het LVS zoals Begrijpend lezen, Woordenschat, Rekenen-Wiskunde en Begrijpend luisteren zijn eveneens opgavenbanken in gebruik.

Een opgavenbank is nadrukkelijk niet eenvoudigweg een verzameling opgaven of items waaruit een toetsconstructeur min of meer naar willekeur een aantal items selecteert om een nieuwe toets te construeren. In deze paragraaf wordt beschreven wat de vereisten zijn om van een deugdelijke en psychometrisch goed gefundeerde opgavenbank te kunnen spreken.

Unidimensionaal continuüm
Het algemene uitgangspunt is dat elke deelvaardigheid van taalverzorging, dus spelling niet-werkwoorden, spelling werkwoorden, interpunctie en grammatica, kan worden opgevat als een unidimensionaal continuüm (de reële lijn), en dat elke leerling voorgesteld kan worden als een punt op die lijn, met andere woorden: als een getal. Het getal drukt de mate van vaardigheid uit op dat specifieke deelgebied, waarbij een groter getal wijst op een grotere vaardigheid. Het doel van de meetprocedure – het afnemen van een toets – is de plaats van de leerling op dit continuüm zo nauwkeurig mogelijk te bepalen. De uitkomst van de meetprocedure bestaat strikt genomen uit twee grootheden: de eerste is de schatting van de plaats van de leerling op het vaardigheidscontinuüm. De tweede grootheid geeft aan hoe nauwkeurig die schatting is, en heeft dus de status van een standaardfout, te vergelijken met de standaardmeetfout uit de klassieke testtheorie.

Latente vaardigheid
De antwoorden van een leerling op de items worden beschouwd als indicatoren van de vaardigheid, hetgeen ruwweg betekent dat men verwacht dat alle items in de vier opgavenbanken taalverzorging meten.
De vaardigheid zelf wordt als niet-observeerbaar beschouwd, en daarom gewoonlijk omschreven als een latente vaardigheid.

‘Moeilijkheid’ in de Item Respons Theorie
Hoewel items dezelfde vaardigheid meten, kunnen ze toch systematisch van elkaar verschillen.
Het belangrijkste verschil tussen de items is hun moeilijkheidsgraad. In de klassieke testtheorie wordt moeilijkheidsgraad uitgedrukt met een zogenaamde P-waarde, de proportie correcte antwoorden op het item in een welbepaalde populatie van leerlingen. In de Item Respons Theorie (IRT) die voor het construeren van de opgavenbanken wordt gebruikt, hanteert men echter een andere definitie van moeilijkheid: ruwweg gesproken is het de mate van vaardigheid die nodig is om het item goed te kunnen beantwoorden.
Dit verschil in definitie van de moeilijkheidsgraad tussen klassieke theorie en IRT is uitermate belangrijk: men kan verwachten dat de P-waarde van een item in groep 8 groter zal zijn dan in groep 6, waardoor duidelijk wordt dat de P-waarde een relatief begrip is: ze geeft de moeilijkheid aan van een item in een bepaalde populatie. Binnen de IRT is de moeilijkheid van een item gedefinieerd in termen van de onderliggende vaardigheid, zonder enige referentie aan een bepaalde populatie van leerlingen. Zo kan men ook de uitspraak begrijpen dat in de IRT vaardigheid en moeilijkheid op eenzelfde schaal liggen.

Kansmodel
De ruwe omschrijving van de moeilijkheidsgraad die in de vorige alinea werd gehanteerd (de mate van vaardigheid die nodig is om het item goed te kunnen beantwoorden) behoeft enige verdere uitwerking. Men zou deze omschrijving kunnen opvatten als een drempel: heeft een leerling die mate van vaardigheid niet, dan kan hij het item niet juist beantwoorden; heeft hij die drempel wel gehaald, dan geeft hij (gegarandeerd) het juiste antwoord. Deze interpretatie weerspiegelt een deterministische kijk op het antwoordgedrag van de leerling, die echter in de praktijk geen stand houdt, omdat eruit volgt dat een leerling die een moeilijk item correct beantwoordt geen fout kan maken op een gemakkelijker item. Daarom wordt in de IRT een kansmodel gebruikt: hoe groter de vaardigheid, des te groter de kans dat een item juist wordt beantwoord. De moeilijkheidsgraad van een item wordt dan gedefinieerd als de mate van vaardigheid die nodig is om met een kans van precies een half een juist antwoord te kunnen produceren.

Kalibratie

In het voorgaande stuk zijn nogal wat veronderstellingen ingevoerd (unidimensionaliteit; alle items zijn indicatoren voor dezelfde vaardigheid; kansmodel) die niet zonder meer voor waar kunnen worden aangenomen; er moet aangetoond worden dat al die veronderstellingen deugdelijk zijn. Dit 'aantonen' gebeurt met statistische gereedschappen waar in de volgende paragraaf dieper op wordt ingegaan. Maar voor de items in een toets gebruikt kunnen worden, moet ook geprobeerd worden de waarden van de moeilijkheidsgraden te achterhalen. Dit gebeurt met een statistische schattingsmethode die wordt toegepast op de itemantwoorden die bij een steekproef van leerlingen zijn verzameld. Het hele proces van moeilijkheidsgraden schatten en verifiëren of de modelveronderstellingen houdbaar zijn, wordt kalibratie of ijking genoemd; de steekproef van leerlingen die hiervoor wordt gebruikt heet kalibratiesteekproef.

Afnamedesigns

Meestal bevat een opgavenbank meer items dan een doorsnee toets, waardoor het praktisch niet doenlijk is om alle items aan alle leerlingen voor te leggen. Elke leerling in de kalibratiesteekproef krijgt derhalve slechts een (klein) gedeelte van de items uit de opgavenbank voorgelegd. Dit gedeeltelijk voorleggen gebeurt aan de hand van een zogeheten 'onvolledig design'. Dit moet met de nodige omzichtigheid gebeuren. Verderop wordt ingegaan op het afnamedesign dat voor de kalibratie is gebruikt, de geïnteresseerde lezer wordt verwezen naar Eggen (1993).

Belangrijke implicaties gekalibreerde opgavenverzameling

Als de kalibratie met succes uitgevoerd is, is het resultaat een zogenoemde gekalibreerde itembank. In dat proces worden de items die niet passen bij de verzameling uit de collectie verwijderd. De opgavenbank bevat voor elk item niet alleen zijn feitelijke inhoud, maar ook zijn psychometrische eigenschappen, en de statistische zekerheid dat alle items dezelfde vaardigheid aanspreken. Dit houdt onder meer het volgende in:

1 In principe kan met een willekeurige selectie items uit de bank de vaardigheid worden gemeten bij een willekeurige leerling. In principe, want een willekeurige toets die uit de itembank wordt getrokken zal in de praktijk meestal niet voldoen omdat de meetresultaten (de schatting van de vaardigheid) onvoldoende nauwkeurig zullen zijn. Voor een nauwkeuriger meting (bij een gegeven aantal items in de toets) moeten de moeilijkheidsgraden van de items in overeenstemming gebracht worden met het vaardigheidsniveau van de leerlingen.
2 Om een schatting te kunnen maken van de verdeling van de vaardigheid in een welomschreven populatie, worden selecties van items voorgelegd aan aselecte steekproeven van leerlingen uit populaties die van belang zijn voor de normering. In het geval van Taalverzorging 3.0 zijn dat steekproeven van leerlingen op de verschillende normeringsmomenten vanaf medio groep 6 tot en met medio groep 8. Daarbij maakt het, behoudens wat bij 1 is vermeld over nauwkeurigheid, niet uit welke selectie van items aan een leerling binnen een normeringsgroep wordt afgenomen. Een van de eigenschappen van gekalibreerde itembanken is immers dat met elke selectie items de vaardigheid van leerlingen kan worden bepaald. Voor een voorbeeld hiervan, zie Staphorsius (1994). In de praktijk komt dit meestal neer op het schatten van gemiddelde en standaardafwijking in de veronderstelling dat de vaardigheid normaal verdeeld is. Met deze schattingen kunnen dan ook schattingen gemaakt worden van de percentielen in de populatie.
3 Aan leerlingen die niet tot de betreffende referentiepopulatie behoren, kan dezelfde toets worden voorgelegd. De toetsscore wordt omgezet in een schatting van de vaardigheid en deze schatting kan geplaatst worden in de vaardigheidsverdeling van de populatie. Een leerling met achterstand in groep 7 kan een toets maken die normaliter aan groep 6 wordt voorgelegd, en zijn vaardigheidsschatting kan behalve met de populatie van groep 7 ook vergeleken worden met de percentielen in de populatie van groep 6, met bijvoorbeeld de uitspraak: "De vaardigheid van deze leerling komt overeen met de mediane vaardigheid in groep 6".
4 De vergelijking die in het voorgaande gemaakt is, kan evengoed plaatsvinden als de (achterstands)-leerling een andere toets (i.e. een selectie uit de opgavenbank) maakt dan de toets die normaliter aan groep 6 wordt voorgelegd. Immers, het kalibratieonderzoek heeft aangetoond dat alle items dezelfde vaardigheid meten. Een nieuwe toets meet dus dezelfde vaardigheid, zodat schattingen die van verschillende toetsen afkomstig zijn zinvol met elkaar kunnen worden vergeleken.

Tot zover de nadere bepaling van het begrip ‘opgavenbank’. In de volgende hoofdstukken van dit deel van de verantwoording worden de begrippen die hierboven aan de orde zijn geweest nader uitgewerkt en toegelicht voor de opgavenbank Taalverzorging. De verantwoording van de inhoudelijke constructie van deze opgavenbank staat in hoofdstuk 3. In hoofdstuk 4 wordt (onder andere) de psychometrische constructie van de opgavenbanken besproken (kalibratie).

2.4.1.2 Het gehanteerde meetmodel

In het normeringsonderzoek is gebruikgemaakt van een op de itemresponstheorie (IRT) gebaseerd meetmodel zoals dat bij Cito gebruikelijk is. Dergelijke modellen verschillen in een aantal opzichten nogal sterk van de klassieke testtheorie (Verhelst, 1993; Verhelst & Kleintjes, 1993; Verhelst & Glas, 1995). Bij de klassieke testtheorie staan de toets en de toetsscore centraal. Het theoretisch belangrijkste begrip in deze theorie is de zogeheten ware score, de gemiddelde score die de persoon zou behalen indien de test een oneindig aantal keren onder dezelfde condities zou worden afgenomen. Die notie geeft een van de belangrijkste (praktische) obstakels van deze theorie voor ons onderzoek weer: het is problematisch om toetsscores te vergelijken die verkregen zijn in een onvolledig design. Hoewel er methoden bestaan binnen de klassieke testtheorie om toetsscores te equivaleren (Engelen & Eggen, 1993), schiet deze benadering tekort als het gaat om de centrale vraag: hoe wordt duidelijk dat de equivalering zinvol is? Op die vraag heeft IRT een antwoord.

In de IRT staat het te meten begrip of de te meten eigenschap centraal. De IRT beschouwt het antwoord op een item als een indicator voor de mate waarin die eigenschap aanwezig is. Het verband tussen eigenschap en itemantwoord is van probabilistische aard en wordt weergegeven in de zogenoemde itemresponsfunctie. Die geeft aan hoe groot de kans is op een correct antwoord als functie van de onderliggende eigenschap of vaardigheid. Formeler: zij X_i de toevalsvariabele die het antwoord op item *i* voorstelt. X_i neemt de waarde 1 aan in geval van een correct antwoord en 0 in geval van een fout antwoord. Als symbool voor de vaardigheid wordt θ (theta) gekozen. De vaardigheid θ is niet rechtstreeks observeerbaar. Dat zijn alleen de antwoorden op de opgaven. Dat is de reden waarom θ een 'latente' variabele wordt genoemd[2]. De itemresponsfunctie $f_i(\theta)$ is gedefinieerd als een conditionele kans:

$$f_i(\theta) = P(X_i = 1 \mid \theta) \qquad (2.1)$$

Een IRT-model is een speciale toepassing van (2.1) waarbij aan de functie $f_i(\theta)$ een meer of minder specifieke functionele vorm wordt toegekend. Een eenvoudig en zeer populair voorbeeld is het Raschmodel (Rasch, 1960) waarin $f_i(\theta)$ gegeven is door

$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \qquad (2.2)$$

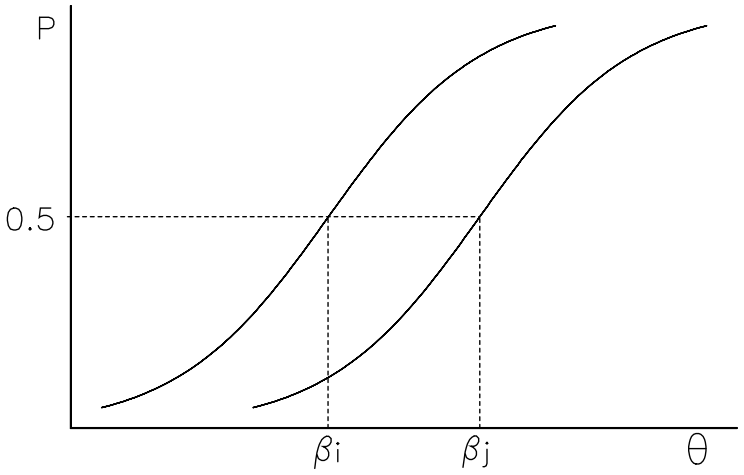
waarin β_i de moeilijkheidsparameter van item *i* is. Dat is een onbekende grootheid die geschat wordt uit de observaties. De grafiek van (2.2) is weergegeven in figuur 2.1 voor twee items, *i* en *j*, die in moeilijkheid verschillen. Deze figuur illustreert dat de itemresponsfunctie een stijgende functie is van θ: hoe groter de vaardigheid, des te groter de kans op een juist antwoord. Indien de latente vaardigheid precies gelijk is aan de moeilijkheidsparameter β_i, volgt

$$f_i(\beta_i) = \frac{\exp(\beta_i - \beta_i)}{1 + \exp(\beta_i - \beta_i)} = \frac{1}{1 + 1} = \frac{1}{2} \qquad (2.3)$$

[2] Dit maakt duidelijk waarom men de modellen die ressorteren onder de IRT, ook wel aanduidt met 'latente trek'-modellen.

Daaruit volgt onmiddellijk een interpretatie voor de parameter β_i: het is de 'hoeveelheid' vaardigheid die nodig is voor de kans van precies een half om het item *i* juist te beantwoorden. Uit de figuur blijkt duidelijk dat voor item *j* een grotere vaardigheid nodig is om diezelfde kans te bereiken, maar dit is hetzelfde als te zeggen dat item *j* moeilijker is dan item *i*. De parameter β_i kan dus terecht omschreven worden als de moeilijkheids-parameter van item *i*. De implicatie van het bovenstaande is dat 'moeilijkheid' en 'vaardigheid' op dezelfde schaal liggen.

Figuur 2.1 Twee itemresponscurven in het Raschmodel



Formule (2.2) is geen beschrijving van de werkelijkheid, het is een hypothese over de werkelijkheid die getoetst kan worden op haar houdbaarheid. Hoe zo'n toetsing grofweg verloopt, is te verduidelijken aan de hand van figuur 2.1. Daaruit blijkt dat, voor welk vaardigheidsniveau dan ook, de kans om item *j* juist te beantwoorden steeds kleiner is dan de kans op een juist antwoord op item *i*. Hieruit volgt de statistisch te toetsen voorspelling dat de verwachte proportie juiste antwoorden op item *j* kleiner is dan op item *i* in een willekeurige steekproef van personen. Splitst men nu een grote steekproef in twee deelsteekproeven, een 'laaggroep', met de vijftig procent laagste scores, en een 'hooggroep', met de vijftig procent hoogste scores, dan kan men nagaan of de geobserveerde P-waarden van de opgaven in beide deelsteekproeven op dezelfde wijze geordend zijn. Daarvan kan strikt genomen alleen sprake zijn als, in termen van de klassieke testtheorie uitgedrukt, alle opgaven eenzelfde discriminatie-index hebben. Dat echter blijkt lang niet altijd zo te zijn, ook in ons geval niet. Veel van de items blijken dan ook niet beschreven te kunnen worden met het Raschmodel. Daarom is bij dit instrument gekozen voor een ander IRT-model.

Alvorens het hier gebruikte model te introduceren, is eerst een kanttekening nodig bij het schatten van de moeilijkheidsparameters in het Raschmodel. Een vaak toegepaste schattingsmethode is de 'conditionele grootste aannemelijkheidsmethode' (in het Engels: Conditional Maximum Likelihood, verder aangeduid als CML). Die maakt gebruik van het feit dat in het Raschmodel een afdoende steekproefgrootheid ('sufficient statistic') bestaat voor de latente variabele θ, namelijk de ruwe score of het aantal correct beantwoorde items. Dat betekent grofweg dat, indien de itemparameters bekend zijn, alle informatie die het antwoordpatroon over de vaardigheid bevat, kan worden samengevat in de ruwe score; het doet er dan verder niet meer toe welke opgaven goed en welke fout zijn gemaakt. Hieruit vloeit voort dat de conditionele kans op een juist antwoord op item *i*, gegeven de ruwe score, een functie is die alleen afhankelijk is van de itemparameters en onafhankelijk van de waarde van θ[3]. De CML-schattingsmethode maakt gebruik van deze functie.

[3] Een gedetailleerde uiteenzetting hierover kan men vinden in Verhelst, 1992.

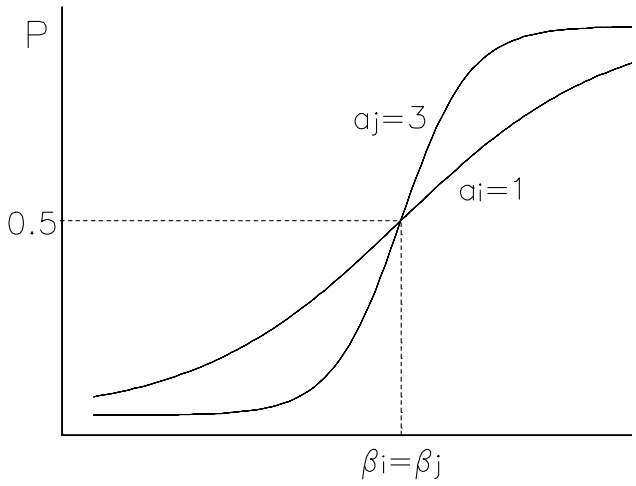
Deze methode maakt geen enkele veronderstelling over de verdeling van de vaardigheid in de populatie, en is ook onafhankelijk van de wijze waarop de steekproef is getrokken.

De CML-schattingsmethode is echter niet bij elk meetmodel toepasbaar. In het zogeheten éénparameter logistisch model (One Parameter Logistic Model, afgekort: OPLM) is CML mogelijk. Dit model is, anders dan het Raschmodel, wel bestand tegen 'omwisseling' van 'proporties juist' in verschillende steekproeven (Glas & Verhelst, 1993; Eggen, 1993; Verhelst & Kleintjes, 1993). De itemresponsfunctie van het OPLM is gegeven door

$$f_i(\theta) = \frac{\exp[\, a_i(\theta - \beta_i)\,]}{1 + \exp[\, a_i(\theta - \beta_i)]}, \qquad (2.4)$$

waarin a_i de zogenoemde discriminatie-index van het item is. Door deze indices te beperken tot (positieve) gehele getallen, en door ze a priori als constanten in te voeren, is het mogelijk CML-schattingen van de itemparameters β_i te maken. In figuur 2.2 is de itemresponscurve weergegeven van twee items *i* en *j*, die even moeilijk zijn maar verschillend discrimineren.

Figuur 2.2 Twee itemresponscurven in het OPLM: zelfde moeilijkheid, verschillende discriminatie



De schattingen worden berekend met het computerprogramma OPLM (Verhelst, Glas & Verstralen, 1995). Dit programma voert ook statistische toetsen uit op grond waarvan kan worden bepaald of het model de gegevens adequaat beschrijft. Omdat een aantal van deze toetsen bijzonder gevoelig is voor een verkeerde specificatie van de discriminatie-indices, zijn de uitkomsten van deze toetsen bruikbaar als modificatie-indices: ze geven een aanwijzing in welke richting deze discriminatie-indices moeten worden aangepast om een betere overeenkomst tussen model en gegevens te verkrijgen. Kalibratie van items volgens het OPLM is dan ook een iteratief proces waarin alternerend de modelfit van items wordt onderzocht door middel van statistische toetsing en de waarden van de discriminatie-indices worden aangepast op grond van de resultaten van deze toetsen. Deze aanpassingen geschieden in de praktijk op basis van een en hetzelfde gegevensbestand. Er kan dus kanskapitalisatie optreden. Indien een steekproef een voldoende grootte heeft, is het effect van deze kanskapitalisatie echter gering (Verhelst, Verstralen & Eggen, 1991).

Voor de schatting van de populatieverdeling wordt gebruikgemaakt van de 'marginale grootste aannemelijkheidsmethode' (in het Engels: Marginal Maximum Likelihood, verder afgekort als MML). Deze schattingsmethode veronderstelt naast (2.2) ook nog dat de vaardigheid θ in de populatie een bepaalde

verdeling heeft. De meeste computerprogramma's die IRT-analyses kunnen uitvoeren, veronderstellen een normale verdeling. Bovendien stelt deze methode de voorwaarde dat de steekproef die voor de schatting gebruikt wordt uit die verdeling een aselecte steekproef is. Omdat leerlingen bovendien gevolgd worden, is het mogelijk gelijktijdig de verdelingen op de verschillende normeringsmomenten te schatten.

3 Beschrijving van de toetsen

3.1 Opbouw en structuur van de toetsen

Het toetspakket Taalverzorging voor groep 6, 7 en 8 uit het Cito Volsysteem primair en speciaal onderwijs bestaat uit de volgsysteemtoetsen: M6/E6, M7/E7 en M8. Per jaargroep is er één toets beschikbaar die genormeerd is voor de twee reguliere afnamemomenten in het jaar, het zogeheten M-moment (halverwege het schooljaar: half januari/half februari) en het E-moment (aan het einde van het schooljaar: juni). De toets M7/E7 bijvoorbeeld, kent twee afnamemomenten, maar bevat dezelfde inhoud. De school beslist op welk moment de toets wordt afgenomen: het M-moment óf het E-moment. Alleen in groep 8 is er maar een afnamemoment: M8.

Opbouw
De toetsen Taalverzorging voor groep 6, 7 en 8 bestaan per jaargroep telkens uit drie (groep 6) of vier taken (groep 7 en 8) van elk 20 opgaven. Deze taken dienen bij voorkeur te worden afgenomen op verschillende dagdelen, zodat de leerlingen geconcentreerd aan de taken kunnen werken. De afnameduur is circa 30 minuten per taak.

De toets voor groep 6 is als volgt ingedeeld:

taak 1	spelling niet-werkwoorden	20 opgaven
taak 2	interpunctie	20 opgaven
taak 3	grammatica	20 opgaven

De toetsen voor groep 7 en 8 zijn als volgt ingedeeld:

taak 1	spelling niet-werkwoorden	20 opgaven
taak 2	interpunctie	20 opgaven
taak 3	grammatica	20 opgaven
taak 4	spelling werkwoorden	20 opgaven

Vorm
De opgaven in de toetsen Taalverzorging zijn meerkeuzeopgaven met vier antwoordalternatieven.
Met behulp van verschillende typen opgaven worden de vaardigheden gemeten die de leerlingen nodig hebben om tot een goed verzorgde schriftelijke taalproductie te komen.
De opgaven zijn opgenomen in een opgavenboekjes met respectievelijk 3 of 4 taken die zowel een klassikale als individuele afname mogelijk maken. Door het gebruik van meerkeuzeopgaven wordt het nakijken en het bepalen van de toetsscore zo objectief en efficiënt mogelijk gehouden. De leerlingen beantwoorden de vragen door het antwoord op het antwoordblad te noteren. Hierdoor zijn de boekjes meerdere jaren te gebruiken.

Keuze van een passende toets: toetsen op maat
De vaardigheid in de deelgebieden van taalverzorging van leerlingen in een groep loopt vaak sterk uiteen. Als gevolg daarvan zal eenzelfde toets taalverzorging voor een deel van de leerlingen op niveau zijn, maar voor sommige leerlingen erg moeilijk of erg gemakkelijk. Met name voor een aantal leerlingen van niveau IV en voor de leerlingen van niveau V (of de leerlingen van niveau D en E) zijn de toetsen van het eigenlijke afnamemoment (bijvoorbeeld de M7/E7-toets voor leerlingen medio of eind groep 7) aan de moeilijke kant. Voor een aantal leerlingen van niveau I (of niveau A) zijn de toetsen echter aan de gemakkelijke kant. Voor leerlingen die zich minder snel of juist sneller ontwikkelen dan de gemiddelde leerling, is het belangrijk om het niveau van de toets af te stemmen op het niveau van de leerling in plaats van op het aantal jaren onderwijs

dat de leerling gevolgd heeft. Dit noemen we toetsen op maat. Zo wordt op de meest betrouwbare manier de vaardigheid van de leerling gemeten. En uiteraard is het maken van een toets op maat prettiger voor de leerlingen. Voor het toetsen op maat wordt gebruikgemaakt van de onderliggende vaardigheidsschaal. Deze schaal maakt het mogelijk om de resultaten van leerlingen die verschillende toetsen voor een bepaald leergebied maken toch met elkaar te vergelijken. Ook kan zo de ontwikkeling van individuele leerlingen in de tijd worden gevolgd. De onderliggende meettechniek voorziet er namelijk in dat iedere ruwe score – op welke toets van een deelgebied van Taalverzorging deze score ook behaald is – kan worden omgezet in een score op één en dezelfde vaardigheidsschaal. Leerlingen kunnen daardoor bijvoorbeeld een toets maken die hoort bij een vorig afnamemoment (een E7-leerling maakt een toets E6) of een volgend afnamemoment (een E7-leerling maakt de toets M8). Het is wel van belang om de inhoud van de toets die past bij het functioneringsniveau van de leerling met het onderwijsaanbod te vergelijken. Een leerling die lesstof uit groep 6 krijgt aangeboden, kan een functioneringsniveau >M7 scoren. Dit betekent *niet* dat de leerling de stof die niet is aangeboden (bijvoorbeeld bepaalde categorieën die specifiek voor groep 7 gelden) al volledig beheerst.

Afname
De toets wordt in principe klassikaal afgenomen door de leerkracht of IB' er. De afname start met een klassikale instructie. De toetsmap bevat hiervoor een afnamekaart met afname-instructies. De afname is niet aan tijd gebonden. De leerlingen werken de desbetreffende taak in het opgavenboekje door en geven de antwoorden aan op het antwoordblad. Aan het eind van de toets controleren de leerlingen de antwoorden waarna de leerkracht/IB'er de antwoordbladen ophaalt.
De leerkracht/IB' er kan ervoor kiezen om de toets individueel af te nemen bij leerlingen met concentratie-problemen, leerlingen die langzamer dan gemiddeld werken of bij leerlingen die afwezig waren bij de klassikale afname. Belangrijk is dat de leerkracht of IB' er zich ook bij een individuele afname aan de afname-instructies houdt.
In de toetsmap is een handleiding opgenomen die zich richt op de organisatorische kant van de afname en op de verwerking en interpretatie van de toetsresultaten. In de handleiding is extra aandacht besteed aan het afnemen van de toetsen conform de afname-instructies. Er is geëxpliciteerd welke aanpassingen de leerkracht eventueel zelf kan doen en welke invloed dat heeft op de vergelijkbaarheid van de scores.

Scoring
Voor het handmatig nakijken van de toetsen kan gebruikgemaakt worden van een nakijkkaart en/of een nakijkmal met goede antwoorden, die in de toetsmap is opgenomen. Indien gewenst kan de leerkracht/IB' er in het Computerprogramma LOVS de opgaven die fout beantwoord zijn aanklikken. Op basis van het aantal goede antwoorden, de toetsscore, wordt een inschatting gemaakt van de vaardigheid op de deelgebieden van Taalverzorging van de leerlingen. De leerkracht/IB'er kan ook het aantal goede antwoorden invoeren in het Computerprogramma LOVS. De toetsscore wordt zo automatisch omgezet naar de bijbehorende vaardigheidsscore met een score-interval ofwel betrouwbaarheidsinterval. Een andere optie is om met behulp van de omzettingstabellen in de toetsmap of op Cito Portal de vaardigheidsscore bij de behaalde toetsscore op te zoeken. Men kan ook gebruikmaken van een leerlingadministratiepakket van een andere partij dan Cito.

Verwerking resultaten en interpretatie
Na de toetsafname en de scoring van de leerlingantwoorden kunnen de toetsresultaten door de leerkracht of IB'er verwerkt worden op speciaal ontwikkelde rapportageformulieren per deelgebied Taalverzorging, zoals leerlingrapporten en groepsoverzichten. Deze rapportages zijn beschikbaar via Cito Portal.

Niet alleen de scoring van de toetsen kan met behulp van de computer worden uitgevoerd, ook de categorieënanalyse kan via de computer gedaan worden. Dit kan door het aantal goede antwoorden in te voeren in het Computerprogramma LOVS. Op basis van de prestaties van de leerling berekent het computerprogramma een verwacht aantal goed per deelgebied. De categorieënanalyse geeft inzicht in de deelgebieden waarop een individuele leerling of een groep leerlingen zwakker of juist sterker scoort dan verwacht.

Op schoolniveau kan een IB'er en/of directeur met de computer een dwarsdoorsnede en trendanalyses opvragen. Met behulp van deze overzichten kan de kwaliteit van het gegeven onderwijs op groeps- en schoolniveau geanalyseerd worden.

In de handleiding in de toetsmap worden in hoofdstuk 4 de interpretatie- en analysemogelijkheden op leerling- en groepsniveau behandeld. In hoofdstuk 5 van de handleiding komt de interpretatie op schoolniveau aan bod. De handleiding gaat in op de inhoudelijke interpretatie van de rapportages. In de handleiding bij het Computerprogramma LOVS staan de aanwijzingen over de wijze waarop de rapportages zijn op te vragen en welke keuzemogelijkheden de school hierbij heeft.
In de toetsmaterialen zijn twee niveau-indelingen opgenomen, waarmee de leerkracht de scores van een leerling kan vergelijken met die van een grote groep leerlingen (zie ook paragraaf 2.3). De leerkracht kan een keuze maken uit een indeling in de niveaus:
– I tot en met V;
– A tot en met E.

Daarnaast heeft de leerkracht de mogelijkheid om functioneringsniveaus op te vragen. De functioneringsniveaus geven aan met welke gemiddelde leerling de vaardigheidsscore van de getoetste leerling vergelijkbaar is. Een functioneringsniveau E6 betekent bijvoorbeeld dat de vaardigheidsscore van de leerling overeenkomt met de score van de gemiddelde leerling eind groep 6. De indeling in functioneringsniveaus is oorspronkelijk ontwikkeld voor het speciaal (basis)onderwijs, om meer inzicht te krijgen in het niveau van de leerlingen met forse leerachterstanden. Mede dankzij de komst van het 'passend onderwijs' ontstond ook bij regulier onderwijs de wens om functioneringsniveaus te gebruiken en zijn de functioneringsniveaus opgenomen in de rapportages.

Analyse van resultaten: Categorieënanalyse
Voor analyses op leerlingniveau is een speciale rapportage ontwikkeld: de categorieënanalyse. Per jaargroep biedt het toetspakket de mogelijkheid om leerlingen te volgen op de verschillende deelgebieden van Taalverzorging. Op basis van de vaardigheidsscore van de leerling en het toegekende niveau (I t/m V, A t/m E of functioneringsniveau) berekent het Computerprogramma LOVS een verwachting van het aantal goed beantwoorde opgaven. De verwachte aantallen worden vergeleken met het daadwerkelijke aantal goede antwoorden.
De categorieënanalyse is bedoeld als hulpmiddel voor de leerkracht om na te gaan of de leerling, gegeven zijn vaardigheidsniveau, evenwichtig presteert op elk van de verschillende of deelgebieden van de toetsen Taalverzorging. In de grafische presentatie bij de categorieënanalyse kan een leerkracht zien of een leerling bij een bepaald deelgebied beter of zwakker scoort dan verwacht. Wanneer een categorieënanalyse[4] aanwijzingen geeft dat een leerling bij één of meerdere deelgebieden zwakker scoort dan verwacht, dan is voor de leerkracht het bijbehorende advies om aan de hand van bijvoorbeeld eigen observaties en de resultaten op methodetoetsen dit beeld te verifiëren en om de antwoorden bij de opgaven uit het betreffende deelgebied nader te bekijken. Indien nodig kan de leerkracht een diagnostisch gesprek voeren om meer informatie te verkrijgen over welke fouten de leerling binnen dit deelgebied maakt.

Naast een 'categorieënanalyse leerling' is er ook een zogenoemde 'categorieënanalyse groep'. In deze laatste rapportage staan eerst per leerling alle resultaten uit de 'categorieënanalyse leerling' overzichtelijk onder elkaar voor de hele groep. Vervolgens staat in een tabel aangegeven hoeveel leerlingen uit die groep per categorie beneden en hoeveel leerlingen boven verwachting scoren, inclusief de gemiddelde afwijking naar beneden/boven.

[4] De term categorieënanalyse kan hier verwarrend werken omdat er binnen de deelgebieden ook sprake is van categorieën (denk bijvoorbeeld aan spelling). Aangezien we bij de toetsen Rekenen en wiskunde ook spreken van categorieënanalyse is de term ook bij de toetsen Taalverzorging gehanteerd. We bedoelen echter deelgebieden.

In de handleiding bij het Computerprogramma LOVS is voor de leerkrachten een uitvoerige beschrijving opgenomen van de categorieënanalyse en de interpretatie van de uitkomsten. Ook in hoofdstuk 4 van de handleiding in de toetsmap staan aanwijzingen over de interpretatie en het gebruik van de 'categorieën-analyse leerling' en de 'categorieënanalyse groep'.
Naar de categorieënanalyse is geen empirisch onderzoek verricht en deze moet dan ook puur gezien worden als een handreiking richting de leerkracht. De statistiek geeft aan hoe groot de verschillen zijn tussen verwacht en geobserveerd en of op basis van kansrekening aan de verschillen belang kan worden gehecht. Of er daadwerkelijk conclusies voor het onderwijs uit afgeleid kunnen worden, hangt af van nadere analyse en interpretatie van de antwoorden van de leerling.
De toetsen en rapportagemogelijkheden maken deel uit van een systeem van leerlingzorg waarbij een school werkt volgens de cyclus signaleren, analyseren, plannen en handelen. De rapportages richten zich hierbij op de eerste twee fases van deze cyclus.

3.2 Inhoudsverantwoording

In het ontwikkelproces van de toetsen zijn een aantal fasen te onderscheiden:
- domeinbeschrijving en uitwerking in deelgebieden en categorieën;
- itemconstructie;
- proeftoetsing en kalibratie-analyses;
- samenstelling van de te normeren toetsen;
- normeringsonderzoek;
- samenstelling van de definitieve toetsen.

We zullen deze fasen hieronder nader toelichten.

De onderstaande informatie vormt een aanvulling op de inhoudsverantwoording die is opgenomen in de handleiding van het toetspakket Taalverzorging voor groep 6 tot en met 8. In hoofdstuk 6 daarvan staat een uitgebreide inhoudsbeschrijving per deelgebied en een serie overzichten die leerkrachten zicht geven op de doorgaande lijn bij de onderscheiden onderwerpen. Met behulp van die overzichten kunnen de leerkrachten de scores van leerlingen inhoudelijk interpreteren. De paragrafen bestaan uit grafieken waarop de p50- en p80-kanspunten van de items in de toetsen, geordend op basis van P-waarde, zijn afgebeeld, alsmede de vaardigheidsverdelingen op de verschillende afnamemomenten. Bij de grafieken horen overzichten waarbij de opgaven eveneens zijn geordend op basis van P-waarden. Met een willekeurige vaardigheidsscore als uitgangspunt kan de leerkracht uit de overzichten afleiden welke opgaven van dat onderdeel bij die vaardigheidsscore goed beheerst worden, welke matig en welke onvoldoende.

3.2.1 Domeinbeschrijving en uitwerking deelgebieden en categorieën

Uitwerking domeinbeschrijving
Op basis van de domeinbeschrijving (zie paragraaf 2.4.1) zijn de deelgebieden en categorieën geselecteerd die relevant zijn voor de drie jaargroepen. Deze onderwerpen komen aan bod in de meest gebruikte taal- en spellingmethodes. Bij het construeren van de opgaven en het samenstellen van de toets is gekeken naar de wijze waarop en de mate waarin de deelgebieden en categorieën in de methodes naar voren komen. Voorafgaand aan de constructie van de opgaven hebben we vastgesteld dat de leerlijnen van de methodes op hoofdlijnen aan elkaar gelijk zijn en dat groepen leerlingen die met een andere spelling- of taalmethode werken de betreffende stof aangeboden hebben gekregen. De toetsen kunnen derhalve bij elke methode voor de bovenbouw gebruikt worden.

De toetsmatrijs voor de toetsen Taalverzorging voor groep 6 tot en met 8 is samengesteld op basis van het referentiekader Nederlandse taal (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a), de kerndoelen Nederlandse taal (Ministerie van OCW, 2006), de tussendoelen gevorderde geletterdheid voor de middenbouw (Aarnoutse & Verhoeven, 2003), de Leerstoflijnen begrippenlijst en taalverzorging (van der

Beek & Paus, 2011), de meest gebruikte taal- en spellingmethodes en recente publicaties over taalverzorging. Voor een uitvoerige beschrijving van de inhoud van de toetsen en de methodeanalyse verwijzen we naar de Inhoudsverantwoording in het toetspakket (Cito, 2015). In onderstaand overzicht staat per deelgebied aangegeven welke leerstof in de toetsen aan bod komt.

De toetsen Taalverzorging voor groep 6 tot en met 8 bevatten opgaven uit drie, respectievelijk vier subdomeinen of deelgebieden: spelling niet-werkwoorden, spelling werkwoorden (vanaf groep 7), interpunctie en grammatica. De keuze voor het niet aanbieden van taak spelling werkwoorden in groep 6 vindt haar oorspong in het onderwijsaanbod. Werkwoordspelling wordt voor het eerst aangeboden in groep 6, maar nog slechts in beperkte mate. Pas in groep 7 is het onderwijsaanbod toereikend geweest om de vaardigheid spelling werkwoorden te kunnen toetsen.

Verdeling van de opgaven per deelgebied over de toetsen
De verdeling over de deelgebieden van de aantallen opgaven die zijn gepubliceerd, komt overeen met de gewenste toetsmatrijs die we opgesteld hebben voor de toetsen voor groep 6, 7 en 8. Ieder deelgebied heeft per leerjaar een gelijke lengte: 20 opgaven. Het is het minimale aantal opgaven waarmee een betrouwbare vaardigheidsscore gegeven kan worden. We hebben gekozen voor het aantal van 20 opgaven uit het oogpunt van gebruiksvriendelijkheid voor de leerkracht en de leerling. Uitgangspunt was dat aan een leerling in groep 7 en 8 in totaal 80 opgaven Taalverzorging voorgelegd kunnen worden.

Tabel 3.1 Gerealiseerde aantallen uit toetsmatrijs

Taalverzorging	**Gerealiseerd**		
	groep 6	**groep 7**	**groep 8**
Spelling niet-werkwoorden	20	20	20
Interpunctie	20	20	20
Grammatica	20	20	20
Spelling werkwoorden		20	20

Spelling
De taalkundige indeling in vijf klassen van spellingproblemen heeft de basis gevormd voor een overzicht van spellingcategorieën (zie bijlage 3). Hiermee sluiten de taken Spelling van de toetsen Taalverzorging aan bij de indeling die is gehanteerd in het Referentiekader Taal bij het domein ‘Begrippenlijst en Taalverzorging’ (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a; Van der Beek & Paus, 2011). Op basis van een methode-analyse is de indeling van spellingcategorieën vervolgens verder verfijnd tot een overzicht van spellingcategorieën dat gebruikt is voor de taken Spelling van de toetsen Taalverzorging voor groep 6 tot en met 8 en de toetsen Spelling 3.0 voor groep 3 tot en met 8.
Alle doelwoorden in de spellingtaken horen bij een bepaalde spellingcategorie. De spellingcategorie geeft aan welke spellingmoeilijkheid er in het woord zit. Van elke categorie is nagegaan of de bijbehorende spellingkwestie aan de orde komt in de huidige spellingmethoden en zo ja, op welk moment voor het eerst. In het toetspakket is het categorieënoverzicht voor groep 6 tot en met 8 opgenomen, als bijlage bij de handleiding.

In het referentiekader taal en rekenen (Expertgroep Doorlopende Leerlijnen, 2009a) wordt voor de ordening van spellingproblemen een indeling in vijf klassen vermeld die is gebaseerd op een foutenanalyse door Schijf (2009).
1 alfabetisch, d.w.z. fouten in de alfabetische spelling, bijvoorbeeld spelling niet-werkwoorden cat. 12 woorden met -eer, -oor, -eur: beer, poort, deur;
2 orthografisch, bijvoorbeeld spelling niet-werkwoorden cat. 19 woorden met -eeuw, -ieuw: sneeuw, nieuw;
3 lexicaal-morfologisch, bijvoorbeeld spelling niet-werkwoorden cat. 29 woorden met -tie: vakantie, garantie;

4 grammaticaal-morfologisch, bijvoorbeeld spelling werkwoorden cat. 3a voltooid deelwoord: geleefd, gewoond;
5 logografisch, bijvoorbeeld spelling niet-werkwoorden cat. 25 woorden met -ie als i: gitaar, minuut.

De opgaven niet-werkwoorden vallen in de klasse alfabetisch, orthografisch en lexicaal-morfologisch.
De opgaven werkwoorden behorend volledig tot de klasse grammaticaal morfologisch.

Tabel 3.2 Verdeling spellingcategorieën spelling niet-werkwoorden en werkwoorden groep 6, 7 en 8

spellingklasse	**Verdeling spellingcategorieën**					
	gewenst M6/E6	**gerealiseerd M6/E6**	**gewenst M7/E7**	**gerealiseerd M7/E7**	**gewenst M8**	**gerealiseerd M8**
alfabetisch	3	1	2	2	0	0
orthografisch	12	13	8	8	8	5
lexicaal-morfologisch	10	10	15	14	15	14
grammaticaal morfologisch	0	0	40	40	40	40
logografisch	15	16	15	16	17	21

Het is duidelijk dat de meest voorkomende spellingcategorieën voor de bovenbouw binnen de klassen lexicaal-morfologisch en logografisch vallen. De klassen alfabetisch en orthografisch horen meer thuis in het onderwijs voor de onderbouw. In de proeftoetsen hebben we wel getracht deze opgaven ook op te nemen, maar op psychometrische gronden hebben we deze opgaven grotendeels moeten laten vervallen. Zodoende hebben we de verdeling die we beoogd hadden, niet helemaal kunnen realiseren.

Interpunctie
De indeling in drie groepen die Renkema (2002) hanteert, hebben wij omgevormd naar twee hoofd-categorieën van leestekens, aangezien de laatste groep slechts bestond uit aanhalingstekens. Wij zijn tot de volgende indeling gekomen:

- Zinsgeleders en markeerders van een citaat. Tot deze groep behoren leestekens die grenzen binnen zinnen aangeven. Het gaat hierbij om de komma, de puntkomma, de dubbele punt, de liggende streepjes of haakjes. Daarnaast zijn aan deze groep de aanhalingstekens toegevoegd.
- Zinseindetekens. Dit zijn leestekens die het einde van een zin markeren. Het gaat om de punt, het vraagteken en het uitroepteken.

Deze indeling in twee groepen van leestekens heeft de basis gevormd voor onze toetsmatrijs Interpunctie. Hiermee sluiten de taken Interpunctie van de toetsen Taalverzorging aan bij de indeling die is gehanteerd in het Referentiekader Taal bij het domein ‘Begrippenlijst en Taalverzorging’ (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a; Van der Beek & Paus, 2011). Op basis van een methode-analyse is de indeling vervolgens verder verfijnd. Van elk leesteken is nagegaan of dit aan de orde komt in de huidige taalmethoden en zo ja, op welk moment voor het eerst. In het toetspakket is het overzicht van de gehanteerde leestekens voor groep 6 tot en met 8 opgenomen, als bijlage bij de handleiding. U treft het overzicht ook aan in bijlage 3.

Tabel 3.4 Verdeling Interpunctiecategorieën groep 6, 7 en 8

Interpunctie

Hoofdcategorie	**Verdeling interpunctiecategorieën**					
	M6/E6 gewenst	**M6/E6 gerealiseerd**	**M7/E7 gewenst**	**M7/E7 gerealiseerd**	**M8 gewenst**	**M8 gerealiseerd**
Zinseindetekens	5	5	10	10	10	9
Zinsgeleders en citaatmarkeerders	15	15	10	10	10	11

Bij Interpunctie zien we dat de gewenste verdeling over de hoofdcategorieën grotendeels gerealiseerd is. In groep 6 worden voornamelijk zinseindetekens aangeboden in het onderwijs. Vanaf groep 7 nemen de 'zinsgeleders en zinsmarkeerders' een grotere plaats in, in het onderwijsaanbod.

Grammatica

Voor grammatica hanteren wij een indeling met twee hoofdcategorieën van grammaticale begrippen die zowel redekundig als taalkundig ontleden omvat:

- woordbenoeming met de categorieën zelfstandig en bijvoeglijk naamwoord, lidwoord, werkwoord;
- zinsvormen en zinsontleding met de categorieën onderwerp, persoonsvorm, lijdend voorwerp, gezegde en hoofdzin/bijzin.

Deze indeling in twee groepen van grammaticale begrippen heeft de basis gevormd voor onze toetsmatrijs Grammatica. Hiermee sluiten de taken Grammatica van de toetsen Taalverzorging gedeeltelijk aan bij de indeling van de Begrippenlijst die is gehanteerd in het Referentiekader Taal bij het domein 'Begrippenlijst en Taalverzorging' (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a; Van der Beek & Paus, 2011). De begrippenlijst bevat begrippen en concepten die leerlingen en docenten nodig hebben om over taal van gedachten te wisselen. Een groot deel van deze begrippen heeft betrekking op grammaticale kennis, de zogenoemde grammaticale begrippen die nodig zijn voor de werkwoordspelling. Juist deze begrippen hebben een centrale plaats gekregen in de taken Grammatica. De begrippen die betrekking hebben op bijvoorbeeld stilistiek en semantiek hebben we buiten beschouwing gelaten bij het opstellen van de toetsmatrijs omdat deze geen betrekking hebben op grammatica in het algemeen.
Na het vaststellen van de indeling in twee hoofdcategorieën is op basis van een methode-analyse de indeling vervolgens verder verfijnd. Van elk grammaticaal begrip is nagegaan of dit aan de orde komt in de huidige taalmethoden en zo ja, op welk moment voor het eerst. In het toetspakket is het overzicht van de gehanteerde grammaticale begrippen voor groep 6 tot en met 8 opgenomen, als bijlage bij de handleiding. U treft het overzicht ook aan in bijlage 3.

Tabel 3.5 Verdeling Grammaticacategorieën groep 6, 7 en 8

Grammatica

Hoofdcategorie	**Verdeling grammaticacategorieën**					
	M6/E6 gewenst	**M6/E6 gerealiseerd**	**M7/E7 gewenst**	**M7/E7 gerealiseerd**	**M8 gewenst**	**M8 gerealiseerd**
Woordbenoeming en benoeming werkwoorden	11	11	12	14	8	9
Zinsvormen en zinsontleding	9	9	8	6	12	11

Bij grammatica is duidelijk te zien dat de gewenste verdeling over het algemeen gerealiseerd is. In groep 7 zien we een iets groter verschil dan in de groepen 6 en 8, maar op psychometrische gronden hebben we besloten de indeling enigszins aan te passen. Opvallend is dat de verhouding tussen woordbenoeming en zinsvormen en zinsontleding kantelt na groep 7. In groep 7 is het aantal subcategorieën binnen woordbenoeming en benoeming substantieel hoger in het onderwijsaanbod dan binnen zinsvormen en zinsontleding. Pas in groep 8 worden alle onderscheiden subcategorieën in het onderwijs aangeboden en hebben we deze ook in de taak terug laten komen.

Met betrekking tot de samenstelling van alle taken taalverzorging kunnen we stellen dat ieder deelgebied van taalverzorging een groot aantal (sub)categorieën bevat. Bij het samenstellen van de definitieve taken zijn met betrekking tot itemselectie de volgende inhoudelijke criteria aangehouden: alle onderscheiden taalverzorgings-categorieën behoren in principe in de taken taalverzorging voor te komen. Verder moet de verdeling van opgaven over categorieën en taken zo gelijkmatig mogelijk zijn om de inhoudsgebieden zo veel mogelijk te kunnen dekken. Een (sub)categorie zou minimaal een keer moeten voorkomen, en in een aantal gevallen twee, drie of vier keer, al naar gelang het aantal categorieën per deelgebied.
Het eerste criterium leverde geen probleem op. Het tweede criterium, een zo gelijkmatig mogelijke verdeling van het aantal opgaven per categorie, is waar mogelijk aangehouden. Factoren die van invloed waren op de verdeling zijn o.a.: het verschil in aantal categorieën per deelgebied, en het verschil in moeilijkheid van bepaalde categorieën. Sommige categorieën vereisen immers een hoger vaardigheidsniveau dan andere.
Er is telkens opnieuw een afweging gemaakt op basis van inhoudelijke en psychometrische informatie, waarbij psychometrische argumenten het soms wonnen van inhoudelijke. In een paar gevallen is daarom de verdeling van opgaven niet helemaal gelijkmatig. De uiteindelijke verdeling van aantallen opgaven per categorie is een zo goed mogelijk compromis tussen eisen van inhoudelijke aard en psychometrische kwaliteit.

Verdeling van de referentieniveaus over de toetsen
Bij de toetsen Taalverzorging is ervoor gekozen om pas vanaf groep 7 een referentieniveau te bepalen.
In groep 7 bieden we de taak spelling werkwoorden voor het eerst aan en wordt het vierde domein van taal, zoals het in het Referentiekader (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a)beschreven staat, pas volledig gedekt. De samenstelling van de toets is als volgt in zijn werk gegaan: in eerste instantie zijn opgaven geselecteerd op subdomein en categorie, vervolgens op moeilijkheid en uiteindelijk is aan de selectie een referentieniveau toegekend.
De toekenning van het referentieniveau is afhankelijk van de inhoud van de opgave en niet van de moeilijkheid. Bij het bepalen van het referentieniveau van een item maken we gebruik van de inhouds-omschrijving zoals gehanteerd in de publicatie van het referentiekader Taal en Rekenen (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a).

Om te bepalen welk referentieniveau Taalverzorging een leerling behaald heeft, is het noodzakelijk dat in de taken voldoende opgaven van beide referentieniveaus opgenomen zijn. Psychometrisch gezien is een minimum vereiste van 20 opgaven 1F en 15 opgaven 2F vastgesteld. Aan die eis voldoen de toetsen taalverzorging in groep 7 en 8 ruimschoots. De minimum aantallen zijn in deze verhouding (4:3) vastgesteld om te waarborgen dat de ankering die door middel van het referentieonderzoek met de Ankersets Taalverzorging tot stand is gekomen, te waarborgen. Deze ankering wordt in de volgende hoofdstukken beschreven. De reden dat er minder items 2F vereist zijn, ligt in het feit dat referentieniveau 2F ook niveau 1F omvat.
De optimale verdeling die gebaseerd is op de verhoudingen zoals psychometrisch wenselijk is en die in de *kolom gewenst* staat, is niet behaald in groep 8, maar de aantallen wijken slechts in geringe mate daarvan af. In groep 8 is het niet voor de hand liggend om aanzienlijk meer opgaven van referentieniveau 1F op te nemen, aangezien we bij de opgaven 1F al snel tegen een plafondeffect aanlopen.

Tabel 3.6 Referentieniveaus in de toetsen Taalverzorging voor groep 6 tot en met 8

Referentieniveaus

	Gewenst		Gerealiseerd	
	groep 7	**groep 8**	**groep 7**	**groep 8**
1F	45 (min. 20)	43 (min. 20)	45	41
2F	35 (min. 15)	37 (min. 15)	35	39

3.2.2 Itemconstructie, onderzoeken en selectie van opgaven voor de toetsen Taalverzorging

Itemconstructie
Alle opgaven die in de toetsen Taalverzorging zijn opgenomen werden voor deze toetsen geconstrueerd door een speciaal hiervoor samengestelde constructiegroep. De groep bestond uit leerkrachten uit het basisonderwijs en een taalverzorgingexpert. De constructiegroepleden construeerden opgaven bij de verschillende taalverzorgingcategorieën. Zij kregen voor elk van de deelgebieden uitgebreide constructievoorschriften. Een toetsdeskundige heeft waar nodig de opgaven aangepast om hieraan te voldoen.

Spelling
Voor spelling hebben we een nieuwe meerkeuzeopgavenvorm ontwikkeld. Aan de hand van meerkeuzevragen die bestaan uit vier ‘gelijke’ zinnen gaan leerlingen op zoek naar de zin waarin de twee dikgedrukte woorden *juist* gespeld zijn. Hiermee hebben we gehoor gegeven aan de roep uit het veld om niet langer leerlingen te vragen het fout gespelde woord aan te laten wijzen. De opgaven zijn zo samengesteld dat de alternatieven van de meerkeuzevragen betrekking hebben op twee spellingcategorieën. Dat betekent dat in elke opgave over het algemeen meerdere spellingcategorieën zijn opgenomen. Hiermee wordt een brede dekking van het domein gerealiseerd. In bijlage 4 treft u een aantal voorbeelden aan van de opgaven spelling.

In het toetsen van spellingvaardigheid is onderscheid te maken tussen actieve spelling (dicteeopgaven) en passieve spelling (meerkeuzeopgaven). In de bovenbouw van het basisonderwijs worden leerlingen steeds meer geacht hun eigen schrijfwerk – en vaak ook dat van medeleerlingen – na te kijken. Vandaar dat passieve spelling met name in de bovenbouw ook van belang is.
In de derde generatie toetsen Spelling van het Cito Volgsysteem primair en speciaal onderwijs is ervoor gekozen om in de toetsen Spelling alleen actieve spelling op te nemen en in de toetsen Taalverzorging alleen passieve spelling. Wij bieden nu zowel meerkeuze- als dicteeopgaven aan in onze toetsen. Alleen hebben we de twee vormen uit elkaar gehaald en in aparte toetsen opgenomen. Analyses van de itemgegevens van de tweede generatie toetsen lieten zien dat dicteeopgaven en meerkeuzeopgaven op één schaal pasten.
We toetsen nog steeds dezelfde vaardigheid: er is continuïteit tussen de toetsen uit de tweede generatie en de nieuwe toetsen. Uit de evaluatieformulieren van de onderzoeken blijkt dat zowel leerkrachten als leerlingen tevreden zijn over de gehanteerde nieuwe opgavevorm en de opsplitsing in actieve en passieve spelling.

Interpunctie
Voor interpunctie hebben we verschillende opgaventypen ontwikkeld die we afwisselend aanbieden.
Bij interpunctie hebben we er bewust voor gekozen om nergens leestekens of hoofdletters in de opgaven te plaatsen, omdat hiermee mogelijk het antwoord op andere opgaven kan worden weggeven. De meeste leerlingen weten bijvoorbeeld dat na een punt een hoofdletter geplaatst moet worden. Als we een punt zouden plaatsen in opgaven die de hoofdletter bevragen, zouden de opgaven erg eenvoudig worden.
Bij interpunctie beschouwen we inzicht in de zinsstructuur als een onderliggende vaardigheid. Als er geen leestekens staan, moet de leerling de zinstructuur proberen te doorgronden om te bepalen waar de leestekens en hoofdletters behoren te staan. Als een leerling zelf een tekst gaat schrijven, zal hij ook op basis van de regels voor interpunctie en zinsbouw de leestekens en hoofdletters plaatsen. Dat is wat wij met deze opgaven proberen te bewerkstelligen. Uit de evaluatieformulieren van de onderzoeken blijkt dat leerkrachten

aanvankelijk moeten wennen aan de opgaven interpunctie waarin geen leestekens staan, maar dat de leerlingen er geen moeite mee hebben.
We hebben getracht de verschillende opgaventypes zo veel mogelijk gelijkmatig op te nemen in de toetsen. In bijlage 4 treft u een aantal voorbeelden aan van de verschillende opgaventypen. De verschillende opgaventypen kunnen als volgt omschreven worden:

- Opgaven met vier zinnen waarover een vraag wordt gesteld.
- Opgaven met een of meerdere stamzinnen waarover een vraag wordt gesteld en waarbij de antwoordalternatieven gedeeltelijke variaties van deze zin zijn.

Grammatica
Voor de taken Grammatica hebben we verschillende meerkeuzeopgaventypen ontwikkeld die we afwisselend aanbieden. We maken gebruik van verschillende opgaventypen om zo de flexibiliteit van de toepassing van de grammaticale kennis te toetsen. Het opgaventype hangt ook af van de te bevragen categorie. De ene categorie leent zich beter voor een bepaald opgaventype dan het andere. We hebben getracht de verschillende varianten van de opgaventypes zo evenwichtig mogelijk te verdelen over de toetsen. In bijlage 4 treft u een aantal voorbeelden aan van de verschillende opgaventypen. De verschillende opgaventypen kunnen als volgt omschreven worden:

- Opgaven met vier zinnen waarin steeds een woord is onderstreept waarover een vraag wordt gesteld.
- Opgaven met een stamzin waarover een vraag wordt gesteld en waarbij de antwoordalternatieven gedeeltelijke variaties van deze zin zijn.

Proeftoetsing en kalibratie-analyses
De opgaven zijn eerst in proefafnames voorgelegd aan leerlingen in de jaargroep waarvoor ze bedoeld waren. Het primaire doel van dergelijke proefafnames is het verkrijgen van informatie over de moeilijkheid van de afzonderlijke opgaven. Tevens kunnen opgaven met een laag discriminerend vermogen geïdentificeerd en verwijderd worden. Dit zijn opgaven die geen of onvoldoende onderscheid maken tussen vaardigere en minder vaardige leerlingen. Daarnaast hebben wij de proefafname aangegrepen als een mogelijkheid om aan de deelnemende leerkrachten te vragen of zij inhoudelijke of andersoortige bezwaren hadden tegen bepaalde opgaven.
Bij proeftoetsingen is in 2013 halverwege de leerjaren 6, 7 en 8 een aantal opgaven voorgelegd aan leerlingen. Het proeftoetsdesign was een onvolledig design maar het was wel verbonden door middel van blokjes met ankeropgaven. Op het afnamemoment medio groep 6 zijn 211 opgaven geproeftoetst verdeeld over 7 boekjes. Op het afnamemoment medio groep 7 zijn 280 nieuwe opgaven verdeeld over 7 boekjes en op het afnamemoment medio groep 8 272 nieuwe opgaven verdeeld over 6 boekjes. Elke deelnemende school maakte meerdere taken. Een boekje voor leerjaar 6 bestond uit drie taken: een taak spelling niet-werkwoorden (20 opgaven), een taak interpunctie (20 opgaven) en een taak grammatica (20 opgaven). Een boekje voor leerjaar 7 bestond naast deze taken ook nog uit een taak spelling werkwoorden (20 opgaven). Een boekje voor leerjaar 8 bestond uit dezelfde taken als een boekje voor leerjaar 7, met als verschil dat de taak grammatica uit 30 opgaven bestond in plaats van 20. De nieuwe opgaven werden door minimaal 280 leerlingen maar het merendeel van de opgaven is door minimaal 600 leerlingen gemaakt.
Na de afnames zijn de antwoorden van de leerlingen op de toetsen geanalyseerd met behulp van het programmapakket One Parameter Logistic Model (OPLM; Verhelst, 1993; Verhelst en Glas, 1995). Zie voor een algemene technische beschrijving van dit model paragraaf 2.4.2.
Bij de analyses is de kwaliteit van de afzonderlijke items en van de totale opgavenverzameling voor een afnamemoment in kaart gebracht. Itemparameters en discriminatieparameters zijn geschat. Bij de analyses van de antwoorden van de leerlingen op de opgaven is nagegaan of de verschillende onderdelen een beroep doen op hetzelfde complex aan vaardigheden. Dat bleek (voor de meeste opgaven) het geval te zijn; opgaven die niet voldeden vielen af.
Na de uitwerking van de opgaven door toetsdeskundigen van Cito zijn de opgaven gescreend door praktijkdeskundigen uit het (S)BO. Hierbij is erop gelet dat de opgaven geschikt zijn voor een zo groot mogelijke groep leerlingen, ook voor leerlingen met extra onderwijsbehoeften. Opgaven waar de praktijkdeskundigen opmerkingen bij hadden, hebben we waar mogelijk aangepast of verwijderd.
Normeringsonderzoek

Op basis van de psychometrische analyses en de evaluaties van de proeftoetsing hebben we de opgaven geselecteerd voor het normeringsonderzoek. De psychometrische criteria betroffen met name de moeilijkheidsgraad en discriminatieparameter. Voor een evenwichtige samenstelling van de toetsen hebben we gelet op de verdeling van items over de verschillende taalverzorgingcategorieën. Waar mogelijk hebben we bij de opgavenselectie rekening gehouden met de opmerkingen die de leerkrachten gemaakt hebben. Er waren soms opgaven waarvan leerkrachten aangaven ze niet geschikt te vinden voor leerlingen in een bepaalde groep. Meestal is een dergelijke opgave dan ook niet opgenomen in het normeringsonderzoek. In een enkel geval zijn we daarvan afgeweken, omdat er anders te weinig geschikte opgaven zouden overblijven in de betreffende categorie.
Alle opgaven met een acceptabele moeilijkheid (in klassieke termen een P-waarde tussen .40 en .90) (r_{ir} vanaf .20) kwamen in principe in aanmerking voor opname in de normeringsonderzoeken Taalverzorging. In een enkel geval hebben we toch een opgave opgenomen met een P-waarde lager dan .40.

De opgaven die na de proefafname geselecteerd waren plus een aantal extra geconstrueerde opgaven werden vervolgens ingedeeld voor opname in de normeringsonderzoeken. De normeringsonderzoeken vonden plaats in 2014 en 2015, waarbij in 2014 halverwege en eind leerjaar 6 en 7 genormeerd is en in 2015 halverwege en eind leerjaar 7 en 8.
In tegenstelling tot de proefafnames, waar opgaven willekeurig over toetsboekjes waren verdeeld, zijn in de normeringsonderzoeken de taken zodanig samengesteld dat ze in principe al de vorm en inhoud hadden van de definitief uit te geven taken. In elke taak zaten bijvoorbeeld opgaven van uiteenlopende moeilijkheid. Ondanks proeftoetsing van opgaven kan het altijd voorkomen dat sommige opgaven in een normerings-onderzoek alsnog onvoldoende functioneren. Daarom komen de taken in het normeringsonderzoek niet volledig overeen met de definitief uit te geven taken.

Samenstelling definitieve toetsen
Na het normeringsonderzoek is van alle opgaven opnieuw de P-waarde en de r_{ir} bepaald. Ook nu kwamen in principe alle opgaven met een acceptabele moeilijkheid (in klassieke termen een P-waarde tussen .40 en .90) die door de vaardigere leerlingen vaker goed werden gemaakt dan door de minder vaardige leerlingen (r_{ir} vanaf .20) in aanmerking voor opname in de definitieve toetsen Taalverzorging.

Tijdens de selectie voor de definitieve toetsen is in eerste instantie gekeken of de eerder samengestelde taken gehandhaafd konden worden. Er vielen slechts enkele opgaven af vanwege een te lage r_{ir}, een te hoge of juist een te lage P-waarde. Soms viel een opgave af die psychometrisch gezien goed functioneerde, maar die tot een taalverzorgingscategorie behoorde die al voldoende vertegenwoordigd was in de toets. Daarentegen werden soms opgaven gehandhaafd die eigenlijk wat te gemakkelijk of juist te moeilijk waren, maar waarvoor in de betreffende taalverzorgingsscategorie geen beter functionerende alternatieven voorhanden waren. Bij elke individuele opgave vond een afweging plaats op zowel inhoudelijke als psychometrische gronden. Daarbij zijn ook reacties vanuit het veld meegenomen.

Bij de samenstelling van de toetsen is rekening gehouden met een mogelijk volgorde-effect: de volgorde van de opgaven in de definitieve toetsen wijkt minimaal af van die van het normeringsonderzoek. Onderzoek naar de beperkte wijzigingen in volgorde liet zien dat de wijziging in de volgorde geen effect had.

Bij het samenstellen van de definitieve toetsen zijn de volgende *inhoudelijke* criteria aangehouden:
- Als in de taal- en of spellingmethoden in een bepaald leerjaar bepaalde categorieën werden behandeld, dan wilden wij die categorieën in de toets terug laten komen.
- De verdeling van opgaven over categorieën en taken moest zo gelijkmatig mogelijk zijn.
- De verdeling van opgaven over referentieniveaus moest zo gelijkmatig mogelijk zijn.

De toetsen bevatten opgaven van uiteenlopende moeilijkheidsgraad. De toetsen zijn hierdoor geschikt om verschillen tussen leerlingen in beeld te brengen. Een goede illustratie hiervan en van de samenstelling van de toetsen zijn de figuren in bijlage 5: p50- en p80-kanspunten van de opgaven in de toetsen voor groep 6, 7 en 8 in relatie tot de gemiddelde vaardigheidsscore voor de afnamemomenten. In deze figuren is zichtbaar

dat de toetsen opgaven bevatten van uiteenlopende moeilijkheidsgraad. In de figuren is de verdeling van de opgaven over de taken van de toetsen visueel weergegeven. De balkjes in de figuren geven het p50- (onderkant van het balkje) en p80-kanspunt (bovenkant van het balkje) van elke opgave aan. Het p50-punt geeft de vaardigheidsscore aan waarbij er sprake is van een kans van 50% om de betreffende opgave goed te beantwoorden.

Bij de toetsen M6 Interpunctie en M6 Spelling niet-werkwoorden ligt bijvoorbeeld het merendeel van de balkjes op en rond de gemiddelde vaardigheidsscore behorende bij medio groep 6. Er is een goede spreiding in moeilijkheidsgraad van de opgaven: er zijn makkelijke opgaven (die liggen onder de lijn van M6), opgaven van gemiddelde moeilijkheid (doorkruisen de lijn van M6) en (erg) moeilijke opgaven (liggen boven de lijn van M6). Bij deze toetsen kunnen ook de betere leerlingen laten zien wat ze in huis hebben. Voor het afnamemoment E6 liggen de opgaven hoger op de vaardigheidsschaal. Zoals we zagen in tabel 2.1 is er sprake van een lichte groei in vaardigheid tussen de momenten M6 en E6. Bij de toets M7 Spelling werkwoorden zien we dat veel opgaven boven de lijn van M7 liggen. Het is zichtbaar dat deze toets relatief veel moeilijke opgaven bevat. Het is voor leerlingen het eerste moment dat werkwoordspelling in de toetsen Taalverzorging aan bod komt. De andere deelgebieden geven echter een beeld waarbij de spreiding rond de gemiddelde vaardigheidsgroei ligt.

3.3 Statistische beschrijving

3.3.1 Itemkenmerken: moeilijkheidsgraad en interne consistentie

Wat de moeilijkheid van de opgaven betreft: voor de opgavenselectie geldt het uitgangspunt dat de P-waarden bij voorkeur tussen 0,40 – 0,90 moeten liggen en dat de opgaven van Taalverzorging gemiddeld een P-waarde tussen de 0,55 en 0,75 hebben. In tabel 3.3 rapporteren we de geschatte range van P-waarden en de geschatte gemiddelde P-waarde van de opgaven voor de verschillende meetmomenten van de toets Taalverzorging voor groep 6 tot en met 8. Daarnaast zijn ook gegevens opgenomen over de R_{it}-waarden van de opgaven, waarbij de toetsscore over het betreffende onderdeel het uitgangspunt was voor de berekening van de coëfficiënt. R_{ir}-waarden zijn wellicht te prefereren omdat zij een realistischer beeld geven van de correlatie met de schaalscore, maar helaas zijn ons geen normgegevens bekend voor R_{ir}. Voor R_{it}-waarden kent het COTAN-beoordelingssysteem (Commissie Test Aangelegenheden Nederland van het Nederlands Instituut van Psychologen, zie Evers et al., 2010) wél kwaliteitscriteria.

Voor alle deeltoetsen over de leerjaren op beide toetsmomenten blijken de P-waarden redelijk overeen te komen met de gekozen uitgangspunten. De gemiddelden liggen respectievelijk op 0,58 en 0,75. Voor de minima per meetmoment geldt dat slechts voor drie items de P-waarde onder de 0,40 uitkomt: een item M7 spelling niet-werkwoorden, een item M8 spelling niet-werkwoorden en een item M8 interpunctie. De maximale P-waarden per meetmoment komt bij één moment (E6) iets boven de 0,90 uit, maar het betreft een uitzondering: het gaat om slechts één item spelling niet-werkwoorden. Alle andere items komen onder de 0,90 uit. De gemiddelde R_{it}-waarden zijn voor alle deeltoetsen over de leerjaren op beide toetsmomenten te kenschetsen als ‘goed’ (gemiddelde $R_{it} > 0,30$). Praktisch alle opgaven kennen een waarde van 0,30 of hoger. Er zijn slechts enkele opgaven die in de categorie voldoende vallen (tussen 0,20 en 0,30). Dat zijn twee opgaven spelling werkwoorden die voorkomen in de toetsen M7/E7 (R_{it} 0,27 en 0,26) en een opgave spelling werkwoorden in de toets M8 R_{it} 0,27). In de tabellen is de volgorde gehanteerd van voorkomen in de toetsen.

Tabel 3.7a[5] *Range en gemiddelde van p- en R_{it}-waarden voor de toets M6 voor de verschillende taalverzorgingsonderdelen*

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
M6 IP	0,5-0,85	0,67	0,36-0,56	0,46	20

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
M6 SN	0,44-0,88	0,69	0,34-0,59	0,43	20

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
M6 GR	0,49-0,67	0,58	0,31-0,56	0,47	20

Tabel 3.7b Range en gemiddelde van p- en R_{it}-waarden voor de toets E6 voor de verschillende taalverzorgingsonderdelen

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
E6 IP	0,57-0,90	0,74	0,37-0,55	0,46	20

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
E6 SN	0,5-0,92	0,75	0,38-0,54	0,44	20

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
E6 GR	0,57-0,77	0,69	0,35-0,6	0,51	20

[5] In de tabellen is de volgorde van de deelgebieden gehanteerd zoals ze in de toetsen voorkomen: 1. interpunctie; 2. spelling niet-werkwoorden; 3. grammatica; 4. spelling werkwoorden (alleen in groep 7 en 8).

Tabel 3.7c Range en gemiddelde van p- en R_{it}-waarden voor de toets M7 voor de verschillende taalverzorgingsonderdelen

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
M7 IP	0,55-0,81	0,7	0,35-0,59	0,52	20

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
M7 SN	0,37-0,86	0,71	0,34-0,5	0,42	20

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
M7GR	0,49-0,79	0,66	0,34-0,59	0,46	20

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
M7 SW	0,42-0,84	0,6	0,26-0,44	0,34	20

Tabel 3.7d Range en gemiddelde van p- en R_{it}-waarden voor de toets E7 voor de verschillende taalverzorgingsonderdelen

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
E7 IP	0,59-0,85	0,73	0,33-0,55	0,49	20

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
E7 GR	0,55-0,83	0,71	0,37-0,61	0,49	20

	P-waarden		Rit-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
E7 SW	***0,47-0,88***	***0,65***	***0,27-0,46***	***0,36***	***20***

Tabel 3.7e Range en gemiddelde van p- en R_{it}-waarden voor de toets M8 voor de verschillende taalverzorgingsonderdelen

	P-waarden		Rit-waarden		N items
	Range	Gem.	Range	Gem.	
M8 IP	0,3-0,89	0,73	0,38-0,6	0,5	20

	P-waarden		Rit-waarden		N items
	Range	Gem.	Range	Gem.	
M8 SN	0,39-0,86	0,69	0,3-0,51	0,42	20

	P-waarden		Rit-waarden		N items
	Range	Gem.	Range	Gem.	
M8 GR	0,55-0,85	0,71	0,33-0,61	0,47	20

	P-waarden		Rit-waarden		N items
	Range	Gem.	Range	Gem.	
M8 SW	0,42-0,83	0,59	0,28-0,46	0,37	20

3.3.2 Verdeling van de ruwe scores

In tabel 3.4 zijn de verdelingskarakteristieken gegeven van de ruwe scores op de verschillende toets-momenten. De gemiddelden komen uiteraard overeen met wat men bij een gegeven aantal items mag verwachten bij de gekozen (gemiddelde) moeilijkheidsgraad. De gemiddelde moeilijkheidsgraad voor de afnamemomenten bij bijvoorbeeld interpunctie loopt op van 0,58 (M6) tot 0,71(M8). De verdelingen zijn over het algemeen linksscheef (vergelijk de negatieve waarden in de kolom 'skewness'), de ene wat meer dan de andere. De verdelingen zijn ééntoppig en lijken vrij sterk op elkaar. Voor een grafische weergave zie de histogrammen van de scores op beide momenten in de figuren 3.1. Duidelijk te zien is dat in de grafische weergave vrijwel alle deelgebieden dezelfde curve vertonen en dat de waarden zijn opgeschoven tussen medio- en de eindafname. Bij grammatica zien we tussen M6 en E6 een meer symmetrische verdeling die minder gepiekt is. Dit is ook in tabel 3.4c waar te nemen bij de waarden voor Skewness en Kurtosis.
Bij Spelling werkwoorden M7, E7 en M8 zien we een vergelijkbare symmetrische verdeling. Zowel de taak Grammatica als de taken Spelling werkwoorden hebben een iets hogere moeilijkheidsgraad waardoor de scores dichterbij elkaar komen te liggen.

Tabel 3.4a Verdelingskenmerken van de afnamemomenten M6, E6, M7, E7, M8 Interpunctie

Interpunctie Meetmoment	Aantal Opgaven	M	SD	Skewness	Kurtosis
M6	20	13,4	4,26	-0,618	-0,288
E6	20	14,8	3,92	-0,941	0,433
M7	20	13,9	4,65	-0,748	-0,243
E7	20	14,6	4,24	-0,881	0,156
M8	20	14,5	4,28	-0,870	0,130

Tabel 3.4b Verdelingskenmerken van de afnamemomenten M6, E6, M7, E7, M8 Spelling niet-werkwoorden

Spelling niet-werkwoorden Meetmoment	Aantal Opgaven	M	SD	Skewness	Kurtosis
M6	20	13,8	3,86	-0,683	-0,010
E6	20	15,0	3,65	-0,950	0,603
M7	20	14,1	3,61	-0,738	0,203
E7	20	14,7	3,59	-0,893	0,525
M8	20	13,8	3,71	-0,649	-0,023

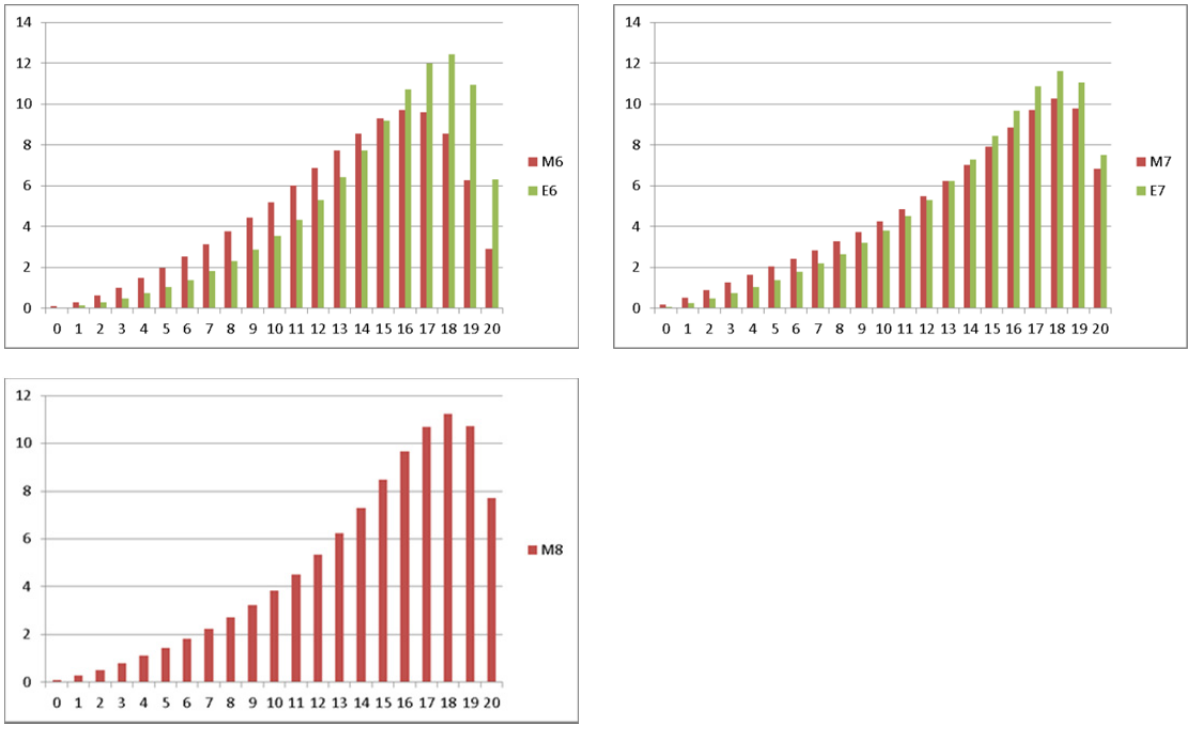
Tabel 3.4c Verdelingskenmerken van de afnamemomenten M6, E6, M7, E7, M8 Grammatica

Grammatica Meetmoment	Aantal Opgaven	M	SD	Skewness	Kurtosis
M6	20	11,7	4,56	-0,304	-0,737
E6	20	13,9	4,62	-0,781	-0,186
M7	20	13,1	4,22	-0,581	-0,342
E7	20	14,3	4,29	-0,861	0,092
M8	20	14,2	4,15	-0,791	-0,015

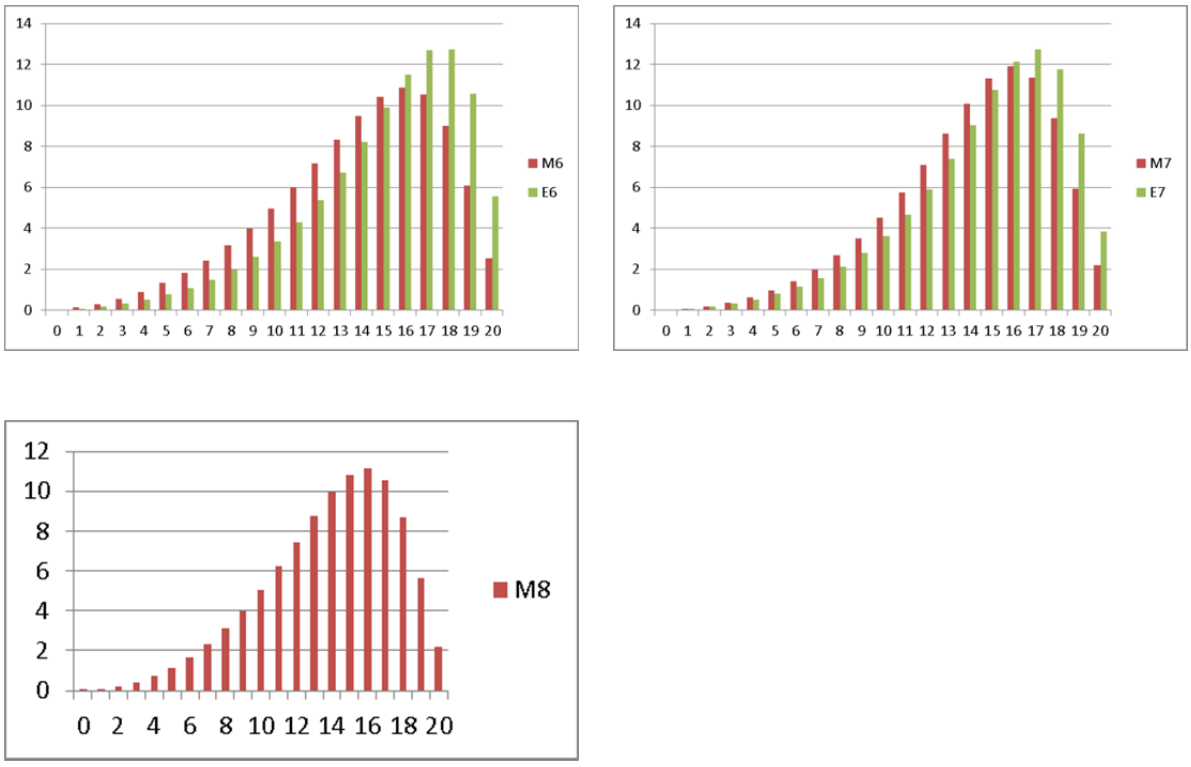
Tabel 3.4d Verdelingskenmerken van de toetsmomenten M7, E7, M8 Spelling werkwoorden

Spelling werkwoorden Meetmoment	Aantal Opgaven	M	SD	Skewness	Kurtosis
M7	20	12,0	3,18	-0,242	-0,275
E7	20	12,9	3,26	-0,391	-0,184
M8	20	11,7	3,56	-0,250	-0,401

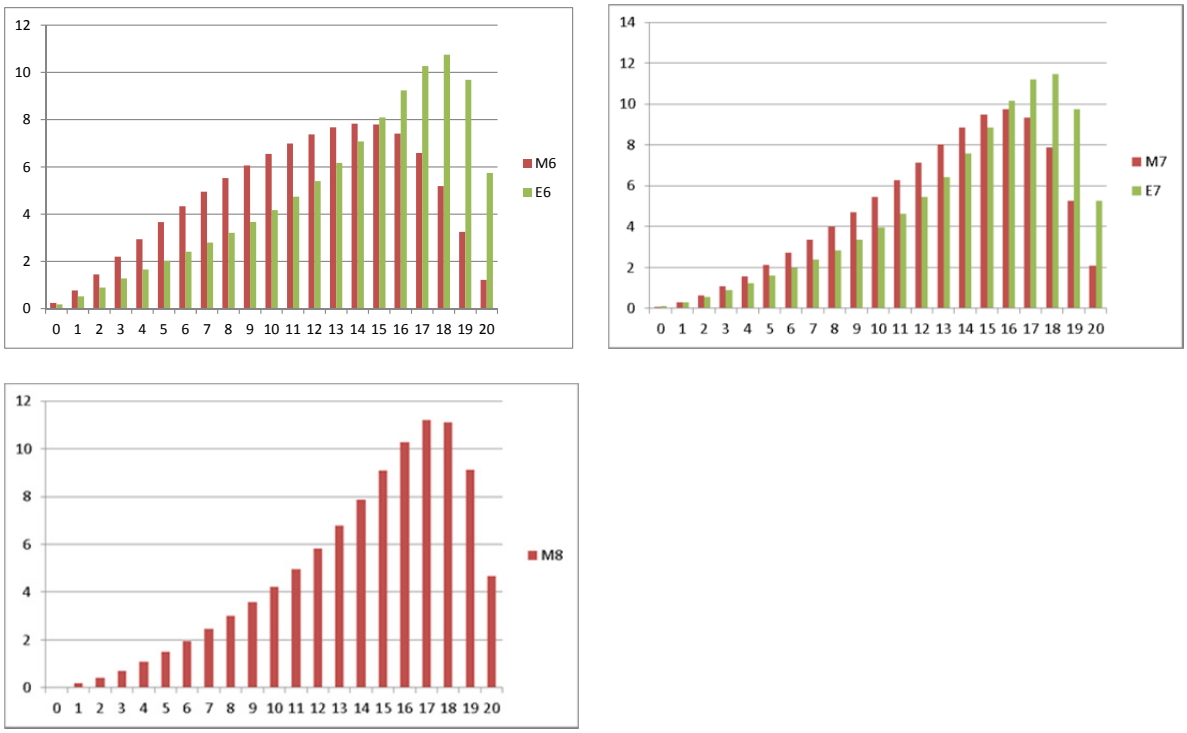
Figuur 3.1a Histogrammen van de toetsscores op de afnamemomenten M6, E6, M7, E7 en M8 Interpunctie



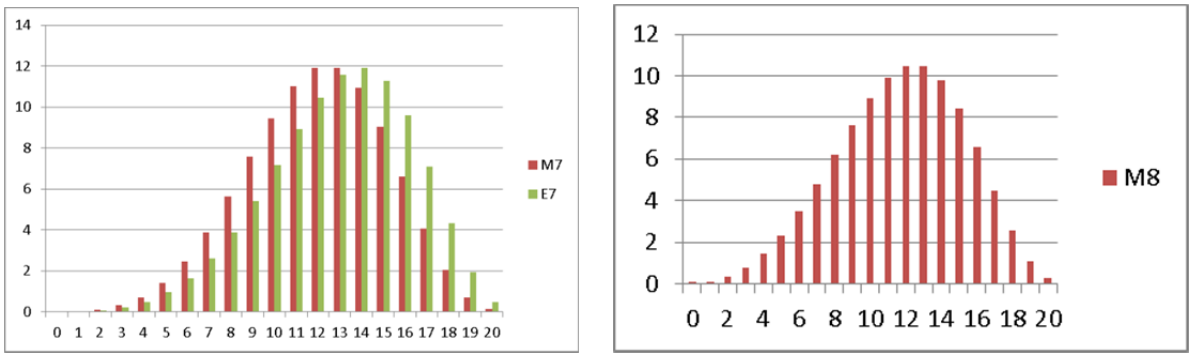
Figuur 3.1b Histogrammen van de toetsscores op de afnamemomenten M6, E6, M7, E7 en M8 Spelling niet-werkwoorden



Figuur 3.1c Histogrammen van de toetsscores op de afnamemomenten M6, E6, M7, E7 en M8 Grammatica



Figuur 3.1d Histogrammen van de toetsscores op de afnamemomenten M7, E7 en M8 Spelling werkwoorden



4 Kalibratie en normering

4.1 Opzet en verloop van het kalibratie- , normerings- en referentieonderzoek

Met het oog op het ontwikkelen van de toetsen Taalverzorging voor groep 6 tot en met 8 zijn in 2012 opgaven geconstrueerd voor de vier deelgebieden. In mei 2013 zijn deze opgaven in een kalibratieonderzoek (proefonderzoek) voorgelegd aan leerlingen van groep 6 tot en met 8 op een groot aantal scholen om gegevens te verzamelen over de kwaliteit en de moeilijkheid van de opgaven. Aansluitend zijn bij een landelijke normgroep referentiegegevens verzameld door de psychometrisch en inhoudelijk meest geschikte opgaven voor te leggen aan leerlingen op de normeringsmomenten medio en einde schooljaar.
De normeringen voor het M-moment vonden plaats in januari en begin februari 2014 en 2015, de normeringen voor het E-moment vonden plaats in mei/begin juni 2014 en 2015.

Het kalibratieonderzoek
Het kalibratieonderzoek levert gegevens op over de kwaliteit en de moeilijkheid van de opgaven. In het kalibratieonderzoek, dat aan de opgavenbanken ten grondslag ligt, is uitgegaan van een onvolledig maar 'verbonden' design: niet alle leerlingen in de steekproef van het kalibratieonderzoek maakten alle opgaven. Opgaven werden verdeeld over taken en aan elke leerling werden meerdere taken voorgelegd. De taken die gezamenlijk aan een groep leerlingen worden voorgelegd, worden 'boekjes' ('booklets') genoemd.
De verschillende boekjes overlappen elkaar. Deze overlap zorgt ervoor dat het design verbonden is, een noodzakelijke voorwaarde om CML-schattingen van de itemparameters te kunnen bepalen. Er is sprake van overlap tussen de boekjes binnen een jaargroep, maar ook tussen de verschillende jaargroepen. Er zijn bijvoorbeeld blokjes van 5 items die in drie jaargroepen zijn afgenomen.

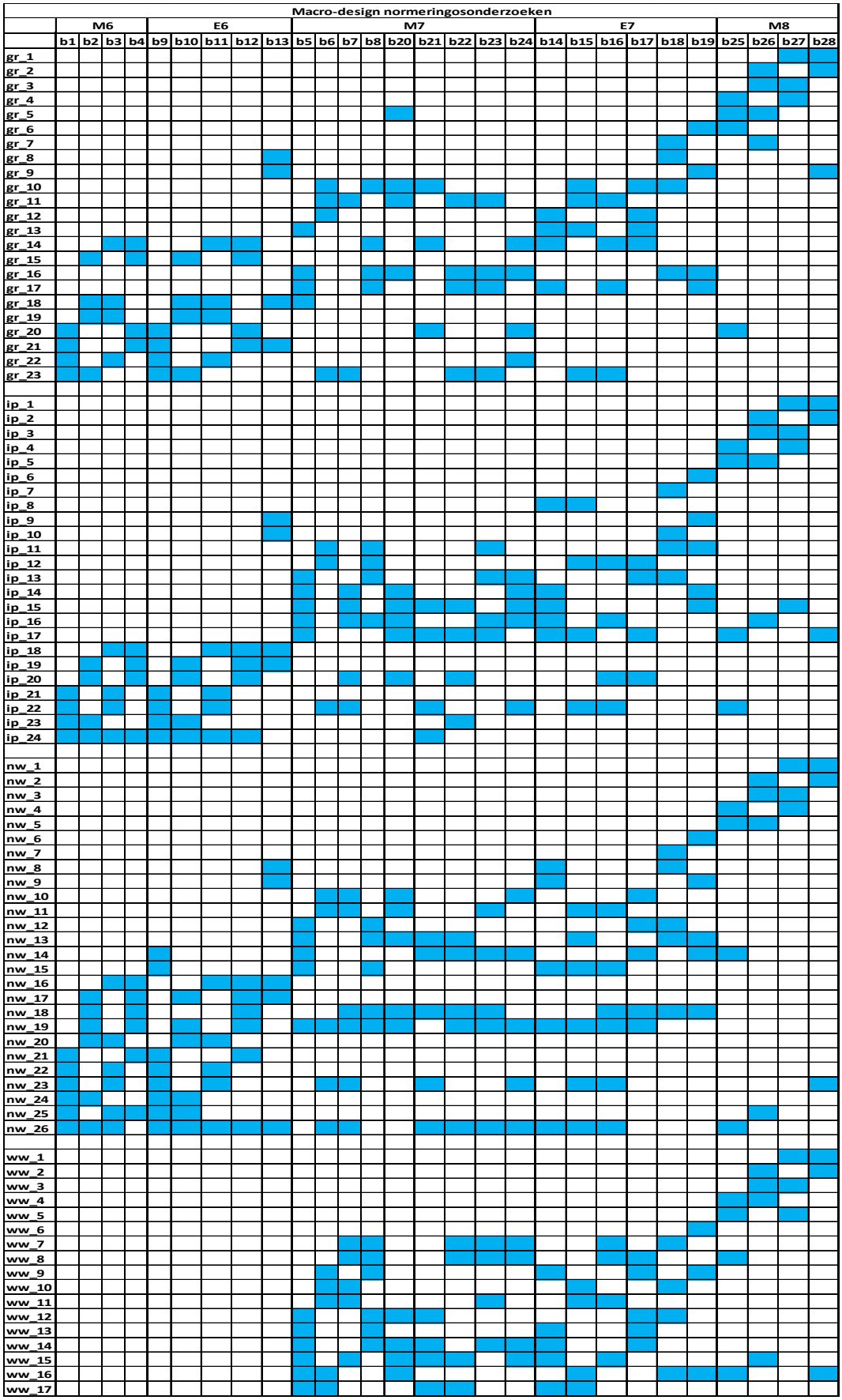
Op grond van de resultaten van de kalibraties die zijn uitgevoerd op de data die in 2013 in het kader van de proeftoetsing werden verzameld, zijn items geselecteerd voor drie toetsen, te weten de toets voor leerjaar 6 (toets M6/E6), leerjaar 7 (toets M7/E7) en leerjaar 8 (toets M8), bedoeld voor de afnamemomenten M en E in de betreffende leerjaren. De voor deze toetsen bedoelde items werden samen met een aantal aanvullende items in 2014 en 2015 voorgelegd aan 6200 leerlingen van groep 6, 7 en 8 in perioden van het schooljaar die overeenkwamen met de genoemde afnamemoment (medio: januari/begin februari; eind: mei/begin juni).
Het ging in totaal (dus inclusief de aanvullende opgaven) om 418 items.
We kozen ervoor om de opgaven in deze normeringsonderzoek op dusdanige wijze aan de leerlingen voor te leggen dat het mogelijk was in de kalibraties alle items met elkaar te verbinden. Dit werd gerealiseerd door bij de afnames het onvolledige, maar geheel verbonden design te hanteren dat is weergegeven in tabel 4.1.
Normaliter is een dergelijk design in de normeringsfase niet meer zo nodig omdat in de proeftoetsing al voldoende kalibratiegegevens beschikbaar zijn gekomen. In het geval van de toetsen Taalverzorging deden we dit voor alle zekerheid wél. Daarbij waren de volgende overwegingen bepalend:

- De toetsen Taalverzorging zijn betrekkelijk nieuw, zodat het niet mogelijk was de toetsen in de nieuwe samenstelling (dus bestaande uit opgaven voor spelling, interpunctie en grammatica) te verbinden met LVS-toetsen van eerdere leerjaren.
- Er was lange tijd onzekerheid over de keuze van de vaardigheidsschalen; uiteindelijk is gekozen voor kalibratie op basis van vier verschillende vaardigheidsschalen voor spelling niet-werkwoorden, spelling werkwoorden, interpunctie en grammatica.
- Misschien nog wel de belangrijkste reden om te kiezen voor een geheel verbonden design is het feit dat de referentiesets voor Taalverzorging pas in 2015 beschikbaar kwamen. Omdat de referentiecesuren op zowel de toetsen voor leerjaar 6 als die voor leerjaar 7 moesten worden overgebracht was het noodzakelijk de opgaven van deze toetsen onderling te kunnen verbinden.

Om het design beter te kunnen toelichten, halen nu het normeringsonderzoek M7 van januari 2015 even naar voren. Er zijn in totaal 155 items opgenomen in het onderzoek. Deze items waren verdeeld over vijf verschillende opgavenboekjes in een onvolledig, maar 'verbonden' design. Elk boekje bestond uit 80 items,

20 items per deelgebied. Elk item kwam in principe in twee boekjes voor. Soms kwam een blokje maar een keer voor in het design van deze jaargroep, maar dat blokje werd dan wel weer opgenomen in het onderzoek van een aangrenzend afnamemoment (E6 of E7). Het gemiddeld aantal leerlingantwoorden per item was 426.
Scholen uit de getrokken steekproef werden via een brief voor dit onderzoek uitgenodigd. De scholen ontvingen opgavenboekjes, antwoordbladen en een handleiding voor de docent. De toetsen taalverzorging werden afgenomen door de leerkracht aan de hand van de handleiding. De ingevulde antwoordbladen werden door Cito verwerkt en de scholen ontvingen een rapportage met de resultaten per leerling.
Voor de andere normeringsonderzoeken is een soortgelijke opzet gehanteerd. Het minimale aantal boekjes per normeringsonderzoek was vier (M6 en M8) en het maximale aantal boekjes was zes (bij het referentie-onderzoek E7; in dit onderzoek werden ook de ankeropgaven uit de referentiesets meegenomen).

Figuur 4.1 Afnamedesign normeringsonderzoek groep 6, 7, 8



Tabel 4.1 Aantal leerlingen per afnamemoment in de normeringsonderzoeken M6/E6, M7/E7 en M8

Afname-moment	Aantal leerlingen	Aantal scholen
M6	799	31
E6	1323	50
M7	1812	56
E7	1346	50
M8	941	35

De in tabel 4.1 genoemde aantallen waren ruimschoots voldoende voor het doel van het onderzoek (kalibreren en normeren). Hiervoor was een minimum van 400 waarnemingen per item vereist; voor het merendeel van de opgaven werd aan deze minimumeis voldaan. Bij een gering aantal items waren er minder dan 400 waarnemingen. Voor elk onderdeel is dit het geval bij 10 items. Doordat een gedeelte van de items op meerdere tijdstippen is afgenomen, is het aantal waarnemingen over het geheel genomen aanzienlijk hoger dan in tabel 4.2 is weergegeven. de getallen in de tabel zijn de aantallen waarnemingen per item per tijdstip. Voor M6 waren er gemiddeld 439 waarnemingen, voor M7 852 en M8 471. In tabel 4.2 staan de aantallen waarnemingen per item.

Tabel 4.2 Aantal waarnemingen per item in de normeringsonderzoeken M6/E6, M7/E7 en M8

	M6				E6				M7				E7				M8			
	k*	min	max	gem	k	min	max	gem	k	min	max	gem	k	min	max	gem	k	min	max	gem
IP	33	388	799	460	38	253	1070	628	44	375	1223	823	58	241	919	443	40	451	490	471
NW	34	388	799	458	44	253	1323	588	37	391	1630	1006	55	241	1138	490	38	451	490	471
GR	33	388	411	401	43	253	796	516	37	375	1061	781	47	241	897	464	39	451	490	471
WW									46	391	1412	799	54	241	870	452	36	451	490	471

* Het aantal opgaven duiden we met *k* aan.

Voor de normeringsonderzoeken M6/E6, M7/E7 en M8 werden scholen geworven na het trekken van een representatieve steekproef, waarbij rekening gehouden werd met verdeling over strata, schoolgrootte, regio en verstedelijking (zie paragraaf 4.2). Aangenomen werd dat de verdeling over sekse ongeveer 50/50 zou zijn; de verdeling is achteraf gecontroleerd: zie tabel 4.13.

Het referentieonderzoek

De toetsen Taalverzorging zijn geankerd aan de openbare Ankersets Taalverzorging 1F en 2F die ontwikkeld zijn door het College voor Toetsen en Examens (het CvTE) om toetsontwikkelaars, uitgeverijen en onderzoekers in de gelegenheid te stellen de landelijke prestatiestandaard van de referentieniveaus over te brengen op hun eigen producten taalverzorging door middel van een ankeronderzoek.
Het ankeronderzoek hebben we in juni 2015 uitgevoerd in groep waarbij leerlingen zowel (een deel van) de ankersets als (een deel van) de toetsen voor groep 7 gemaakt hebben. Door de volledige verbondenheid van het design konden de prestatiestandaarden op de toetsen voor groep 7 en 8 worden overgebracht.

Keuze en representativiteit van de opgaven

We hebben getracht een zo representatief mogelijk anker samen te stellen dat paste bij de inhoud van de toetsen Taalverzorging en bij de doelpopulatie. Alle subdomeinen of deelgebieden en vrijwel alle opgaven-types waren vertegenwoordigd. We hebben ervoor gekozen om de dicteeopgaven spelling niet mee te nemen als anker omdat er in de toetsen Taalverzorging alleen meerkeuzeopgaven voorkomen.
De ankers 1F en 2F bestonden ieder uit 20 opgaven. Voor elk referentieniveau waarop gerapporteerd moet worden, gelden volgens de richtlijnen zoals beschreven in de publicatie toelichting Ankeronderzoek met Ankersets (2015) dat een anker uit minimaal 15 opgaven per referentieniveau moet bestaan. Verder moet een anker per referentieniveau representatief zijn voor het hele domein. We hebben van alle subdomeinen vijf opgaven 1F en vijf opgaven 2F opgenomen. Daarmee wordt aan deze richtlijnen voldaan.
Er waren geen redenen om aan te nemen dat de ankeropgaven zich afwijkend zouden gedragen in de te ankeren toets, aangezien de constructen vergelijkbaar zijn. Het was niet nodig om meer ankeropgaven op te nemen dan de hierboven genoemde aantallen. Verder kwamen ook de afnamecondities overeen.
De ankerset is net als de toetsen, gebaseerd op een afname van opgaven op papier.

Ankertoetsdesign met intern anker

We hebben een toetsversie samengesteld waarbij de opgaven uit het anker gecombineerd zijn met alle opgaven uit de toets Taalverzorging voor groep 7. Er zijn in totaal 200 items opgenomen in het onderzoek. Deze items waren verdeeld over zes verschillende opgavenboekjes in een onvolledig, maar 'verbonden' design. Elk boekje was verbonden met de M7/E7- en M8-normeringsonderzoeken. Een boekje bestond uit 80 items, 20 items per deelgebied. Elk item kwam in twee boekjes voor. In vier boekjes is een blokje van 5 items uit de Openbare ankerset (in figuur 4.2 OS1 en OS2) opgenomen, aangevuld met 10 uitgave-items en 5 items uit de toets M8 of 5 items van een door Cito ontwikkelde marktset Taalverzorging (in figuur 4.2 MS1 en MS2). Elk ankerblokje bestond uit zowel items 1F als 2F.
Scholen uit de getrokken steekproef werden via een brief voor dit onderzoek uitgenodigd. De scholen ontvingen opgavenboekjes, antwoordbladen en een handleiding voor de docent. De toetsen taalverzorging werden afgenomen door de leerkracht aan de hand van de handleiding. De ingevulde antwoordbladen werden door Cito verwerkt en de scholen ontvingen een rapportage met de resultaten per leerling.
In figuur 4.2 is toegelicht hoe de ankeritems verbonden zijn met de opgaven in de toets.

Figuur 4.2 Afnamedesign referentieonderzoek E7

	Referentie onderzoek juni 2015						LVS normeringsonderzoeken		
	refM7_1	refM7_2	refM7_3	refM7_4	refM7_5	refM7_6	M7	E7	M8
OS_1									
OS_2									
MS_1									
MS_2									
A1									
A2									
A3									
A4									
A5									
A6									
A7									
A8									
A9									
A10									
A11									
A12									

Steekproeftrekking en dataverzameling
Aan het referentieonderzoek deden 1346 leerlingen mee. Per opgave hebben we gemiddeld 462 waarnemingen verzameld. Dat is ruim voldoende om te voldoen aan de richtlijnen die het CvTE opgesteld heeft om een stabiele ankering te bewerkstelligen (minimaal 400 waarnemingen per item). Over de representativiteit van de steekproef van het referentieonderzoek is meer te vinden in de volgende paragrafen over de samenstelling en representativiteit van de normeringssteekproeven, ook die ten aanzien van het afnamemoment E7 waar het referentieonderzoek aan is gekoppeld.

4.2 Samenstelling van de normeringssteekproef en representativiteit

Representativiteit van de normgroepen M6/E6, M7/E7 en M8
De representativiteit van de steekproeven voor de normeringsonderzoeken M6/E6, M7/E7 en M8 zijn geëvalueerd op verdeling over zogenoemde strata, schoolgrootte, regio, verstedelijking en sekse.

Strata
De strata die door Cito worden gehanteerd, zijn gebaseerd op de percentages gewichtenleerlingen. Het stratum kan worden opgevat als een indicatie voor het aantal achterstandsleerlingen van een school. De formatieomvang van een school wordt in belangrijke mate bepaald door het formatiegewicht van de leerlingen die samen de schoolpopulatie vormen. Voorheen werden de leerlingen gecategoriseerd in vijf formatiegewichten die een combinatie vormden van opleidingsniveau, sociaal-economische status en etnische herkomst van de ouders. Inmiddels is echter de categorisering naar leerlinggewichten vervangen door een andere regeling waarbij nog slechts drie niveaus worden onderscheiden en waarbij etniciteit geen rol meer speelt. De definities van de categorieën zijn aangescherpt en uitsluitend gebaseerd op het opleidingsniveau van de ouders. In tabel 4.3 staat de gewichtenregeling toegelicht.

Tabel 4.3 Gewichtenregeling

Gewicht	Toelichting
0,0	Leerling van wie één van de ouders of beide ouders een opleiding heeft gehad van minimaal 3 jaar vmbo-gemengde leerweg/vmbo theoretische leerweg/mavo, minimaal 2 jaar havo/vwo, mbo, hbo of universiteit
0,3	Leerlingen van wie beide ouders of de ouder die belast is met de dagelijkse verzorging een opleiding heeft gehad van maximaal lbo/vbo, praktijkonderwijs of vmbo basis- of kaderberoepsgerichte leerweg
1,2	Leerlingen van wie beide ouders een opleiding hebben gehad van maximaal basisonderwijs of (v)so-zmlk of van wie één van de ouders een opleiding heeft gehad van maximaal basisonderwijs of (v)so-zmlk en de andere ouder maximaal lbo/vbo praktijkonderwijs of vmbo basis- of kaderberoepsgerichte leerweg

Het percentage gewichtsleerlingen op een school wordt gebruikt om strata te construeren die gebruikt worden voor de steekproeftrekking van de onderzoeken voor het leerlingvolgsysteem. Daarnaast wordt de representativiteit van de steekproef aan de hand van een indeling in vier categorieën geëvalueerd.
De omschrijving van de categorieën is te vinden in tabel 4.4. In deze tabel zijn de categorieën gecombineerd met twee categorieën voor schoolgrootte (scholen van 200 leerlingen of minder versus scholen van meer dan 200 leerlingen). De representativiteit naar schoolgrootte wordt afzonderlijk geëvalueerd.

Tabel 4.4 Indeling strata

stratum	aantal leerlingen	aantal leerlingen	% gewichts-leerlingen	% scholen
1	>200	>200	> 20	6.2
2	<=200	<=200	> 20	9.6
3	>200	>200	> 10 & <= 20	8.7
4	<=200	<=200	> 10 & <= 20	11.8
5	>200	>200	> 5 & <= 10	11.6
6	<=200	<=200	> 5 & <= 10	12.3
7	>200	>200	<= 5	21.1
8	<=200	<=200	<= 5	18.7

Voor elke school is het percentage leerlingen ('p') met een gewicht afwijkend van 0,0 bepaald. Gebaseerd op deze gewichten zijn er vier groepen scholen gevormd. De CFI-gegevens van oktober 2014/2015 zijn als basis voor het steekproefkader voor de normeringen van M6/E6, M7/E7 en M8 genomen. De verdeling van de scholen over de strata wordt weergegeven in tabel 4.5.

Tabel 4.5 Scholen uit steekproeven M6, E6, M7, E7 en M8*, naar percentage leerlingen met een afwijkend leerlinggewicht*

Categorie		Landelijk		M6		E6		M7		E7		M8	
		%	Aantal	%	Aantal	%	Aantal	%	Aantal	%	Aantal	%	
1	P≤5%	42,6	8	28,6	18	40,0	23	37,7	18	40,9	13	43,3	
2	5%<P≤10%	22,7	6	21,4	8	17,8	12	19,7	8	18,2	7	23,3	
3	10%<P≤20%	19,4	10	35,7	14	31,1	21	34,4	14	31,8	8	26,7	
4	P>20%	15,4	4	14,3	5	11,1	5	8,2	4	9,1	2	6,7	

M6 χ^2_2 (3, N = 799) = 5,18; p = 0,16; ϕ = 0,08
E6 χ^2_2 (3, N = 1323) = 4,26; p = 0,23; ϕ = 0,06
M7 χ^2_2 (3, N = 1812) = 9,73; p = 0,02; ϕ = 0,07
E7 χ^2_2 (3, N = 1346) = 5,05; p = 0,17; ϕ = 0,06
M8 χ^2 (3, N = 941) = 2,30; p = 0,51; ϕ = 0,05

In de steekproef is categorie 3 (10% < P ≤ 20%) bij elk van de normeringsmomenten licht oververtegenwoordigd. Voor de verschillende normeringsmomenten is dat terug te zien in de significante chi-kwadraat-waarden. Verder is categorie 1 (P≤5%) bij het normeringsomoment M6 licht ondervertegenwoordigd. Voor alle andere categorieën zijn de steekproefafwijkingen niet significant. Met betrekking tot de significante chi-kwadraat-waarden kan gesteld worden dat bij steekproeven van deze grootte significantie niet veel informatie geeft. Het is beter om de effectgrootte ϕ als uitgangspunt te nemen.

Formule 4.1 Berekening van de effectgrootte ϕ

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

We zien dat de effectgroottes onder de .10 liggen en daarmee zeer klein zijn (cf. Cohen, 1988). Ze zijn met 0,08 nog het hoogst bij het afnamemoment M6. Deze effectgroottes zijn te interpreteren als klein tot

middelgroot. De conclusie is niettemin dat de normeringssteekproeven een redelijk goede afspiegeling vormen van de populatie. Om deze reden is statistische weging van de resultaten van de normeringssteekproef niet nodig.

Schoolgrootte

Wat betreft *schoolgrootte* kan onderscheid worden gemaakt tussen grote scholen (> 200 leerlingen) en kleine scholen (≤ 200 leerlingen). De percentages grote en kleine scholen in de steekproeven kunnen worden vergeleken met de percentages in de populatie De verschillen zijn terug te vinden in tabel 4.6.

Tabel 4.6 Scholen uit steekproef M6/E6, M7/E7 en M8 *naar schoolgrootte*

	Landelijk					**Normering**					
	Populatie	**M6**		**E6**		**M7**		**E7**		**M8**	
	%	**Aantal**	**%**	**Aantal**	**%**	**Aantal**	**%**	**Aantal**	**%**	**Aantal**	**%**
Groot	52,40%	18	64,3	28	62,2	38	62,3	28	63,6	18	60,0
Klein	47,60%	10	35,7	17	37,8	23	37,7	16	36,4	12	40,0

M6 χ^2_2 (1, N = 799) = 1,59; p = 0,21; ϕ = 0,04
E6 χ^2_2 (1, N = 1323) = 1,74; p = 0,19; ϕ = 0,04
M7 χ^2_2 (1, N = 1812) = 2,40; p = 0,12; ϕ = 0,04
E7 χ^2_2 (1, N = 1346) = 2,23; p = 0,14; ϕ = 0,04
M8 χ^2 (1, N = 941) = 0,70; p = 0,40; ϕ = 0,03

Voor de steekproeven zijn de verschillen niet significant (De steekproefverdelingen bij alle normeringsmomenten zijn vrijwel identiek en overeenkomstig de populatieverdeling. Aangenomen wordt daarom dat de scholen in de steekproeven nagenoeg representatief zijn naar schoolgrootte.

Regio

Bij de definitie van de variabele *regio* is uitgegaan van de CBS-indeling naar landsdeel. Dit betekent dat er vier regio's onderscheiden zijn. Regio *noord* omvat de provincies Groningen, Friesland en Drenthe; regio *oost* de provincies Overijssel, Gelderland en Flevoland; regio *west* de provincies Utrecht, Noord-Holland, Zuid-Holland en Zeeland en regio *zuid* de provincies Noord-Brabant en Limburg.

De populatieverdeling van de scholen en de scholen in de steekproeven voor M6, E6, M7, E7 en M8 naar regio staat in tabel 4.7.

Tabel 4.7 Scholen uit steekproeven M6, E6, M7, E7 en M8 naar regio

Regio	**Landelijk**	**M6**		**E6**		**M7**		**E7**		**M8**	
	%	**Aantal**	**%**	**Aantal**	**%**	**Aantal**	**%**	**Aantal**	**%**	**Aantal**	**%**
Noord	15,2	3	9,7	10	20,0	7	12,5	10	20,0	8	22,9
Oost	24,6	4	12,9	7	14,0	8	14,3	7	14,0	6	17,1
West	41,7	16	51,6	19	38,0	23	41,1	19	38,0	12	34,3
Zuid	18,5	8	25,8	14	28,0	18	32,1	14	28,0	9	25,7

M6 χ^2_2 (3, N = 799) = 4,85; p = 0,18; ϕ = 0,08
E6 χ^2_2 (3, N = 1323) = 4,15; p = 0,25; ϕ = 0,06
M7 χ^2_2 (3, N = 1812) = 5,26; p = 0,15; ϕ = 0,05
E7 χ^2_2 (3, N = 1346) = 4,15; p = 0,25; ϕ = 0,06
M8 χ^2 (3, N = 941) = 2,69; p = 0,44; ϕ = 0,05

In de steekproef zijn de scholen voor alle normeringmomenten in regio oost ondervertegenwoordigd en in regio zuid oververtegenwoordigd. Voor bijna alle normeringsmomenten (behalve M7) is regio noord ook oververtegenwoordigd. Met betrekking tot de significante chi-kwadraat-waarden is het raadzaam om de effectgrootte ϕ als uitgangspunt te nemen.
We zien dat de effectgroottes onder de .10 liggen en daarmee zeer klein zijn (cf. Cohen, 1988). Ze zijn met 0,08 nog het hoogst bij het afnamemoment M6. Deze effectgroottes zijn te interpreteren als klein tot middelgroot. Aangenomen wordt daarom dat voor alle afnamemomenten de steekproeven van scholen redelijk representatief zijn naar regio.

Verstedelijking
De populatieverdeling van de scholen en de verdelingen van de scholen in de steekproeven naar verstedelijking (urbanisatiegraad) staan in tabel 4.8. Het betreft hier een indeling in vijf categorieën die bij het CBS gebruikelijk is.

Tabel 4.8 Scholen uit steekproeven M6, E6, M7, E7 en M8 naar verstedelijking

Regio	Landelijk	M6		E6		M7		E7		M8	
	%	Aantal	%	Aantal	%	Aantal	%	Aantal	%	Aantal	%
niet	19,7	2	6,5	7	14,0	8	14,3	7	14,0	5	14,3
weinig	19,4	9	29,0	14	28,0	15	26,8	14	28,0	11	31,4
matig	19,7	10	32,3	13	26,0	18	32,1	13	26,0	12	34,3
sterk	22,4	5	16,1	10	20,0	8	14,3	10	20,0	5	14,3
zeer sterk	12,2	5	16,1	6	12,0	7	12,5	6	12,0	2	5,7

M6 χ^2_2 (4, N = 799) = 5,91; p = 0,21; ϕ = 0,04
E6 χ^2_2 (4, N = 1323) = 1,88; p = 0,76; ϕ = 0,04
M7 χ^2_2 (4, N = 1812) = 6,71; p = 0,15; ϕ = 0,06
E7 χ^2_2 (4, N = 1346) = 1,88; p = 0,76; ϕ = 0,04
M8 χ^2 (4, N = 941) = 6,81; p = 0,15; ϕ = 0,09

Voor de afnamemomenten M6, M7 en M8 geldt dat de steekproefverdelingen statistisch gezien significant afwijken van de verdeling van scholen in de populatie. Voor alle afnamemomenten zien we een oververtegenwoordiging bij weinig en matig verstedelijkte gebieden en een ondervertegenwoordiging bij niet verstedelijkte gebieden. Met betrekking tot de significante chi-kwadraat-waarden hebben we ook hier naar de effectgrootte ϕ gekeken.
We zien dat de effectgroottes onder de .10 liggen en daarmee zeer klein zijn (cf. Cohen, 1988). Ze zijn met 0,09 nog het hoogst bij het afnamemoment M8. Deze effectgroottes zijn te interpreteren als klein tot middelgroot. Aangenomen wordt daarom dat voor alle groepen de scholen in de steekproeven redelijk representatief zijn naar verstedelijking.

Sekse
Bij de variabele *sekse* is een tweedeling naar jongens en meisjes gehanteerd. De verdeling van alle leerlingen en de leerlingen in de verschillende steekproeven naar sekse staat in tabel 4.9. In de steekproeven voor de verschillende normeringsmomenten zijn de leerlingen nagenoeg verdeeld zoals de landelijke verdeling van leerlingen. Hier is dan ook geen significant verschil geconstateerd.

Tabel 4.9 Leerlingen uit steekproeven M6, E6, M7, E7 en M8 naar sekse

	Landelijk	M6		E6		M7		E7		M8	
Geslacht	**%**	**Aantal**	**%**	**Aantal**	**%**	**Aantal**	**%**	**Aantal**	**%**	**Aantal**	**%**
jongen	49,8	363	49,3	602	49,6	801	48,5	561	49,4	377	47,3
meisje	50,2	374	50,7	611	50,4	851	51,5	575	50,6	420	52,7

M6 χ^2_2 (1, N = 799) = 0,007; p = 0,79; ϕ = 0,01
E6 χ^2_2 (1, N = 1323) = 0,007; p = 0,93; ϕ = 0,002
M7 χ^2_2 (1, N = 1812) = 1,05; p = 0,30; ϕ = 0,02
E7 χ^2_2 (1, N = 1346) = 0,06; p = 0,80; ϕ = 0,006
M8 χ^2 (1, N = 941) = 1,91; p = 0,17; ϕ = 0,04

De deelnemende scholen zijn dus voor alle achtergrondvariabelen redelijk representatief te noemen voor de populatie van scholen. Statistische weging is om die reden dan ook niet nodig. Hetzelfde geldt op leerlingniveau voor de verdeling naar sekse.

4.3 Kalibratie

4.3.1 De kalibratieprocedure

Met kalibratie wordt bedoeld dat we kengetallen zoeken bij de items die de antwoorden van de leerlingen goed representeren. Hoe de kengetallen gezocht worden ligt deels vast door het gekozen model (zie paragraaf 2.4.2.2). Hoe succesvol deze operatie is, kan statistisch getoetst worden. Eenvoudig gezegd, schatten we in OPLM met de CML-methode de itemparameters en controleren we of deze de data goed voorspellen. Voor een exacte beschrijving van de statistische toetsen die in OPLM gebruikt worden, hun eigenschappen en feitelijke implementatie in OPLM, verwijzen we naar Verhelst (1993). Hier beperken we ons tot een korte beschrijving van de principes van de statistische toetsen die gebruikt zijn in de kalibratieprocedure.

De statistische toetsen in OPLM hebben goede statistische en asymptotische eigenschappen, daar OPLM behoort tot de exponentiële familie, met de gewogen somscore:

$$s = \sum_{i=1}^{k} a_i x_i \quad , \tag{4.1}$$

als een ‘afdoende statistiek’ voor de vaardigheid . Dit betekent dat alle informatie in de data met betrekking tot de vaardigheid in deze statistiek aanwezig is. Hiervan wordt gebruikgemaakt bij de statistische toetsen in OPLM. Het basisprincipe van de statistische toetsen in OPLM is dat op grond van de afdoende statistiek s de personen in de data kunnen worden gegroepeerd. En binnen deze groepen kan de verwachte proportie goede antwoorden op een item onder het model, $p(\ |s)$, vergeleken worden met de feitelijk geobserveerde proportie goede antwoorden, $prop(\ |s)$. In het polytome geval worden de items gedichotomiseerd, de proportie goede antwoorden verwijst dan naar de hoge itemscore (zie Verhelst, 1993, hoofdstuk 7). Via de basisvergelijking van OPLM kunnen we eenvoudig de conditionele kans op het goed beantwoorden van de items afleiden en daarmee kunnen we $p(\ |s)$ evalueren, $prop(\ |s)$ volgt uit de data. Discrepanties tussen $p(\ |s)$ en $prop(\ |s)$ duiden op schendingen van het model. Deze discrepanties vormen de basis voor de diverse statistische toetsen in OPLM. De toetsingsgrootheid voor de veronderstelde discriminatie-indices is gegeven door

$$M = f_{s\in H}(p(+|s) - prop(+|s)) + f_{s\in L}(prop(+|s) - p(+|s)). \tag{4.2}$$

Deze zogeheten M-toetsen verdelen de scoregroepen in een laag deel (L) en een hoog deel (H) en f is een monotone functie. M-toetsen hebben een duidelijke interpretatie: is M significant positief dan is de veronderstelde steilheid van de ICC (item karakteristieke curve) overschat in het model, is M daarentegen erg laag dan is de index te klein. Verhelst laat zien voor welke functie, f , $M \quad N(0,1)$. In OPLM zijn drie verschillende M-toetsen geïmplementeerd die verschillen in de definitie van de hoge en lage scoregroepen. Naast deze M-toetsen is er een algemene itemtoets die de volgende vorm heeft

$$S = f(p(+ \mid s) - prop(+ \mid s))$$

Deze zogenoemde S-toets heeft een ^{2} verdeling onder het model. Als globale modeltoets is de R1c-toets (Glas, 1988) geschikt. Ook de distributie van alle afzonderlijke S-toetsen komt hiervoor in aanmerking. Als we deze S-toetsen opvatten als onafhankelijk, wat ze strikt genomen niet zijn, dan zouden de overschrijdingskansen uniform verdeeld moeten zijn op het (0,1) interval.
Kortom, als we afzien van de formeel-statistische achtergrond van de gehanteerde toetsen, kan de kalibratieprocedure als volgt worden samengevat:

1. Met behulp van het programma OPCAT stellen we de discriminatie-indices in OPLM in en hercoderen we indien noodzakelijk de antwoordcategorieën in de data.
2. Vervolgens schatten we de itemparameters met behulp van de CML-methode.
3. Met behulp van de M-toetsen controleren we of de discriminatie-indices goed zijn ingesteld.
4. Een volgende controle betreft de overschrijdingskansen van de S-toetsen en een grafische modelcontrole door middel van het programma Wopplot (grafische inspectie van de ICC's).
5. Vervolgens vindt een globale modelcontrole plaats in de vorm van een R1c-toets en de verdeling van de overschrijdingskansen van de S-toetsen.

De stappen 1 tot en met 5 worden een aantal malen doorlopen tot het resultaat bevredigend is. Afhankelijk van de uitkomsten kunnen items worden verwijderd. De opgaven vormen na de kalibratie een gekalibreerde opgavenbank, waarbij de opgaven per onderscheiden vaardigheidsdimensie een beroep doen op hetzelfde complex aan vaardigheden of 'latente trek'.

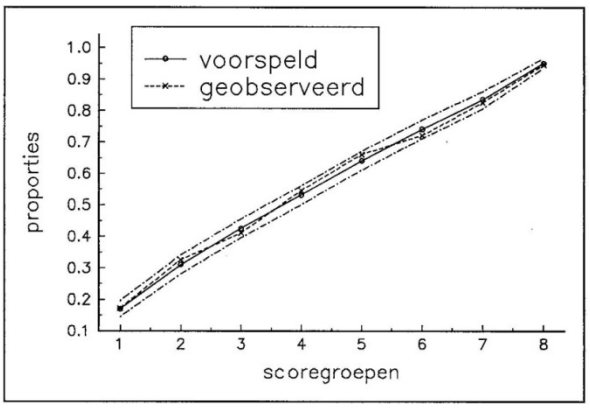
OPCAT voert een aantal statistische toetsen uit op grond waarvan bepaald kan worden of het model een adequate beschrijving geeft van de data. Belangrijk zijn de itemgeoriënteerde *S*-toets en de overall R_{1c}-toets. De *S*-toets is asymptotisch ^{2} verdeeld en is gebaseerd op de verschillen tussen de geobserveerde en verwachte proporties antwoorden in homogene scoregroepen. Een uniforme verdeling van *p*-waarden voor de *S*-toetsen in het interval [0,1] pleit voor passing van het model. De R_{1c}-toets heeft dezelfde onderliggende rationale als de *S*-toets en wordt over het algemeen acceptabel bevonden indien zijn waarde niet groter is dan anderhalf tot hooguit twee keer het aantal vrijheidsgraden.

4.3.2 Resultaten van de kalibratieprocedure: modelfit

Het is niet eenvoudig om de kwaliteit van de kalibratie aan te tonen. De belangrijkste statistische instrumenten om de passing van een opgave in het IRT-model te bewerkstellingen en uiteindelijk te documenteren betreffen de hierboven al besproken S-toetsen. Het lastige daarvan is, dat de toetsing voor een groot deel visueel gebeurt. Dit kunnen we illustreren aan de hand van figuur 4.1 (zie Staphorsius, 1994, blz. 239). Ten behoeve van deze toetsing wordt de totale groep van leerlingen die een verzameling opgaven gemaakt heeft, ingedeeld in een aantal (meestal acht) scoregroepen. Elke groep bestaat uit leerlingen met een ongeveer even hoge score. De geobserveerde proporties juiste antwoorden van deze groepen (telkens gesymboliseerd door een x) zijn door de middelste stippellijn verbonden. De volle lijn daarentegen verbindt de proporties die op grond van de parameterschattingen voorspeld kunnen worden. De twee buitenste lijnen geven het 95%-betrouwbaarheidsinterval aan. De breedte van dit interval is in belangrijke mate afhankelijk van het aantal leerlingen dat de opgave heeft beantwoord. Uit de figuur blijkt heel duidelijk dat de

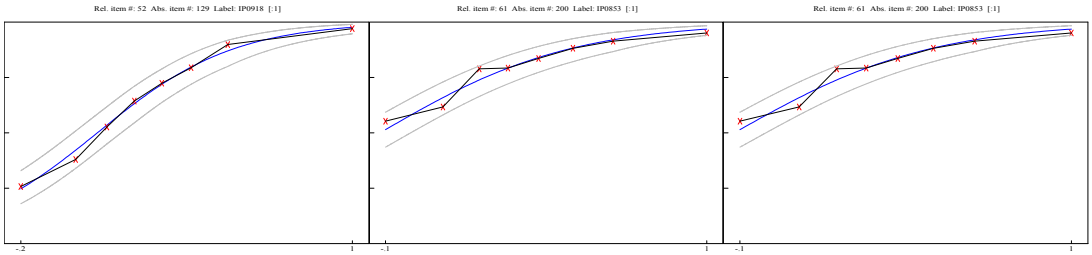
geobserveerde proporties, zoals bedoeld, binnen het 95%- betrouwbaarheidsinterval van de (geschatte) voorspelde proporties liggen, en dit komt in grote lijnen overeen met een niet-significante S-toetsingsgrootheid (Verhelst et al., 1994).

Figuur 4.3 Grafische voorstelling van een Si –toets



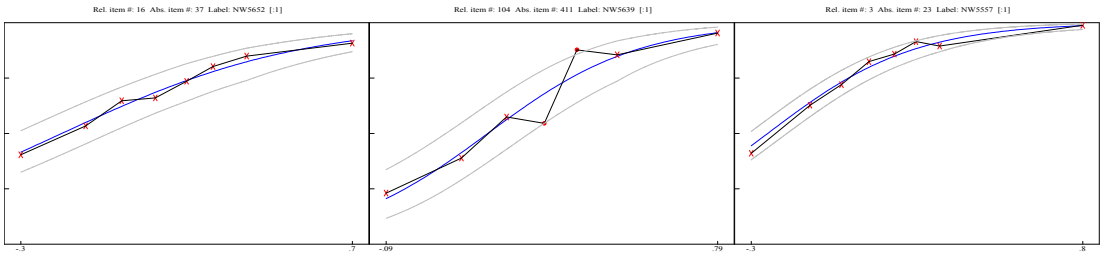
Het is ondoenlijk om voor alle opgaven dergelijke grafische voorstellingen in deze verantwoording op te nemen. Daarom beperken we ons steeds per toetsversie tot het item met de slechtste en de beste S-passing, aangevuld met een qua S-toetsingsresultaat gemiddelde (dat wil zeggen, meest representatieve) passing. De voorbeelden in figuur 4.4 illustreren dat voor beide toetsversies zelfs bij de slechtst passende opgave sprake is van een zeer aanvaardbaar beeld. Er wordt in deze gevallen voor een deel (van de onderscheiden scoregroepen) niet beantwoord aan de eis dat de geobserveerde proportie binnen het 95%-betrouwbaarheidsinterval van de geschatte proporties ligt. Dit beeld doet zich slechts bij enkele opgaven voor die dan ook een uitzondering vormen. De overige opgaven voldoen voor alle scoregroepen wel aan die eis. De afbeeldingen voor de representatieve en best passende opgaven illustreren dit. Dit leidt tot de conclusie dat bij vrijwel alle opgaven in de Taalverzorging-toetsen een grafische voorstelling van de S-toetsing hoort die in grote lijnen met figuur 4.3 overeenkomt; andere opgaven zijn bij de kalibratie niet in de itembank opgenomen. Dit is, zeker gezien de relatief grote aantallen observaties die in het geding zijn, een zeer sterke aanwijzing dat het meetinstrument en het meetmodel dat ontwikkeld is, respectievelijk gebruikt is, adequaat zijn om het gedrag van de leerlingen te verklaren. Bovendien blijkt, en dat is vanuit theoretisch oogpunt nog belangrijker, dat gemeten verschillen in gedrag tussen de leerlingen te verklaren zijn door één unidimensioneel concept.

Figuur 4.4 Voorbeelden van S-toetsen voor de toets Taalverzorging voor alle onderdelen, alle toetsen met de best passende, de slechtst passende en een qua passing representatieve opgave



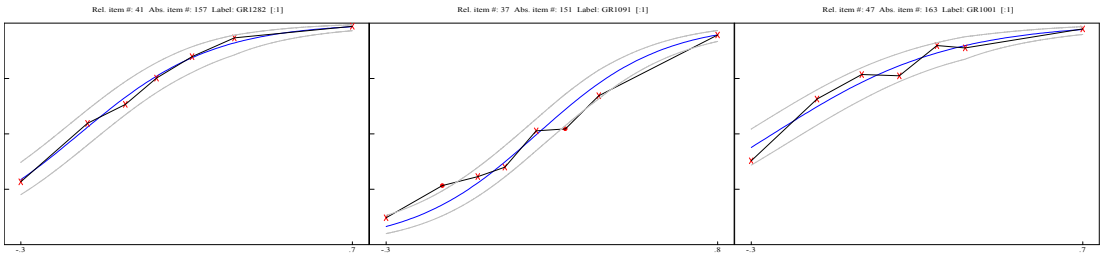
Interpunctie

best passend | slechtst passend | representatieve passing



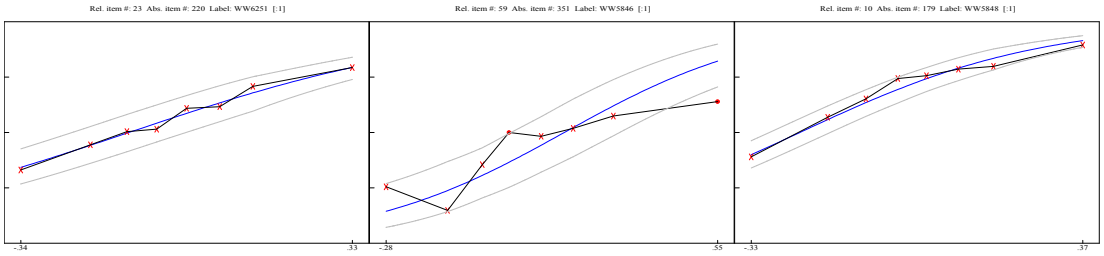
Spelling niet-werkwoorden

best passend | slechtst passend | representatieve passing



Grammatica

best passend | slechtst passend | representatieve passing



Spelling werkwoorden

best passend | slechtst passend | representatieve passing

In feite kan men bij de kalibratie beter varen op deze grafische weergaven dan op toetsingsresultaten in termen van exacte getallen en de significantie daarvan. Niettemin zijn er bij de kalibratie S-toetsen uitgevoerd die een indicatie geven van de kwaliteit van de kalibratie. Daarbij zijn we vooral geïnteresseerd in de

distributie van de overschrijdingskansen van deze verzameling toetsingsresultaten. Zoals eerder aangegeven zouden de overschrijdingskansen gelijkmatig verdeeld moeten zijn binnen het (0,1) interval, uiteraard met zo weinig mogelijk significante resultaten. Tabel 4.10 waarin het (0,1) interval is opgedeeld in tien gelijke stukken, geeft een beeld van de uitkomsten bij een kalibratie van alle 220 opgaven van de deelgebieden Taalverzorging alle items. Daarnaast is aangegeven in hoeveel gevallen de overschrijdingskans kleiner was dan .01, respectievelijk .05. Het is duidelijk dat voor de toets de verdeling redelijk gelijkmatig is over het gehele interval van overschrijdingskansen. Deze resultaten geven een bevestiging van het eerder geschetste beeld, dat met uitzondering van enkele opgaven, sprake is van niet-significante S-toetsen. Zij vormen een kwantitatieve ondersteuning van de conclusie dat de opgaven per deelgebied een unidimensioneel construct representeren.

Tabel 4.10 Verdeling van overschrijdingskansen bij S-toetsen voor alle items per deelgebied

	0.------/-------/-------.1--------.2-------.3-------.4-------.5------.6------.7------.8------.9------1
IP	3 9 5 12 10 9 7 12 6 7 16 12
	0.-------/-------/-------.1--------.2-------.3-------.4-------.5------.6------.7-------.8-----.9------1
NWW	1 4 4 8 8 8 6 10 16 15 11 16
	0.-------/------/--------.1--------.2-------.3-------.4-------.5------.6------.7-------.8-----.9------1
GR	5 8 6 12 11 7 5 13 5 12 8 11
	0.-------/------/--------.1--------.2-------.3--------.4------.5------.6------.7-------.8-----.9------1
WW	1 6 2 7 10 5 9 8 14 6 8 8

In tabel 4.11 zijn de R1c-waarden weergegeven voor de deelgebieden van Taalverzorging waarvoor in tabel 4.10 de resultaten van de S-toetsen (op itemniveau) zijn weergegeven. R1c is een statistiek die zicht geeft op de modelpassing van de toets als geheel. Voor een acceptabele modelfit geldt als vuistregel dat R1c bij voorkeur niet significant (bij α=0,01) zou moeten zijn en niet groter dan ongeveer anderhalf maal het aantal vrijheidsgraden (df).
De modelpassing van de toets voldoet aan een van deze voorwaarden. Voor alle deelgebieden geldt dat de R1c minder dan anderhalf maal het aantal vrijheidsgraden bedraagt. De toetsingsgrootheid is significant. Aan dit laatste moet bij steekproeven met een dergelijke omvang (ca. 6000 leerlingen in totaal) niet te veel waarde worden gehecht.

Tabel 4.11 R1c-waarden voor alle items taalverzorging per deelgebied

Deelgebied	R1c	df	p
IP	2883,2	1956	<0,005
NWW	2740,6	2095	<0,005
GR	2904,0	1663	<0,005
WW	1763,6	1324	<0,005

Ten slotte bespreken we nog een methode om de modelpassing te verantwoorden die wordt besproken in het COTAN Beoordelingssysteem (Evers et al., 2010). Het betreft hier een poging om de nauwkeurigheid van de itemparameterschattingen te beoordelen op basis van een constante (in het COTAN Beoordelingssysteem met ‘c’ aangeduid) die weergeeft hoe de relatie is tussen de standaardfout van de moeilijkheidsparameter van een item en de standaarddeviatie van de vaardigheidsverdeling van de kalibratiepopulatie.

Het beoordelingssysteem geeft ook richtlijnen voor het beoordelen van de grootte van deze 'c'. Deze dient te worden beoordeeld als goed als de waarde lager is dan of gelijk aan 0,20. Waarden tussen 0,30 en 0,40 kunnen nog als voldoende worden aangemerkt. De standaardfouten van de moeilijkheidsparameters worden dus gedeeld door de standaarddeviatie van de populatie waarin ze zijn afgenomen. Bij drie opgaven spelling van het totale aantal items in de uitgave taalverzorging is de waarde groter dan 0,20. Het gaat om twee items spelling werkwoorden en een item spelling niet-werkwoorden. Op het totale aantal items die zijn opgenomen in de drie toetsen Taalverzorging (220) is dit niet betekenisvol. Bovendien liggen deze waarden onder de 0,30 (dat wil zeggen, volgens de COTAN-normen ergens tussen 'goed' en 'voldoende'). Range en gemiddelde van de waarden van 'c' duiden op een hoge nauwkeurigheid van de itemparameterschattingen.

Tabel 4.12 Nauwkeurigheid van de itemparameterschattingen (constante 'c')

Deelgebied	Constante 'c'	
	Range	Gemiddelde
IP	0,044 – 0,137	0,075
NW	0,038 – 0,218	0,085
GR	0,012 – 0,182	0,083
WW	0,06 – 0,255	0,123

Op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de toetsen Taalverzorging van het Cito Volgsysteem primair en speciaal onderwijs (LVS) voor alle afnamemomenten de kalibratie geslaagd is. Hiermee is het laatste woord nog niet gezegd over de validiteit, maar het kalibratieonderzoek brengt in ieder geval een essentieel aspect van het validiteitsvraagstuk naar voren: de rechtvaardiging van wat in de meeste toetstoepassingen gebruikelijk is, namelijk het reduceren van alles wat de leerling heeft geantwoord tot een enkele toetsscore (of afgeleid daarvan, een enkele schatting van zijn onderliggende vaardigheid). De kalibratie-analyse, als puur formeel proces, kan geen uitspraken doen over de inhouds-validiteit of over de constructvaliditeit als antwoord op de vraag: hoe kan worden aangetoond dat het concept dat de items in de bank meten dekkend is voor en samenvalt met het construct dat we in de toetsen Taalverzorging proberen te meten (zoals dat in het didactisch en het wetenschappelijk forum wordt bedoeld)? In hoofdstuk 6 over validiteit zal worden nagegaan of de gemeten concepten inderdaad overeenkomen met het begrip zoals bedoeld. De vraag is dan in het geval van het onderdeel Taalverzorging: kunnen de unidimensionele concepten onder de opgaven in de opgavenbank van de vier deelgebieden van Taalverzorging inderdaad worden opgevat als de deelvaardigheden van 'taalverzorging'? Een geslaagde kalibratie op een unidimensioneel construct beschouwen we als een noodzakelijke voorwaarde voor deze begripsvaliditeit.

4.4 Normeringsresultaten

De volgsysteemtoetsen Taalverzorging voor groep 6 tot en met 8 kennen per leerjaar één toets, genormeerd voor de twee afnamemomenten in het jaar, het zogeheten M-moment (halverwege het schooljaar) en het E-moment (aan het einde van het schooljaar). Aan het normeringsonderzoek M6 namen in totaal ruim 6200 leerlingen deel. Zoals te zien is in tabel 4.1 namen per afnamemoment ruim voldoende leerlingen deel. Op M6 799 leerlingen, op M7 1812, op M8 941 en op E6 en E7 respectievelijk 1323 en1346 leerlingen.

In paragraaf 2.4.2 noemden we de belangrijke implicaties voor een gekalibreerde opgavenverzameling. Het slagen van de kalibratie betekent dat we met een selectie van opgaven uit de opgavenbank de vaardigheid (of beter gezegd, de vier onderscheiden deelvaardigheden) bij een leerling kunnen meten. Hoe nauwkeurig we dat doen, staat beschreven in paragraaf 5.2.
We kunnen nu een schatting maken van de betreffende vaardigheidsverdelingen in een welomschreven populatie, omdat we de toetsopgaven voorgelegd hebben aan aselecte steekproeven van leerlingen uit

populaties die in overeenstemming zijn met de aangeduide afnameperiodes M6, M7, M8 en E6 en E7.
We schatten het gemiddelde en de standaardafwijking in de veronderstelling dat de vaardigheid normaal verdeeld is. Met behulp van deze gegevens kunnen vervolgens ook schattingen gemaakt worden van de percentielen in de populatie, die van belang zijn voor de indeling van leerlingen in de niveaugroepen die zijn beschreven in paragraaf 2.3.

Deze percentielen zijn voor beide afnamemomenten weergegeven in tabel 4.13.
Een overzicht van de geschatte gemiddelden en de standaardafwijkingen van de vaardigheid op de verschillende normeringsmomenten voor de onderzochte populaties is eveneens te vinden in tabel 4.13.
Uit deze tabel blijkt dat de gemiddelde vaardigheid in de verschillende taalverzorgingsaspecten in de periode tussen de afnamemomenten toeneemt, terwijl de spreiding in de scores nagenoeg gelijk is.

Tabel 4.13 Overzicht van de vaardigheidsverdelingen per normeringsmoment per deelgebied

Afname	**IP**											
moment	**N**	**gem.**	**st.dev.**	**P10**	**P20**	**P25**	**P40**	**P50**	**P60**	**P75**	**P80**	**P90**
M6	799	113,0	16,1	92,4	99,4	102,1	108,9	113,0	117,0	123,8	126,5	133,5
E6	1323	120,1	19,1	95,7	104,1	107,3	115,3	120,1	124,9	132,9	136,1	144,5
M7	1812	120,5	16,9	98,8	106,2	109,1	116,2	120,5	124,7	131,8	134,7	142,1
E7	1346	123,2	17,7	100,5	108,3	111,2	118,7	123,2	127,6	135,1	138,0	145,8
M8	941	128,6	19,4	103,8	112,3	115,5	123,7	128,6	133,5	141,7	144,9	153,4

Afname	**NW**											
moment	**N**	**gem.**	**st.dev.**	**P10**	**P20**	**P25**	**P40**	**P50**	**P60**	**P75**	**P80**	**P90**
M6	799	107,4	12,7	91,1	96,7	98,8	104,1	107,4	110,6	115,9	118,0	123,6
E6	1323	112,7	13,8	95,0	101,1	103,4	109,2	112,7	116,2	122,0	124,3	130,4
M7	1812	115,9	13,4	98,7	104,6	106,9	112,5	115,9	119,3	124,9	127,2	133,1
E7	1346	119,3	14,7	100,5	106,9	109,4	115,5	119,3	123,0	129,1	131,6	136,5
M8	941	121,1	13,6	103,7	109,6	111,9	117,6	121,1	124,5	130,2	132,5	138,4

Afname	**GR**											
moment	**N**	**gem.**	**st.dev.**	**P10**	**P20**	**P25**	**P40**	**P50**	**P60**	**P75**	**P80**	**P90**
M6	799	100,1	12,2	84,5	89,8	91,9	97,0	100,1	103,2	108,3	110,4	115,7
E6	1323	108,3	15,2	88,8	95,5	98,0	104,4	108,3	112,1	118,5	121,0	127,7
M7	1812	110,1	14,1	92,0	98,2	100,6	106,5	110,1	113,7	119,6	122,0	128,2
E7	1346	115,7	16,7	94,2	101,6	104,4	111,4	115,7	119,9	125,5	128,5	134,5
M8	941	116,3	15,7	96,2	103,1	105,7	112,3	116,3	120,3	127,5	129,5	136,4

Afname	**WW**											
moment	**N**	**gem.**	**st.dev.**	**P10**	**P20**	**P25**	**P40**	**P50**	**P60**	**P75**	**P80**	**P90**
M6	799											
E6	1323											
M7	1812	99,4	8,0	89,1	92,6	94,0	97,3	99,4	101,4	104,7	106,1	109,6
E7	1346	102,9	9,1	91,2	95,2	96,7	100,5	102,9	105,2	109,0	110,5	114,5
M8	941	104,7	9,4	92,6	96,7	98,3	102,3	104,7	107,0	111,0	112,6	116,7

In figuur 4.5 zijn in een reeks histogrammen de vaardigheidsscores voor de normeringssteekproeven weergegeven, additioneel zijn ook de bijbehorende normaalverdelingen ingetekend. De aanname van een normaal verdeelde vaardigheidsverdeling wordt niet volledig ondersteund door de data. Om te bepalen of de vaardigheidsverdeling normaal verdeeld, is een statistische toets ontwikkeld, de zogeheten R0 toets (Verhelst, Glas & Verstralen, 1995). Voor de onderhavige gevallen staan de waarden van deze toetsen in tabel 4.14.

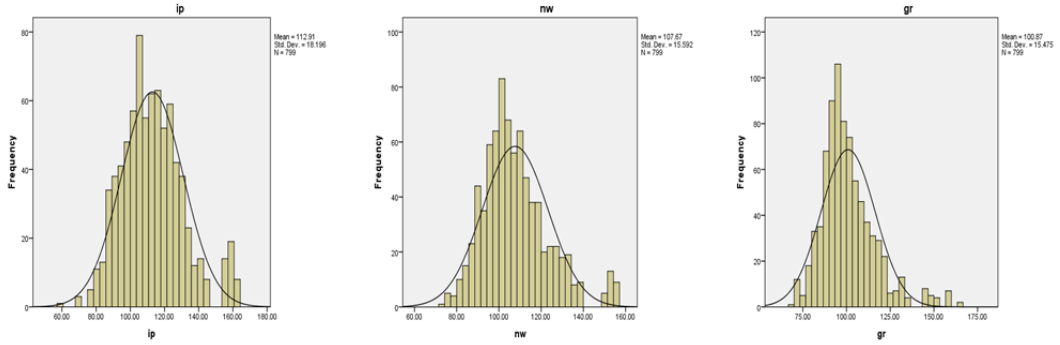
Tabel 4.14 Toets op normaliteit van de vaardigheidsverdelingen per deelgebied Taalverzorging

Deelgebied	**Chi^2**	**df**	**P**	**phi**
IP	2530,5	1730	0	0,64
NWW	2393	1697	0	0,62
GR	2415	1984	0	0,62
WW	1918,6	1228	0	0,56

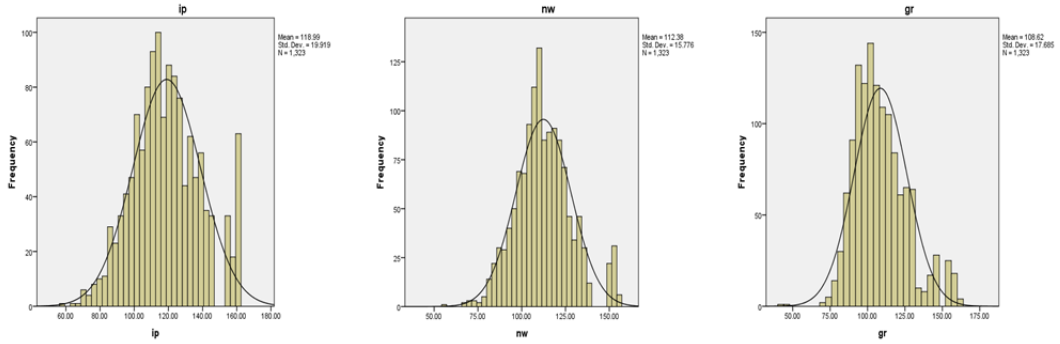
De R0-statistieken geven aan in hoeverre de vaardigheidsverdelingen afwijken van een normale verdeling. De R0-toets is uitgevoerd per deelgebied over alle meetmomenten heen. Zoals te zien is in tabel 4.14 zijn de R0-toetsen (die de vorm hebben van een Chi^2-toets) significant. De waarde van phi laat zien dat de afwijkingen substantieel zijn. De toetsingen zijn echter uitgevoerd over alle meetmomenten heen, terwijl de histogrammen (zie figuur 4.5) laten zien dat de afwijkingen van normaliteit beperkt blijven tot enkele deelgebieden en meetmomenten. Dit impliceert dat de betreffende vaardigheden op enkele normerings-momenten niet zoals verondersteld was, als normaal verdeeld kunnen worden opgevat. In de figuren per meetmoment is te zien dat sommige deelgebieden afwijkingen vertonen in de vorm van histogrambalken die niet goed overeenstemmen met de normale verdelingscurve. Vooral zien we dit heel duidelijk bij E6, M7, E7, M8 Interpunctie en M7 Grammatica, met name daar waar de balken boven de curve uitsteken. We zien de afwijkingen alleen aan de rechterkant (boven percentiel 80), verder lijkt de normaalverdeling goed te passen. Aangezien de afwijkingen beperkt blijven tot de hoogste scores, en binnen het hoogste quintiel worden gecompenseerd in oppervlakte onder de curve, is de conclusie gerechtvaardigd, dat deze geen effect hebben op de niveauverdeling. Bij de niveaubepalingen van de toetsen Taalverzorging wordt door de leerkrachten alleen gebruikgemaakt van de vaardigheidsscores en niveauverdelingen I t/m V of A t/m E en niet van percentielscores. Daarom is besloten de normering op basis van de aanname van normaliteit ook voor de genoemde vaardigheidsscores en meetmomenten te handhaven.

Figuur 4.5 Histogram van de vaardigheidsverdelingen in de normeringssteekproeven M6, E6, M7, E7 en M8 met de beoogde normaalverdelingen per afnamemoment

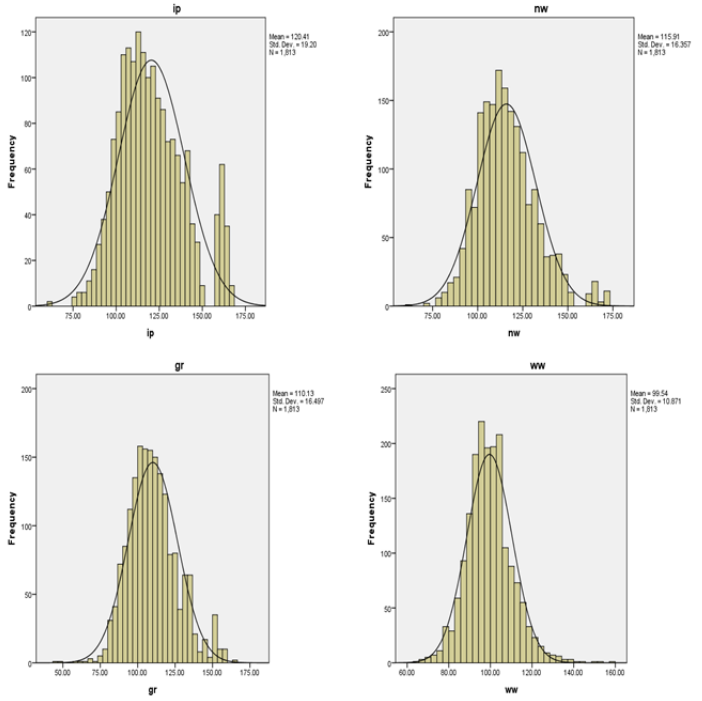
M6



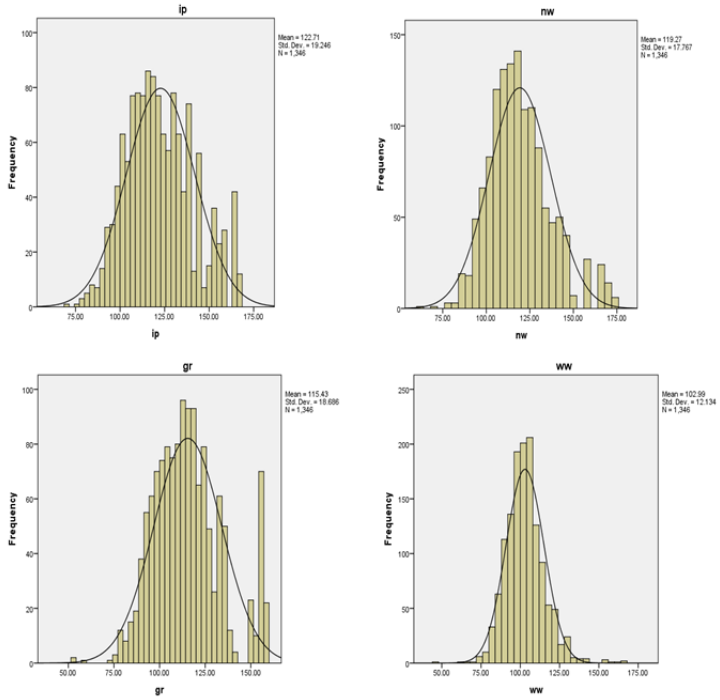
E6



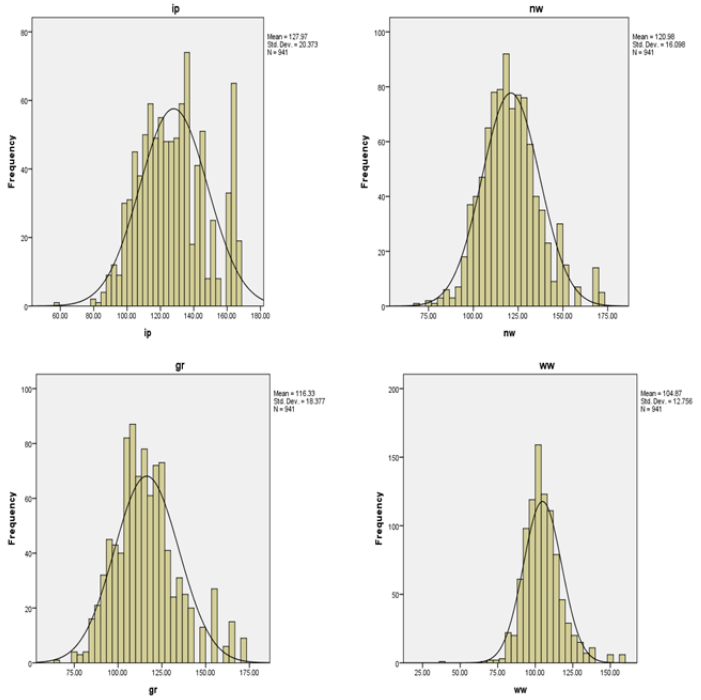
M7



E7



M8



4.5 Referentieonderzoek: het bepalen van de cesuur

In het Referentiekader Taal en Rekenen (Expertgroep Doorlopende Leerlijnen Taal en rekenen, 2009a) is het domein Taalverzorging gespecificeerd als overkoepelende term voor de onderliggende constructen: spelling werkwoorden, niet-werkwoorden, interpunctie en begrippenlijst. Het referentiekader was voor CvTE de leidraad bij het ontwikkelen van de Ankersets. De cesuur op de Ankersets is daarmee vastgesteld op het domein als geheel, en niet op de onderliggende constructen. Ook voor het ontwikkelen van de LVS toetsen Taalverzorging was het Referentiekader de basis. Gezien het doel van de toetsen, namelijk leerlingen volgen

op specifieke onderdelen, ligt het voor de hand voor de rapportage uit te gaan van een schaling van de losse onderdelen (i.e., losse itembanken). Dit omdat de uitkomsten op deze manier het meeste recht doen aan de dimensionele structuur in de data. Aangezien de referentiecesuur niet vastgesteld kan worden op een afzonderlijk onderdeel is er voor het overzetten van de cesuur gebruikgemaakt van een itembank waarin alle vier de onderdelen opgenomen waren.

Om de standaard van de openbare Ankerset Taalverzorging over te brengen naar de toetsen Taalverzorging voor de groepen 7 en 8 hebben we een overbrengingsstrategie toegepast die uitgaat van een vaardigheidsschaal of latente schaal en waarvoor IRT wordt toegepast. De cesuur wordt omgezet in een cesuur op de latente schaal. De score op de te ankeren toets en de ankersetopgaven worden beide afgebeeld op dezelfde latente schaal. De cesuur op de te ankeren toets kan worden bepaald als de score die een verwachte latente schaalscore heeft die het dichtst bij de latente cesuur ligt. Aandachtspunt bij de analyse is dat voor het bepalen van de latente schaal alleen de data op de ankerset gebruikt worden die verzameld zijn bij de doelpopulatie. Voor elke doelpopulatie wordt dus een eigen latente schaal gemaakt.

Voor het bepalen van de cesuur hebben we gebruikgemaakt van de applicatie Bereken Grensscore-Taalverzorging die beschikbaar is gesteld door het CvTE bij de Openbare Ankersets Taalverzorging. De applicatie Bereken Grensscore-Taalverzorging is een instrument dat is ontwikkeld om geautomatiseerd de cesuur over te brengen op toetsen taalverzorging zodat alle toetsaanbieders, ondanks verschillen in toetsvorm en toetslengte, dezelfde landelijke prestatiestandaard hanteren bij het vaststellen van de behaalde referentieniveaus.

Er zijn twee sets door CvTE beschikbaar gesteld: een set met grammatica-opgaven en een set zonder grammatica-opgaven. Het onderdeel grammatica vormt geen verplicht te toetsen onderdeel bij een eindtoets PO. In de algemene toetswijzer taal en rekenen voor eindtoetsen PO (2014) is vastgelegd dat het subdomein grammatica een keuzeonderdeel is. Toetsontwikkelaars die grammatica niet bevragen, moeten de referentiecesuur taalverzorging kunnen overbrengen op hun eigen toets. Maar ook toetsaanbieders die grammatica wel opnemen in hun toets, moeten de referentiecesuur kunnen overbrengen. Daarom zijn er twee referentiecesuren vastgesteld voor het referentieniveau taalverzorging 1F, namelijk exclusief én inclusief het onderdeel grammatica. Wij hebben grammatica wel opgenomen als onderdeel van de toetsen Taalverzorging en daarom hebben wij gebruikgemaakt van de set met grammatica. In tabel 4.15 zijn de cesuren weergeven voor de referentieniveaus Taalverzorging.

Tabel 4.15 Cesuren voor bepaling scores referentieniveaus Taalverzorging

	Cesuur		
	M7	**E7**	**M8**
1F	44	44	40
2F	69	69	66

5 Betrouwbaarheid en meetnauwkeurigheid

5.1 Methoden om de betrouwbaarheid te bepalen

In hoofdstuk 4 is beschreven hoe de kalibratie en normering is uitgevoerd en zijn de resultaten daarvan beschreven. In dit hoofdstuk gaan we nader in op de betrouwbaarheid en de meetnauwkeurigheid van de toetsen Taalverzorging voor groep 6 tot en met 8 voor beide afnamemomenten. Het is mogelijk om de betrouwbaarheid van de toets voor elk meetmoment te schatten door gebruik te maken van het feit dat alle items die zijn opgenomen in de toets OPLM-geschaald zijn. Ook andere beschrijvende gegevens, zoals de gemiddelde score en de standaardmeetfout, zijn te schatten op grond van het feit dat de toets volledig bestaat uit OPLM-gekalibreerde items. Om relevante beschrijvende gegevens bij de toets te genereren, is gebruikgemaakt van het programma OPLAT (Verhelst, Glas en Verstralen, 1995).

In OPLAT wordt een door Verhelst, Glas en Verstralen (1995, pp. 99-100) ontwikkelde coëfficiënt berekend die qua interpretatie een grote overeenkomst vertoont met betrouwbaarheidscoëfficiënten uit de klassieke testtheorie. Het begrip ware score is wat meer geëxpliciteerd, namelijk als de verwachte score op een (vaste) toets, maar dan gezien als functie van de latente variabele θ. Deze verwachte waarde wordt aangeduid met τ(θ). Als bovendien bekend is hoe θ in de populatie verdeeld is, kunnen ook het gemiddelde en de variantie van de ware scores in de populatie bepaald worden. De variantie van de ware scores in de populatie worden aangegeven met het symbool Var(τ). Tussen θ en τ(θ) bestaat een een-op-een relatie, immers de een kan uit de andere berekend worden. Het is echter niet zo dat een persoon met vaardigheid θ per se de toetsscore τ(θ) moet behalen (dat is alleen zo als de toets oneindig lang wordt). De geobserveerde score bij een eenmalige afname zal dan ook een afwijking vertonen van de verwachte score, waardoor met een eenmalige toetsafname niet meer zonder fout de waarde van θ bepaald kan worden. De variantie van de geobserveerde toetsscore wordt aangegeven met Var(t|τ(θ)), en door weer gebruik te maken van de distributie van θ in de populatie kan ook de gemiddelde variantie van de geobserveerde toetsscores berekend worden.

$$Var(t) = E[Var(t \mid \tau(\theta))] \tag{5.1}$$

Deze variantie kan opgevat worden als de (gemiddelde) meetfoutvariantie in de metriek van de geobserveerde scores (t). In analogie met de theorie over de betrouwbaarheid volgt dan

$$MAcc = \frac{Var(\tau)}{Var(\tau) + Var(t)} \tag{5.2}$$

waarin MAcc staat voor 'Accuracy of Measurement'.

5.2 Betrouwbaarheid: resultaten

De tabellen 5.1a t/m d bevatten informatie over de meeteigenschappen van de toetsen Taalverzorging. In de tabel is de maximumscore niet weergegeven, deze is gelijk aan het aantal opgaven dat deel uitmaakt van de totale toets en dat is telkens 20. De eerste kolom geeft de geschatte gemiddelde scores van de leerlingen op de toetsen aan. De tweede kolom bevat informatie over de geschatte standaardmeetfout op de ruwe score van de toets. In de tabel kunnen we ook geschatte betrouwbaarheidscoëfficiënten MAcc en GLB aflezen van de toetsen op de verschillende afnamemomenten. Naast de op basis van IRT berekende MAcc is GLB (‘Greatest Lower Bound’) een betrouwbaarheidsschatter op basis van klassieke testtheorie. In de regel komt GLB hoger uit dan de bekende (Cohens) coëfficiënt alfa (eveneens op basis van klassieke testtheorie) die geldt als een lichte onderschatter van de betrouwbaarheid. MAcc is doorgaans (ongeveer) gelijk aan

coëfficiënt alfa en kent dus ook de lichte vorm van onderschatting die kenmerkend is voor alfa. Kortom, GLB is waarschijnlijk de coëfficiënt die de werkelijke betrouwbaarheid van de toets het beste benadert.

Voor toetsen van het type waar geen zware consequenties voor leerlingen aan verbonden zijn (zoals de toetsen Taalverzorging) geeft de COTAN (Evers et al., 2010) aan dat een betrouwbaarheidscoëfficiënt lager dan 0,70 onvoldoende is, een betrouwbaarheidscoëfficiënt tussen 0,70 en 0,80 voldoende, en een betrouwbaarheidscoëfficiënt hoger dan 0,80 goed. Op grond van dit criterium is de betrouwbaarheid op alle normeringsmomenten voor het deelgebied interpunctie goed te noemen (GLB variërend van 0,81 tot 0,86) voor spelling niet-werkwoorden is de betrouwbaarheid voldoende (GLB variërend van 0,76 tot 0,79), voor grammatica goed (GLB variërend van 0,81 tot 0,86) en voor spelling werkwoorden is alleen het normeringsmoment M8 voldoende (GLB 0,70). Voor M7 en E7 is de betrouwbaarheid niet voldoende gebleken. De lage betrouwbaarheid bleek al uit de item-totaalcorrelaties die in tabel 3.7. Deze zijn qua range en gemiddelde een stuk lager, in combinatie met een vrij lage gemiddelde P-waarde (M7, gem. R_{it}: 0,34 en E7, gem. R_{it} 0,36*)*.
De verklaring hiervoor is drieledig en ligt deels in het onderwijsaanbod, deels in de referentieniveaus en deels in de lengte van de taken. Werkwoordspelling is nog niet veel aan bod gekomen in het onderwijs in groep 7. Om een zo evenwichtig mogelijke verdeling van de categorieën te bewerkstelligen volgens het categorieënoverzicht spelling werkwoorden en recht te doen aan de inhoudelijke beschrijving van de morfologische spelling in het referentiekader, zijn in de taken items opgenomen die voor veel leerlingen nog te moeilijk waren. Er moesten bijvoorbeeld ook al in groep 7 voldoende items van het referentieniveau 2F opgenomen worden. Pas tegen het eind van groep 8 beheersen de leerlingen de werkwoordspelling in grotere mate zoals we ook zien aan de betrouwbaarheid. Verder zou een groter aantal items spelling werkwoorden waarschijnlijk gezorgd hebben voor een hogere betrouwbaarheid. Om de leerlingen niet te veel belasten is er consequent besloten om in alle jaargroepen 20 items per deeltaak op te nemen.
Verder is de gemiddelde betrouwbaarheid van alle deelgebieden samen van de toetsen taalverzorging met een gemiddelde van 0,90 goed.
In de tabel 5.1a t/m d zien we alle toetsen voor de verschillende groepen op afnamemoment M7 en op afnamemoment E7. In tabel 5.2 staan de beschrijvende gegevens van alle deelgebieden samen per jaargroep.

Tabel 5.1 a Beschrijvende gegevens bij de toetsen Taalverzorging voor populatie M6, E6, M7, E7 en M8 onderdeel Interpunctie

IP	Gemiddelde	SD	SE	MAcc	test/hertest	GLB
M6	13,4	4,3	1,88	0,81	0,81	0,82
E6	14,8	3,9	1,75	0,80	0,80	0,81
M7	13,9	4,7	1,77	0,86	0,85	0,86
E7	14,6	4,3	1,74	0,83	0,83	0,84
M8	14,5	4,3	1,65	0,85	0,85	0,86

Tabel 5.1 b Beschrijvende gegevens bij de toetsen Taalverzorging voor populatie M6, E6, M7, E7 en M8 onderdeel Spelling niet-werkwoorden

NW	Gemiddelde	SD	SE	MAcc	test/hertest	GLB
M6	13,8	3,9	1,85	0,77	0,78	0,78
E6	15,0	3,7	1,75	0,78	0,78	0,79
M7	14,1	3,6	1,83	0,75	0,76	0,78
E7	14,7	3,6	1,76	0,77	0,76	0,77
M8	13,8	3,7	1,87	0,75	0,75	0,76

Tabel 5.1 c Beschrijvende gegevens bij de toetsen Taalverzorging voor populatie M6, E6, M7, E7 en M8 Onderdeel Grammatica

GR	Gemiddelde	SD	SE	MAcc	test/hertest	GLB
M6	11,7	4, 6	1,99	0,81	0,81	0,82
E6	13,9	4,6	1,80	0,85	0,85	0,86
M7	13,1	4,2	1,89	0,80	0,80	0,81
E7	14,3	4,3	1,76	0,83	0,83	0,84
M8	14,2	4,2	1,79	0,82	0,81	0,82

Tabel 5.1 d Beschrijvende gegevens bij de toetsen Taalverzorging voor populatie M6, E6, M7, E7 en M8 onderdeel Spelling werkwoorden

WW	Gemiddelde	SD	SE	MAcc	test/hertest	GLB
M6						
E6						
M7	12,0	3,2	2,03	0,60	0,60	0,62
E7	12,9	3,3	1,97	0,64	0,64	0,66
M8	11,7	3,6	2,03	0,68	0,68	0,70

Tabel 5.2 Beschrijvende gegevens bij de toetsen Taalverzorging voor populatie M6, E6, M7, E7 en M8 Taalverzorging totaal (alle onderdelen)

Totaal	Gemiddelde	SD	gamma1	gamma2	SE	MAcc
M6	38,6	9,71	-0,46	-0,28	3,43	0,88
E6	43,6	9,94	-0,86	0,28	3,15	0,90
M7	53,1	12,28	-0,60	-0,07	3,85	0,90
E7	56,7	13,06	-0,83	0,28	3,66	0,92
M8	54,2	12,58	-0,65	0,01	3,77	0,91

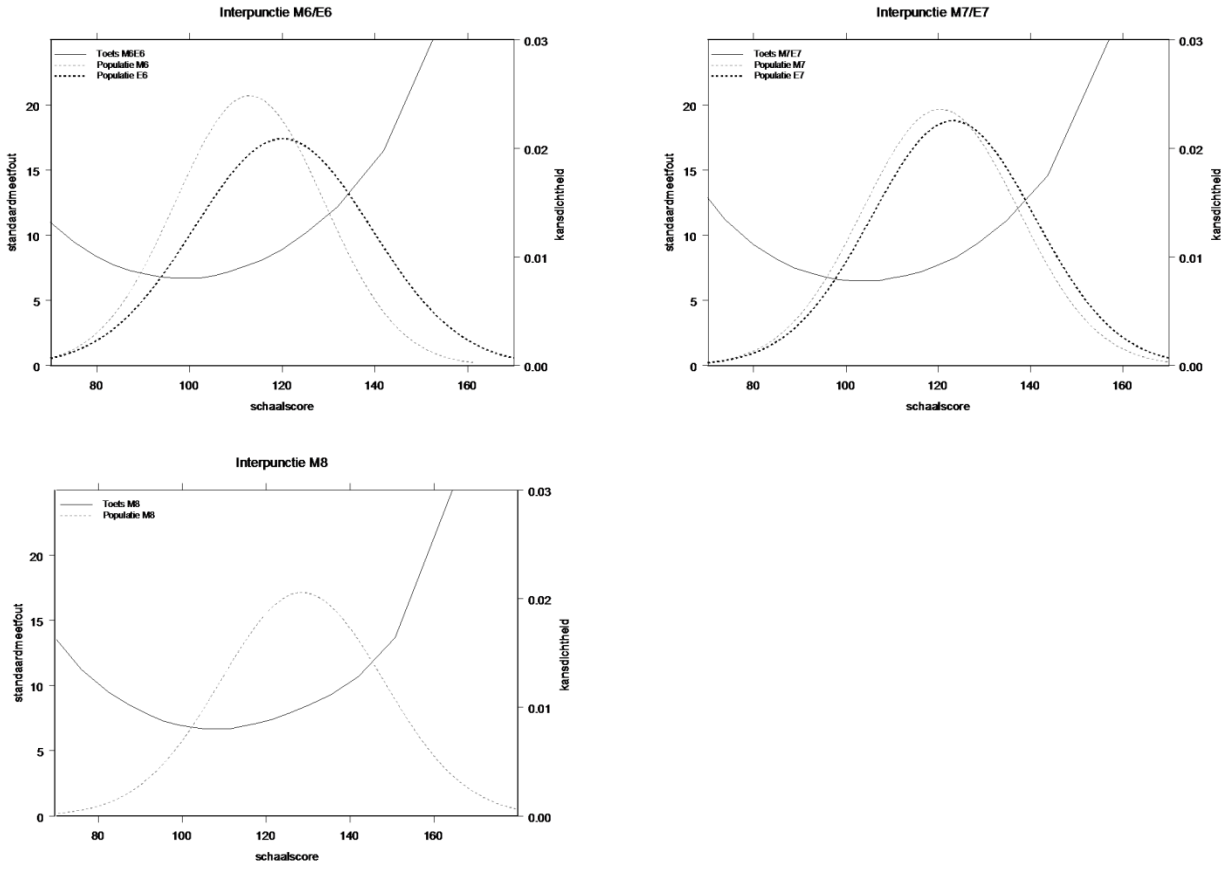
Er heeft geen test-hertest onderzoek plaatsgevonden. De afnamecontext van de LVS-toets Taalverzorging leent zich daar niet goed voor. Het feit dat alle items echter OPLM-gekalibreerd zijn, maakt het mogelijk een hertest te simuleren. We hebben een dubbele afname gesimuleerd van 1.000.000 leerlingen. Daarbij hebben we enerzijds de vaardigheidsverdeling van alle leerlingen, anderzijds alle itemparameters als uitgangspunt genomen. Steeds is een bepaalde vaardigheid aselect uit de verdeling genomen en zijn twee bij deze vaardigheid horende afnames gesimuleerd. Uiteindelijk is de correlatie tussen deze 1.000.000 dubbele (virtuele) afnames berekend. Men kan deze simulatie beschouwen als een test-hertestonderzoek onder ideale condities. De tweede toetsafname is immers volledig onafhankelijk van de eerste toetsafname en wordt niet beïnvloed door de kennis die de leerling mogelijk verworven heeft in de eerste toetsafname. Daarnaast is er geen sprake van invloed van een test-hertest-interval: beide afnames worden gesimuleerd alsof zij op hetzelfde moment plaats zouden vinden. De resultaten zijn weergegeven in de tabellen 5.1a t/m d. De uitkomsten komen vrijwel exact overeen met andere berekende coëfficiënten (MAcc, GLB) en leiden dan ook tot dezelfde conclusies met betrekking tot de betrouwbaarheid van de toetsen Taalverzorging.

5.3 Lokale betrouwbaarheid en meetnauwkeurigheid

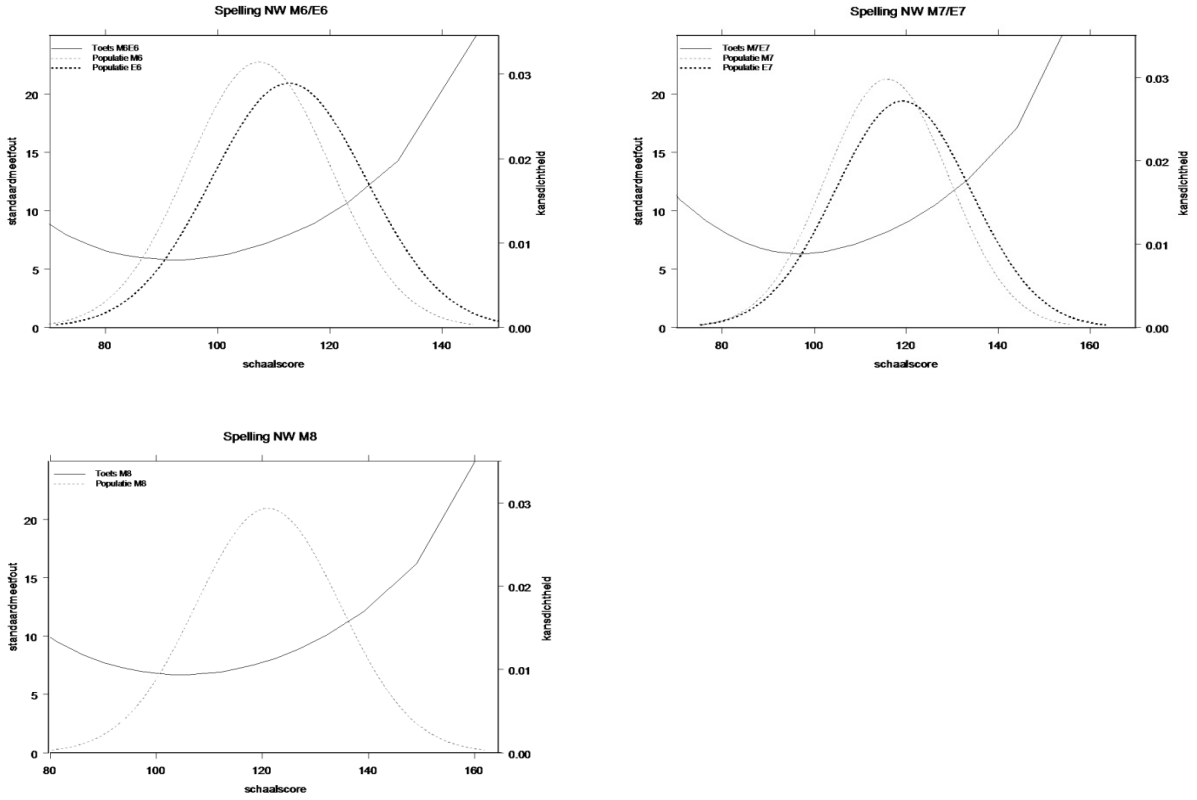
De hiervoor vermelde betrouwbaarheidscoëfficiënten hebben alleen betrekking op de globale meetnauwkeurigheid van de toetsen en geven geen beeld van de lokale meetnauwkeurigheid ervan. De figuren 5.1a t/m d geven grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid bij de toetsen Taalverzorging. In deze figuur staat de grootte van de meetfout op de vaardigheidsschaal afgebeeld (met verdelingskenmerken zoals aangegeven in tabel 4.11).

Ook zijn de kansdichtheidfuncties voor de normgroepen op de verschillende afnamemomenten opgenomen. Deze laten zien hoe de vaardigheid van de leerlingen verdeeld is in de normeringssteekproef. De figuren 5.1a t/m d maken duidelijk dat de meetfout kleiner is in de lagere en gemiddelde vaardigheidsregionen dan in de hogere vaardigheidsregionen. Over het algemeen is er sprake van een kleine meetfout voor de range waarin de vaardigheidsverdelingen op het betreffende afnamemoment grotendeels liggen. Bij Spelling werkwoorden M7/E7 is de verdeling van de meetfouten vlakker dan bij de andere onderdelen. De curve doet niet vermoeden dat de overall-betrouwbaarheid lager is dan bij de andere onderdelen.

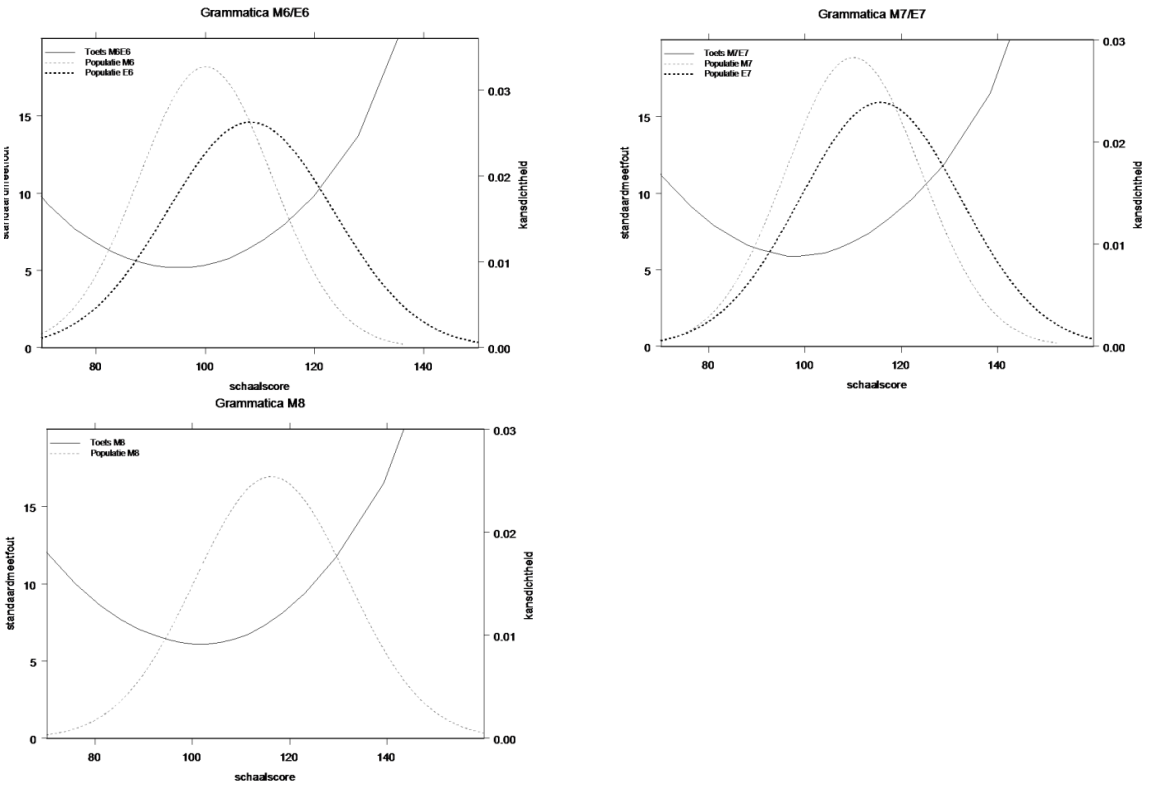
Figuur 5.1a Grootte van de meetfouten voor de toetsen Interpunctie M6/E6, M7/E7 en M8 en de kansdichtheidsfuncties voor de betreffende populaties



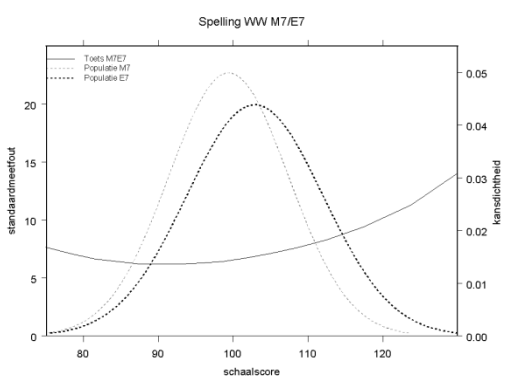
Figuur 5.1b Grootte van de meetfouten voor de toetsen Spelling niet-werkwoorden M6/E6, M7/E7 en M8 en de kansdichtheidsfuncties voor de betreffende populaties

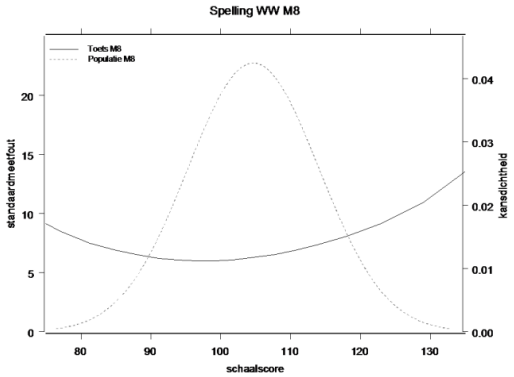


Figuur 5.1c Grootte van de meetfouten voor de toetsen Grammatica M6/E6, M7/E7 en M8 en de kansdichtheidsfuncties voor de betreffende populaties



Figuur 5.1d Grootte van de meetfouten voor de toetsen Spelling werkwoorden M7/E7 en M8 en de kansdichtheidsfuncties voor de betreffende populaties

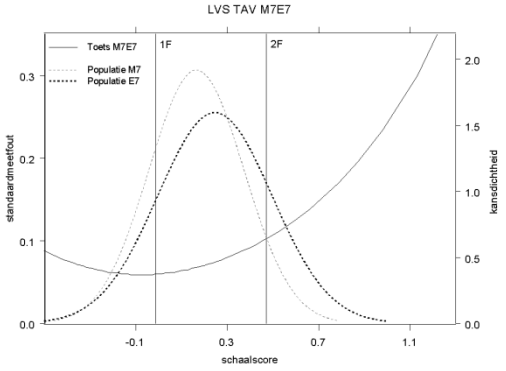




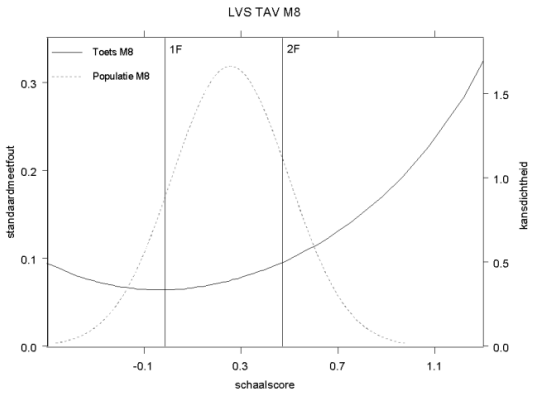
Lokale betrouwbaarheid referentieniveaus

De hiervoor vermelde figuren geven een beeld van de lokale meetnauwkeurigheid. Soortgelijke figuren hebben we ook voor de lokale meetnauwkeurigheid van de gerapporteerde referentieniveaus opgesteld. De figuren 5.2a en b geven grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid bij de gerapporteerde referentieniveaus Taalverzorging. Het is van belang om de lokale meetfout te bepalen als ergens een cesuur wordt gelegd, in dit geval: referentieniveau wel gehaald versus referentieniveau niet gehaald. Het verdient de voorkeur om juist rondom de cesuur zo nauwkeurig mogelijk te meten om het aantal misclassificaties te reduceren. De cesuur is bepaald over alle deelgebieden samen, dus over het hele domein Taalverzorging. In de figuren 5.2 a en b zijn de referentieniveaus 1F en 2F ingetekend. De weergegeven scores op de vaardigheidsschaal (x-as) komen bij M7/E7 en M8 overeen met de ruwe scores die vermeld zijn in hoofdstuk 4 in tabel 4.15. Zoals te zien is in de figuren, is de meetfout klein voor de range waarin de lijnen liggen die de referentieniveaus representeren. We kunnen dus concluderen dat met de toetsen Taalverzorging goed kan worden vastgesteld of de referentieniveaus 1F en 2F al dan niet zijn behaald.

Figuur 5.2a Grootte van de meetfouten van de referentieniveaus voor de toetsen taalverzorging M7/E7 en de kansdichtheidsfuncties voor de betreffende populaties



Figuur 5.2b Grootte van de meetfouten van de referentieniveaus voor de toetsen taalverzorging M8 en de kansdichtheidsfuncties voor de betreffende populaties



Betrouwbaarheidstabellen

De betekenis van de (lokale) meetnauwkeurigheid voor de beslissingen die met de toets genomen worden, is af te leiden uit betrouwbaarheidstabellen. Tabellen 5.3a t/m r laten voor alle afnamemomenten en alle deelgebieden zien hoe vaak de werkelijke vaardigheidsscore in dezelfde scoregroep valt als de geschatte vaardigheidsscore. Zo laat tabel 5.3a zien dat 82,3 procent van de leerlingen die halverwege groep 6 op basis van de M6/E6-toets Interpunctie in scoregroep V geclassificeerd wordt ook met hun werkelijke vaardigheidsscore in deze scoregroep geclassificeerd wordt. De kans dat een V-leerling terecht als V-leerling wordt bestempeld is, met andere woorden, ongeveer 80 procent. Verder laat de linkerkant van tabel 5.3a zien dat 16,2 procent van de leerlingen in scoregroep V een vaardigheidsscore heeft die in werkelijkheid in scoregroep IV valt. De overige getallen in tabellen 5.3a t/m r zijn op dezelfde wijze te interpreteren.

In de onderzoeksliteratuur is weinig geschreven over de beoordeling van betrouwbaarheidstabellen. Wanneer een betrouwbaarheidstabel als goed of voldoende kan worden beschouwd is onduidelijk en wat verwacht mag worden onder ideale omstandigheden is, voor zover ons bekend, niet onderzocht. Daarom worden betrouwbaarheidstabellen vaak samengevat in één of meerdere indices.
Wij kiezen ervoor om twee indices te presenteren te weten: de Marginal Classification Accuracy en de Accuracy plus/minus 1 niveau van Pilliner (1969), beide als totale samenvattende maat, en daarnaast per scoregroep eveneens de Accuracy plus/minus 1 niveau. We hebben voor deze maten gekozen omdat deze betrekkelijk intuïtief te interpreteren zijn. Dat geldt naar onze mening minder voor andere indices.
Zo presenteren we bewust niet de uitkomsten van de analyses per grenswaarde. Dergelijke analyses zullen in de regel resulteren in een positief beeld, maar doordat de dichotomisering bij een classificatieschema met meer dan één grenswaarde volledig kunstmatig is, zijn de uitkomsten naar onze mening nietszeggend. Toetsgebruikers zouden door de uitkomsten slechts op het verkeerde spoor gezet worden.
Bij de interpretatie van de indices maken we gebruik van Pilliner (1969) die als één van de weinigen een ambitieniveau geformuleerd heeft. Hij stelt als ambitieniveau dat 95 procent van de leerlingen in een niveaugroep in werkelijkheid ook in die niveaugroep moet scoren, ***of*** één niveaugroep daarboven ***of*** één niveaugroep daaronder. Dit ambitieniveau is gebaseerd op de veronderstelling dat geen enkele toets perfect meet en dat er dus altijd sprake is van foutieve classificaties. In dat licht is de maximale accuraatheid die op het individuele niveau bereikt kan worden plus of minus één niveaugroep.

Tabel 5.3a Betrouwbaarheidstabel toets Interpunctie M6/E6 op afnamemoment Medio groep 6, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**82,3**	16,2	1,5	0	0	E	**79,1**	18,9	2,0	0	0
IV	23,9	**49**	22,9	4,1	0,1	D	22,8	**51**	24,8	1,4	0
III	4,0	27	**39,9**	24,9	4,1	C	1,8	20,6	**52,7**	23	1,9
II	0,7	7,7	23,4	**38,9**	29,3	B	0,1	2,9	25,1	**46,2**	25,6
I	0,3	1,9	6,8	19,2	**71,8**	A	0	0,5	5,3	20,6	**73,6**

Tabel 5.3b Betrouwbaarheidstabel toets Spelling niet-werkwoorden M6/E6 op afnamemoment Medio groep 6, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**79,3**	18,1	2,4	0,1	0	E	**76,9**	20,5	2,6	0	0
IV	27,6	**44,1**	22,3	5,6	0,4	D	24,5	**47,3**	25,8	2,4	0
III	6,1	26,4	**35,3**	25,7	6,5	C	3,2	22,2	**47,3**	23,9	3,3
II	1,5	9,6	22,6	**35,4**	30,9	B	0,3	4,6	25,5	**41,7**	27,9
I	0,5	2,6	7,6	19,1	**70,1**	A	0,1	0,9	6,4	20,3	**72,2**

Tabel 5.3c Betrouwbaarheidstabel toets Grammatica M6/E6 op afnamemoment Medio groep 6, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**79,1**	18,6	2,2	0,1	0	E	**73,8**	22,3	3,9	0,1	0
IV	23,7	**49,3**	23,2	3,7	0,1	D	23,1	**48,4**	26,6	1,9	0
III	3,1	26,6	**42,6**	24,6	3,1	C	1,9	21,6	**54,2**	21,1	1,2
II	0,3	5,4	23,5	**43,5**	27,3	B	0	1,8	24,6	**51**	22,6
I	0,3	0,8	3,9	14,5	**80,6**	A	0,1	0,2	2,7	15,6	**81,4**

Tabel 5.3d Betrouwbaarheidstabel toets Interpunctie M6/E6 op afnamemoment Eind groep 6, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**84,7**	14,6	0,7	0	0	E	**81,5**	17,7	0,8	0	0
IV	22,7	**52,8**	21,5	2,9	0,1	D	18,8	**56,7**	23,9	0,7	0
III	3,5	26,4	**40,9**	25,1	4,2	C	1,2	20,5	**55,2**	21,6	1,5
II	0,9	8,3	23,6	**37,7**	29,4	B	0,1	3,0	25,4	**45,6**	25,9
I	0,7	3,0	8,0	18,8	**69,4**	A	0,2	1,0	6,7	20,5	**71,6**

Tabel 5.3e Betrouwbaarheidstabel toets Spelling niet-werkwoorden M6/E6 op afnamemoment Eind groep 6, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**81,4**	16,5	1,9	0,1	0	E	**79,6**	18,7	1,7	0	0
IV	26,8	**44,3**	22,8	5,7	0,4	D	23,9	**48,8**	26,8	0,5	0
III	6,8	26,2	**34,3**	25,3	7,3	C	2,6	18,1	**57**	14,7	7,6
II	2,1	10,4	22,1	**33**	32,4	B	0,4	3,8	33,7	**24,9**	37,2
I	1,0	3,7	8,5	18,1	**68,7**	A	0,3	1,5	13,6	13,9	**70,6**

Tabel 5.3f Betrouwbaarheidstabel toets Grammatica M6/E6 op afnamemoment Eind groep 6, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**81,7**	17,6	0,7	0	0	E	**75,8**	22,9	1,2	0	**81,7**
IV	21,8	**55,8**	20,4	2,0	0	D	19,7	**57,7**	22,1	0,5	21,8
III	2,2	25,8	**44,4**	24,7	2,9	C	0,7	19,8	**59,6**	19,1	2,2
II	0,5	7,0	24,9	**41,3**	26,3	B	0	1,9	24,2	**49,5**	0,5
I	0,9	2,3	6,2	15,7	**75**	A	0,4	0,8	4,8	16,3	0,9

Tabel 5.3g *Betrouwbaarheidstabel toets Interpunctie M7/E7 op afnamemoment Medio groep 7, conditioneel voor werkelijke vaardigheidsniveau*

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**82,9**	15,9	1,2	0	0	E	**78,6**	19,9	1,5	0	0
IV	23,2	**50,5**	22,7	3,5	0,1	D	20,5	**53,6**	24,8	1,1	0
III	3,4	26,1	**41,1**	25,4	4,0	C	1,5	20,6	**54,2**	22,1	1,6
II	0,6	7,4	23,6	**39,2**	29,1	B	0,1	2,8	25,8	**47**	24,4
I	0,3	2,1	7,3	19,8	**70,5**	A	0,1	0,6	5,5	21,1	**72,7**

Tabel 5.3h *Betrouwbaarheidstabel toets Spelling niet-werkwoorden M7/E7 op afnamemoment Medio groep 7, conditioneel voor het werkelijke vaardigheidsniveau*

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**80,3**	17,1	2,4	0,2	0	E	**77,4**	19,9	2,6	0	0
IV	28,2	**42**	22,6	6,6	0,6	D	25,8	**45,5**	25,8	2,8	0,1
III	7,8	26,2	**32,9**	25,1	8	C	4,3	22,7	**44,9**	23,9	4,1
II	2,3	10,7	21,8	**32,6**	32,6	B	0,7	5,8	25,3	**38,7**	29,6
I	0,8	3,3	8,1	18,1	**69,8**	A	0,2	1,3	7,2	19,4	**71,8**

Tabel 5.3i *Betrouwbaarheidstabel toets Grammatica M7/E7 op afnamemoment Medio groep 7, conditioneel voor het werkelijke vaardigheidsniveau*

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**82,9**	15,8	1,3	0	0	E	**77,4**	20,4	2,2	0	0
IV	25,5	**49,8**	21,3	3,4	0,1	D	22,6	**50,9**	25	1,4	0
III	3,8	26,7	**40,1**	25,3	4,2	C	1,7	21,1	**53,4**	22	1,7
II	0,7	7,8	23,9	**39,6**	28	B	0,1	2,7	24,6	**46,6**	26
I	0,3	1,7	5,5	15,5	**77**	A	0,1	0,5	4,2	16,1	**79,2**

Tabel 5.3j Betrouwbaarheidstabel toets Spelling werkwoorden M7/E7 op afnamemoment Medio groep 7, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**73,9**	17,5	6,4	2,0	0,3	E	**70,1**	19,9	8,4	1,5	0,1
IV	34,5	**30**	20,6	11,4	3,5	D	33	**31,2**	25,4	8,8	1,6
III	14,4	23,2	**25,5**	22,8	14,1	C	10,9	21,7	**33,5**	23,8	10,1
II	5,0	12,4	19,9	**27,4**	35,3	B	2,4	8,6	23,4	**31,7**	33,8
I	1,1	3,8	8,4	17,3	**69,2**	A	0,4	2,0	8,2	18,9	**70,5**

Tabel 5.3k Betrouwbaarheidstabel toets Interpunctie M7/E7 op afnamemoment Eind groep 7, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**83,6**	15,4	0,9	0	0	E	**79,6**	19,2	1,2	0	0
IV	23,5	**51,3**	21,8	3,3	0,1	D	20,6	**54,7**	23,8	0,9	0
III	3,6	26,9	**40,5**	25	4,0	C	1,4	20,5	**54,4**	22,1	1,6
II	0,8	7,9	23,3	**38,5**	29,4	B	0,1	2,8	24,7	**45,9**	26,5
I	0,6	2,6	7,6	19	**70,2**	A	0,1	0,9	6,1	20,4	**72,5**

Tabel 5.3l Betrouwbaarheidstabel toets Spelling niet-werkwoorden M7/E7 op afnamemoment Eind groep 7, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**81,1**	16,8	2,0	0,1	0	E	**79,7**	18,5	1,9	0	0
IV	27,5	**43,5**	22,5	6,1	0,5	D	23,7	**47,9**	26,1	2,4	0
III	7,5	26,4	**33,2**	25,2	7,8	C	3,8	22,8	**46,1**	23,5	3,7
II	2,4	10,8	21,7	**32,4**	32,7	B	0,7	5,7	25,4	**38,7**	29,5
I	1,0	3,6	8,2	17,7	**69,5**	A	0,3	1,5	7,6	19,4	**71,1**

Tabel 5.3m Betrouwbaarheidstabel toets Grammatica M7/E7 op afnamemoment Eind groep 7, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**79,9**	19,9	0,2	0	0	E	**76,2**	22,7	1,1	0	0
IV	16,7	**64,3**	16,9	2,1	0	D	20,1	**57,5**	21,9	0,5	0
III	1,0	28,4	**42**	24,8	3,7	C	1,0	19,6	**55,7**	22,2	1,5
II	0,3	8,7	24,2	**38,7**	28,1	B	0,1	3,0	26,5	**46,4**	24,1
I	0,5	2,9	6,8	15,9	**73,8**	A	0,3	1,0	5,6	17,1	**76,1**

Tabel 5.3n Betrouwbaarheidstabel toets Spelling werkwoorden M7/E7 op afnamemoment Eind groep 7, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**74,6**	18,3	5,6	1,3	0,1	E	**72,9**	19,7	6,5	0,7	0
IV	32,4	**32,9**	21,6	10,6	2,5	D	31,7	**34,6**	25,9	7	0,8
III	12,4	24,2	**27,1**	23,8	12,4	C	9,0	22,4	**36,4**	24,2	8,1
II	4,1	12	20,2	**28,7**	35	B	1,8	8,0	24,3	**33,7**	32,2
I	1,1	3,8	8,5	17,9	**68,7**	A	0,3	1,8	8,1	19,5	**70,3**

Tabel 5.3o Betrouwbaarheidstabel toets Interpunctie M8 op afnamemoment Medio groep 8, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	84,3	15	0,7	0	0	E	80,9	18,3	0,8	0	0
IV	21,8	53,8	21,8	2,6	0	D	19,2	56,9	23,3	0,6	0
III	2,6	26,2	43,6	24,8	3,0	C	1,0	19,6	57,2	21,1	1,1
II	0,3	6,1	24	42,7	26,8	B	0	1,8	24	50	24,1
I	0,3	1,6	6,3	19,7	72,1	A	0,1	0,4	4,3	20,4	74,7

Tabel 5.3p Betrouwbaarheidstabel toets Spelling niet-werkwoorden M8 op afnamemoment Medio groep 8, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**79,8**	17,2	2,1	1	0	E	**76,6**	20,1	3,2	0,1	0
IV	28,6	**41,7**	13,9	15,2	0,5	D	26	**44,1**	26,5	3,2	0,1
III	9,7	31,1	**18,3**	37	4,0	C	4,2	22,6	**45,2**	24	4,0
II	2,9	13,9	11,7	**46,8**	24,6	B	0,5	5,3	25,5	**40**	28,6
I	0,6	2,8	3,0	23	**70,7**	A	0,1	1,0	6,6	19,6	**72,6**

Tabel 5.3q Betrouwbaarheidstabel toets Grammatica M8 op afnamemoment Medio groep 8, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**85,3**	13,8	0,8	0	0	E	**76,9**	22,4	0,7	0	0
IV	23,8	**51,8**	21,3	3,0	0,1	D	16,4	**61,6**	20,4	1,6	0
III	3,3	26,4	**41,2**	25,2	3,9	C	0,9	27,2	**47,6**	22,9	1,5
II	0,8	7,9	22,7	**37,6**	30,9	B	0,1	4,7	22,4	**45,3**	27,4
I	1,0	2,6	6,2	14,4	**75,8**	A	0,4	1,5	5,1	16,8	**76,2**

Tabel 5.3r Betrouwbaarheidstabel toets Spelling werkwoorden M8 op afnamemoment Medio groep 8, conditioneel voor het werkelijke vaardigheidsniveau

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	**76,3**	17,7	4,9	1,0	0,1	E	**72,8**	20,2	6,3	0,6	0
IV	31,2	**35,3**	22,5	9,4	1,6	D	30,3	**36,4**	26,5	6,3	0,5
III	9,8	24,7	**30,2**	24,9	10,3	C	7,1	22,5	**39,4**	24,6	6,4
II	2,5	10,7	21,7	**32**	33,1	B	1,0	6,6	25,1	**37,3**	30
I	0,4	2,4	7,2	17,8	**72,2**	A	0,1	1,0	6,3	19,2	**73,5**

Tabel 5.4a Samenvattende indices toetsen Interpunctie M6, E6, M7, E7 en M8

IP	**Toets M6/E6, Afnamemoment M6**		**Toets M6/E6, afnamemoment E6**		**Toets M7/E7, afnamemoment M7**		**Toets M7/E7, afnamemoment E7**		**Toets M8, afnamemoment M8**	
	score-groep I t/m V	score-groep A t/m E	score-groep I t/m V	score-groep A t/m E	score-groep I t/m V	score-groep A t/m E	score-groep I t/m V	score-groep A t/m E	scoregroep I t/m V	scoregroep A t/m E
Marginal classification accuracy	56,7	59,0	56,6	59,6	56,5	59,3	57,0	59,3	59,7	62,4
Accuracy plus/ minus 1 niveau (Pilliner)	93,8	96,5	93,5	96,3	93,7	96,7	93,8	96,5	95,3	97,6

Tabel 5.4b Samenvattende indices toetsen Spelling niet-werkwoorden M6, E6, M7, E7 en M8

NW	**Toets M6/E6, Afnamemoment M6**		**Toets M6/E6, afnamemoment E6**		**Toets M7/E7, afnamemoment M7**		**Toets M7/E7, afnamemoment E7**		**Toets M8, afnamemoment M8**	
	score-groep I t/m V	score-groep A t/m E	score-groep I t/m V	score-groep A t/m E	score-groep I t/m V	score-groep A t/m E	score-groep I t/m V	score-groep A t/m E	score-groep I t/m V	score-groep A t/m E
Marginal classification accuracy	53,1	54,9	52,4	56,6	51,6	53,4	52,0	54,0	54,2	53,7
Accuracy plus/minus 1 niveau (Pilliner)	91,5	94,7	90,4	91,7	89,8	93,4	90,1	93,7	88,6	93,7

Tabel 5.4c Samenvattende indices toetsen Grammatica M6, E6, M7, E7 en M8

GR	**Toets M6/E6, Afnamemoment M6**		**Toets M6/E6, afnamemoment E6**		**Toets M7/E7, afnamemoment M7**		**Toets M7/E7, afnamemoment E7**		**Toets M8, afnamemoment M8**	
	score-groep I t/m V	score-groep A t/m E	score-groep I t/m V	score-groep A t/m E	score-groep I t/m V	score-groep A t/m E	score-groep I t/m V	score-groep A t/m E	score-groep I t/m V	score-groep A t/m E
Marginal classification accuracy	60,2	61,8	61,9	63,0	59,8	60,9	61,3	61,1	59,1	60,0
Accuracy plus/minus 1 niveau (Pilliner)	95,6	97,2	95,4	97,6	94,5	96,9	95,0	96,9	94,4	96,4

Tabel 5.4d Samenvattende indices toetsen Spelling werkwoorden M7, E7 en M8

WW	**Toets M7/E7, afnamemoment M7**		**Toets M7/E7, afnamemoment E7**		**Toets M8, afnamemoment M8**	
	scoregroep I t/m V	scoregroep A t/m E	scoregroep I t/m V	scoregroep A t/m E	scoregroep I t/m V	scoregroep A t/m E
Marginal classification accuracy	44,9	45,7	47,0	47,9	49,1	50,5
Accuracy plus/minus 1 niveau (Pilliner)	83,3	86,9	85,2	88,8	87,8	91,2

De samenvattende indices voor alle afnamemomenten en alle deelgebieden zijn te vinden in de tabellen 5.4a t/m d. Waar de betrouwbaarheidstabellen laten zien dat er behoorlijk wat leerlingen zijn die op basis van hun geschatte vaardigheidsscore een niveaugroep te hoog of te laag geplaatst worden, maken de tabellen 5.4a t/m d aannemelijk dat de uitkomsten wel redelijk in lijn liggen met het ambitieniveau zoals dat geformuleerd is door Pilliner (1969). Gemiddeld gezien scoort, afhankelijk van het afnamemoment en de gekozen indeling in scoregroepen, 82 tot 97 procent van de leerlingen in een scoregroep ook in werkelijkheid in die scoregroep, ***of*** één scoregroep daarboven ***of*** één scoregroep daaronder. De *Marginal Classification Accuracy* loopt uiteen van 44 tot 62 procent. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore bij beide afnamemomenten gemiddeld gezien in ruim 50 procent van de gevallen overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. Omdat zowel de *plus/minus 1 niveau-index* als de *Marginal Classification Accuracy* wat lager liggen dan wenselijk is, moet de indeling van leerlingen in score-groepen met de nodige voorzichtigheid geïnterpreteerd worden. De verschillende toetsen Taalverzorging weten vooral de laagst en hoogst scorende leerlingen accuraat te classificeren; in het midden is de accuraatheid van de classificatie minder. Dit pas bij één van de doelen van deze toets: signaleren welke leerlingen extra aandacht of extra uitdaging nodig hebben. Het percentage misclassificaties is bij de middelste scoregroepen het hoogst, te weten bij de scoregroep III, respectievelijk scoregroep C.

Conclusie

De vaardigheidsgroei voor de deelgebieden van Taalverzorging voltrekt zich langzaam in de tijd.
De verschillen tussen vaardigheidsscores op achtereenvolgende meettijdstippen zijn klein, ook al neemt men slechts een maal per jaar een toets af voor deze vaardigheid. Bovendien is er sprake van meetfouten.
De toch al kleine verschillen in vaardigheidsgroei moeten tegen de achtergrond van die meetfouten worden geïnterpreteerd. Dit betekent dat men weliswaar uitspraken kan doen over de vaardigheidsgroei van een leerling, maar dat deze uitspraken met voorzichtigheid dienen te worden gehanteerd. Dit geldt ook wanneer men de progressie van een leerling volgt in termen van vaardigheidsniveaus of een vergelijking maakt met andere leerlingen in termen van vaardigheidsniveaus. Want ook bij de indeling in vaardigheidsniveaus speelt de nauwkeurigheid van de toets een rol. Hoe de leerkracht hier in de praktijk mee om moet gaan wordt toegelicht in de handleiding voor de leerkracht in paragraaf 4.3.

Betrouwbaarheidstabellen referentieniveaus

Tabel 5.5 Betrouwbaarheidstabel voor de referentieniveaus Taalverzorging, conditioneel voor het werkelijke vaardigheidsniveau (in procenten)

M7	scoregroep waarin score valt		
scoregroep	<1F	1F--2F	>2F
<1F	**82,74**	17,26	0,00
1F--2F	5,57	**90,03**	4,40
>2F	0,00	20,58	**79,42**

E7	scoregroep waarin score valt		
scoregroep	<1F	1F--2F	>2F
<1F	**87,08**	12,92	0,00
1F--2F	4,67	**89,47**	5,85
>2F	0,00	17,65	**82,35**

M8	scoregroep waarin score valt		
scoregroep	<1F	1F--2F	>2F
<1F	**86,34**	13,66	0,00
1F--2F	4,18	**90,83**	4,99
>2F	0,00	15,76	**84,24**

Tabel 5.6 Samenvattende gegevens

	M7	E7	M8
marginal classification accuracy	0,88	0,88	0,89
accuracy plus/minus 1 niveau	1,00	1,00	1,00

In tabel 5.6 is te zien dat voor Taalverzorging de meeste leerlingen met hun werkelijke vaardigheidsscore in dezelfde referentieniveaugroep vallen als met hun geschatte vaardigheidsscore. Er zijn echter ook leerlingen die op basis van hun geschatte vaardigheidsscore een niveaugroep te hoog of te laag geplaatst worden. Desalniettemin zijn de uitkomsten in lijn met het ambitieniveau zoals dat geformuleerd is door Pilliner (1969). Praktisch gesproken scoren bijna alle leerlingen in een scoregroep op basis van hun ware score, ook in werkelijkheid in die scoregroep, ***of*** één scoregroep daarboven ***of*** één scoregroep daaronder. De *Marginal Classification Accuracy* (d.w.z. de som van de diagonaal in de verwarringsmatrix als proportie van het totale aantal leerlingen) bedraagt 0,88 en 0,89. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore gemiddeld gezien in bijna 90% procent van de gevallen overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. Hieruit kunnen we concluderen dat de voornaamste functie van de toetsen Taalverzorging voor de groepen 6 tot en met 8 geslaagd is: het zo betrouwbaar mogelijk vastleggen op welk referentieniveau leerlingen in de bovenbouw functioneren.

6 Validiteit

In onderstaande paragrafen komt de validiteit van de LVS-toetsen Taalverzorging aan de orde. Voor leervorderingentoetsen is validiteit in termen van inhoudsvaliditeit bijzonder belangrijk. We bespreken deze in paragraaf 6.1, met name door terug te blikken op de inhoudsverantwoording die aan de orde is geweest in paragraaf 3.2. Juist vanwege het feit dat we daarnaast ook (latente) vaardigheden veronderstellen die aan de basis liggen van de toetsitems is echter ook de begripsvaliditeit van de toetsen van belang. We bespreken deze in paragraaf 6.2. Criteriumvaliditeit is bij de toetsen taalverzorging niet aan de orde: LVS-toetsen kennen geen voorspellende pretenties.

6.1 Inhoudsvaliditeit

De inhoudsvaliditeit van een toets heeft betrekking op de vraag in hoeverre de opgaven in een toets een welomschreven en afgebakend universum representeren van mogelijk in de toets op te nemen opgaven.
De inhoudsvaliditeit van de toetsen Taalverzorging wordt onder meer gegarandeerd door de wijze waarop de opgaven ontwikkeld zijn. In de inhoudsverantwoording (zie paragraaf 3.2) is al aangegeven dat aan de basis van de ontwikkeling van de opgaven de indeling in vier vaardigheidsaspecten (spelling niet-werkwoorden, spelling werkwoorden, grammatica en interpunctie) ligt. Deze indeling is ontwikkeld aan de hand van de visie van Cito-toetsdeskundigen op wat het construct ‘taalverzorging’ inhoudt en is gevoed door documenten van het Referentiekader Taal en Rekenen (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a), Leerstoflijnen begrippenlijst en taalverzorging (van der Beek & Paus, 2011), de kerndoelen Nederlandse taal voor het basisonderwijs (Ministerie van Onderwijs, Cultuur en Wetenschappen, 2006) en de tussendoelen en leerstoflijnen van TULE (TULE, 2008).
De constructie van de opgaven is afgeleid van deze domeinindeling en ook de definitieve selectie van opgaven in de toetsen is gebaseerd op een gewenste verdeling van opgaven binnen en over de verschillende deelgebieden en categorieën van taalverzorging. Beoogd is de toetsen onafhankelijk samen te stellen van de verschillende onderwijsmethoden in die zin dat de getoetste stof in het merendeel van de methodes aan bod is gekomen, maar dat de inhoud van de toetsen niet specifiek aansluit bij een bepaalde methode. Bij de constructie van de opgaven zijn (in de evenwichtig samengestelde constructiegroepen) praktijkdeskundige leerkrachten uit de bovenbouw van het onderwijs betrokken zodat de opgaven voor wat betreft moeilijkheid en voor wat betreft context aansluiten bij het ontwikkelingsniveau van leerlingen van groep 6 tot en met 8. Kortom, de gewenste toetsinhoud – in termen van beoogde aantallen items voor de onderscheiden categorieën – is beargumenteerd op basis van wetenschappelijk goed verdedigbare keuzes omtrent referentiekader, kerndoelen, leerstoflijnen en tussendoelen en in overeenstemming gebracht met de meest gangbare lesmethoden. De opgaven zijn geconstrueerd door leerkrachten uit het basisonderwijs, van een correcte context voorzien en empirisch uitgetest in proef- en normeringsonderzoeken. Voorafgaand aan de normeringsonderzoeken hebben we de toetsen nogmaals voorgelegd aan een klankbordgroep, bestaande uit leerkrachten, intern begeleiders en schoolleiders, zodat ook de laatste versie nog van commentaar voorzien is dat door ons verwerkt is.
In termen van moeilijkheidsgraad kunnen we constateren dat de opgaven uit de toets juist zijn afgestemd op de doelgroep. Het merendeel (90%) van de opgaven is niet te moeilijk (p-waarde < 0,20) of te makkelijk 3 (p-waarde > 0,80) zijn. Dit alles vormt een degelijke basis voor de inhoudsvaliditeit van de toetsen.
De gewenste verdeling over subdomeinen en categorieën, ten slotte, is in de uiteindelijke itemselectie daadwerkelijk gerealiseerd.

6.2 Begripsvaliditeit

In deze paragraaf worden resultaten met betrekking tot verschillende aspecten van begripsvaliditeit besproken. Deze komen overeen met de aspecten die in het COTAN Beoordelingssysteem (Evers et al.,

2010) worden besproken als relevant voor de begripsvaliditeit. Dit zijn achtereenvolgens de aspecten unidimensionaliteit (paragraaf 6.2.1), itemkwaliteit (paragraaf 6.2.2), itembias (paragraaf 6.2.3), convergente en discriminante validiteit (6.2.4) en verschillen tussen relevante subgroepen (6.2.5).

6.2.1 Unidimensionaliteit

In hoofdstuk 4 werd beschreven dat de opgaven die werden ontwikkeld met het oog op de toetsen Taalverzorging na de kalibratie een gekalibreerde opgavenbank vormen bestaande uit vier aparte opgavenbanken voor de onderscheiden onderdelen van Taalverzorging. Bij de analyse van de leerlingantwoorden is nagegaan of de opgaven van de toetsen van de verschillende deelgebieden een beroep doen op hetzelfde complex aan vaardigheden. Opgaven die niet voldeden aan de passingscriteria die we ontleenden aan het toegepaste IRT meetmodel (OPLM) en die we beschreven in paragraaf 4.3.2, zijn uit de opgavenverzameling verwijderd. Het betreft opgaven waarop werd gegokt, opgaven die onjuist geformuleerd zijn, opgaven die een slecht onderscheidend vermogen bleken te hebben, of opgaven die bij nader inzien toch niet alleen de vaardigheid van de subdomeinen of deelgebieden bleken te meten.
We hebben verschillende analyses gerapporteerd met betrekking tot de passing van het onderliggende meetmodel van de toetsen, waaruit blijkt dat die passing bevredigend is. De grafische voorstellingen van de S-toetsen gaven voor de meeste opgaven een bevredigend beeld. Dat is een sterke aanwijzing dat het meetinstrument en het meetmodel dat ontwikkeld is, respectievelijk gebruikt is, adequaat zijn om het gedrag van de leerlingen te verklaren. Het blijkt dat gemeten verschillen in gedrag tussen de leerlingen te verklaren zijn door één unidimensioneel concept per deelgebied van taalverzorging. Zowel de verdeling van overschrijdingskansen bij de S-toetsen als de beoordeling van de modelfit in termen van R1c geven een bevredigend beeld.

Een andere methode om de modelpassing te verantwoorden betreft de beoordeling van de nauwkeurigheid van de itemparameterschattingen op basis van een constante, de ‘c’ uit het COTAN-systeem (Evers et al., 2010) en ondersteunt de al genoemde conclusies. Zowel de ranges als de gemiddelden van de aangetroffen waarden voor ‘c’ (vrijwel alle waarden zijn lager dan 0,20) zijn als ‘goed’ te kwalificeren. Ze duiden op een hoge nauwkeurigheid van de itemparameterschattingen en daarmee op een hoge itemkwaliteit (zie ook verderop).

Op basis van de hierboven beschreven resultaten valt de conclusie te trekken, dat voor de toetsen LVS Taalverzorging voor de afnamemomenten M6, E6, M7, E7 en M8 de kalibraties geslaagd zijn. Het is daarom aannemelijk dat er sprake is van unidimensionaliteit per deelgebied en dat de afzonderlijk gekalibreerde opgavenbanken telkens één latente trek meten. Dat het bij deze latente trekken om de deelvaardigheden gaat die we samen ‘taalverzorging’ noemen, blijkt – naast de conclusies met betrekking tot de inhoudsvaliditeit in de vorige paragraaf – uit de resultaten van analyses die we in de rest van dit hoofdstuk presenteren.

6.2.2 Itemkwaliteit

In deze paragraaf vatten we in tabel 6.1 een aantal gegevens samen die betrekking hebben op de itemparameters van de toetsen Taalverzorging voor groep 6 tot en met 8. Voor een overzicht van alle gegevens per item, zie bijlage 2.
De gemiddelde moeilijkheidsgraad van de toetsen ligt op het (vooraf) gewenste niveau, namelijk voor interpunctie tussen 0,67 (M6) en 0,75 (E6), voor spelling niet-werkwoorden tussen 0,68 (M6) en 0,75 (E6), voor grammatica tussen 0,67 (M6) en 0,71 (E7) en voor spelling werkwoorden tussen 0,59 (M8) en 0,65 (E7). De gemiddelde moeilijkheidsgraad voldoet daarmee aan het gestelde doel, namelijk een optimaal onderscheidend vermogen bij de groep met een lage of gemiddelde vaardigheid (zie verder hoofdstuk 5 over lokale meetnauwkeurigheid), terwijl de toetsen niet als bijzonder moeilijk zullen worden ervaren door de doorsnee leerling. De moeilijkheidsgraad van de afzonderlijke opgaven kent een goede spreiding; er zijn

zowel moeilijke als gemakkelijke opgaven in de toetsen opgenomen. Het merendeel (90%) van de opgaven is niet te moeilijk (p-waarde < 0,20) of te makkelijk (p-waarde > 0,80) zijn.
De samenhang tussen item- en totaalscore is zowel in termen van R_{ir} als in termen van R_{it} weergegeven. Eerstgenoemde kengetallen geven een reëlere inschatting van die samenhang, maar er zijn geen normwaarden voor beschikbaar in het COTAN-beoordelingssysteem (Evers et al, 2010); voor R_{it} is dat wel het geval. In de tabel is te zien dat geen enkele opgave een lagere R_{it}-waarde dan .25 heeft. De laagste waarden treffen we aan bij drie opgaven van de toetsen Spelling werkwoorden, maar de waarden vallen wel binnen de range voldoende. De gemiddelde R_{it}-waarden zijn voor de toetsen te kenschetsen als 'goed' (gemiddelde R_{it} > 0.30).

Tabel 6.1 Samenvatting itemkenmerken voor de toetsen Taalverzorging alle deelgebieden op de afnamemomenten M6, E6,M7, E7 en M8

	P-waarde			R_{it}		
IP	**min**	**max**	**gem**	**min**	**max**	**gem**
M6	0,50	0,85	0,67	0,36	0,56	0,46
E6	0,57	0,90	0,75	0,37	0,55	0,46
M7	0,56	0,82	0,70	0,35	0,59	0,52
E7	0,59	0,85	0,73	0,33	0,55	0,49
M8	0,55	0,85	0,71	0,33	0,61	0,47

	P-waarde			R_{it}		
NW	**min**	**max**	**gem**	**min**	**max**	**gem**
M6	0,44	0,88	0,68	0,38	0,56	0,45
E6	0,51	0,92	0,75	0,38	0,55	0,45
M7	0,37	0,86	0,71	0,34	0,51	0,42
E7	0,41	0,88	0,74	0,35	0,52	0,43
M8	0,39	0,86	0,69	0,30	0,51	0,42

	P-waarde			R_{it}		
GR	**min**	**max**	**gem**	**min**	**max**	**gem**
M6	0,49	0,67	0,58	0,31	0,56	0,47
E6	0,57	0,77	0,69	0,35	0,60	0,51
M7	0,49	0,79	0,66	0,34	0,59	0,46
E7	0,56	0,83	0,71	0,37	0,62	0,49
M8	0,55	0,85	0,71	0,33	0,61	0,47

	P-waarde			R_{it}		
WW	**min**	**max**	**gem**	**min**	**max**	**gem**
M6						
E6						
M7	0,42	0,84	0,60	0,26	0,44	0,34
E7	0,47	0,88	0,65	0,27	0,46	0,36
M8	0,43	0,83	0,59	0,28	0,46	0,37

Zoals eerder vermeld is voor de toetsen ook de constante ‘c’ berekend voor alle deelgebieden en afnamemomenten. De nauwkeurigheid van de itemparameterschattingen is voor alle opgaven en voor beide afnamemomenten als ‘goed’ te kenschetsen.

6.2.3 Convergente en discriminante validiteit

Wanneer we de begripsvaliditeit van de toetsen Taalverzorging voor groep 6 tot en met 8 evalueren kunnen we dit doen door na te gaan in hoeverre de toetsscores samenhang vertonen met de scores op andere leervorderingentoetsen. Als we daarbij op de eerste plaats toetsen kiezen die variëren in de mate van overlap in meetpretentie, krijgen we op deze wijze zicht op de convergente versus discriminante (of divergente) validiteit. Dit is gebeurd door een aantal taaltoetsen te kiezen uit het Cito Volgsysteem primair onderwijs van de tweede generatie (zie paragraaf 6.2.3.1). De LVS-toetsen van de tweede generatie hebben betrekking op alle afnamemomenten tussen M6 en M8.

6.2.3.1 Samenhangen met andere toetsen

Aan de zogeheten ‘volgscholen’ – dit zijn scholen die hebben toegezegd meerdere keren te willen deelnemen aan de proef- en normeringsonderzoeken – die hadden deelgenomen aan de normeringsonderzoeken Taalverzorging, is via e-mail gevraagd om toetsgegevens beschikbaar te stellen voor de vaardigheden Rekenen(Cito, 2006) Spelling (Cito, 2009), Woordenschat (Cito, 2011), Begrijpend lezen (Cito, 2009), Technisch Lezen (Cito, 2010), AVI en DMT (Cito, 2009) via de geautomatiseerde dataretourfunctie van het Computerprogramma LOVS. Al deze toetsen uit de tweede generatie van het LVS zijn door de Cotan (Evers et al, 2010) op alle relevante onderdelen (criteriumvaliditeit is niet van toepassing) met een goed of voldoende beoordeeld.
Cito dataretour is een exporttool die basisscholen in staat stelt om op vrijwillige basis hun LVS-resultaten naar Cito te sturen voor (interne) onderzoeksdoeleinden. Veel basisscholen gaven gehoor aan de oproep (voor aantallen leerlingen zie tabel 6.2).

Tabel 6.2 Aantal deelnemende leerlingen op volgscholen

	aantal leerlingen volgscholen bij normeringen			
	E6 2014	**E7 2014**	**M7 2015**	**M8 2015**
M6 2014	646			
M7 2014		645		
E6 2014			738	
E7 2014				725

Tabel 6.3 Aantal deelnemende leerlingen aan correlatie-onderzoek per onderdeel

	Aantal leerlingen die naast Taalverzorging een of meer andere toetsen gemaakt hebben							
	Rek	**Sp-NW**	**Blz**	**Avi**	**DMT**	**Lees-tempo**	**Sp-ww**	**Wst**
M6	268	271	271	108	195	154		
E6	479	480		198	410	181		
M7	774	774	773	221	602	267	556	
E7	461	484	172	116	403	190	396	472
M8	406	408	408	80	354	136	245	384

Op basis van algemene cognitieve verschillen in aanleg (intelligentie) is er altijd sprake van een zekere samenhang tussen leervorderingen op verschillende vakgebieden. Hoe sterk deze samenhang is, hangt af van het vakgebied. De verwachting is dat we met name een grote samenhang aantreffen tussen de toetsen Spelling van de tweede generatie en de toetsen Spelling als onderdeel van de toetsen Taalverzorging aangezien het om dezelfde vaardigheid gaat. We verwachten verder dat de verschillende deelvaardigheden van taalverzorging redelijk sterk samenhangen met de vaardigheid in technisch lezen en dan vooral met de vaardigheid die naar voren komt bij de toetsen Leestempo. Technisch lezen en de vier deelvaardigheden die onderdeel zijn van Taalverzorging zijn met name 'technische' vaardigheden en geen semantische vaardigheden. Voor een goede vaardigheid spelling, interpunctie en grammatica zijn vooral de kennis en toepassing van regels en conventies van belang.
De sterkste samenhangen verwacht we voor de gearceerde cellen in tabel 6.4: spelling niet-werkwoorden, spelling werkwoorden en leestempo. Zowel spelling als leestempo zijn immers gericht op de herkenning van het woordbeeld. In de toetsvormen die gehanteerd worden, hoeft er niet actief een woord gespeld of verklankt te worden (dit laatste i.t.t. Avi of DMT), maar wordt de receptieve vaardigheid woordbeeldherkenning aangesproken. De vaardigheden waarop een beroep gedaan wordt, zijn sterk verwant.
Voor de overige vakgebieden verwachten we een matige samenhang. Hoewel begrijpend lezen en woordenschat ook tot het taaldomein behoren, zijn deze vaardigheden meer semantisch van aard en daarmee duidelijk te onderscheiden van de deelvaardigheden van taalverzorging. Voor het spellen van woorden is een zekere basiswoordenschat van belang, maar het is niet nodig om alle woordbetekenissen te kennen om woorden correct te kunnen schrijven. Voor de vaardigheden interpunctie en grammatica zou naast technisch lezen toch ook nog begrijpend lezen enigszins van belang kunnen zijn om zinsstructuren te doorgronden. Met name ten aanzien van interpunctie is te verwachten dat er sprake zal zijn van enige samenhang met begrijpend lezen.
De correlatie met rekenen-wiskunde zal naar verwachting matig zijn, omdat het een geheel andere vaardigheid betreft dan de deelvaardigheden van taalverzorging.

In tabel 6.4 worden de (voor attenuatie gecorrigeerde) correlatiecoëfficiënten gerapporteerd tussen de hierboven genoemde toetsen en Taalverzorging op de afnamemomenten M6, E6, M7, E7 en M8.

Tabel 6.4 Correlaties tussen de deelvaardigheden van Taalverzorging en verschillende andere LVS-onderdelen*

M6	Rek	Sp-nw	Blz	Avi	DMT	Leestempo
IP	0,43	0,53	0,55	0,29	0,44	0,46
NW	0,44	0,76	0,44	0,38	0,57	0,69
GR	0,38	0,37	0,42	0,10	0,27	0,16

E6	Rek	Sp-nw	Blz	Avi	DMT	Leestempo
IP	0,55	0,56		0,30	0,43	0,37
NW	0,48	0,80	0,55	0,30	0,67	0,65
GR	0,51	0,41	0,66	0,17	0,39	0,41

M7	Rek	Sp-nw	Blz	Avi	DMT	Leestempo	Sp_ww
IP	0,57	0,70	0,61	0,28	0,36	0,54	0,52
NW	0,58	0,96	0,64	0,53	0,79	0,87	0,68
GR	0,49	0,51	0,56	0,32	0,29	0,40	0,64
WW	0,35	0,66	0,49	0,32	0,51	0,64	0,66

E7	Rek	Sp-nw	Blz	Avi	DMT	Leestempo	Sp_ww	Wst
IP	0,54	0,61	0,59	0,36	0,34	0,44	0,52	0,62
NW	0,54	0,83	0,45	0,33	0,46	0,67	0,52	0,68
GR	0,54	0,49	0,51	0,41	0,34	0,45	0,53	0,57
WW	0,39	0,64	0,23	0,18	0,36	0,54	0,62	0,58

M8	Rek	Sp-nw	Blz	Avi	DMT	Leestempo	Sp_ww	Wst
IP	0,55	0,52	0,66	0,27	0,35	0,54	0,49	0,57
NW	0,46	0,74	0,53	0,16	0,50	0,60	0,47	0,58
GR	0,49	0,44	0,57	0,28	0,33	0,46	0,54	0,60
WW	0,44	0,54	0,52	0,29	0,31	0,58	0,63	0,59

*Deze correlaties zijn gecorrigeerd voor attenuatie

Uit de tabellen blijkt inderdaad dat de correlatie tussen enerzijds spelling van de tweede generatie en anderzijds spelling niet-werkwoorden als deelvaardigheid van taalverzorging heel hoog is. De correlaties van spelling niet-werkwoorden liggen allemaal boven de 0,70 met een uitschieter (0,96) bij M7. Die hoge samenhang is geheel naar verwachting, het gaat hier immers om dezelfde vaardigheid. Voor spelling werkwoorden liggen de waarden tussen 0,62 en 0,66. Met name bij E7 en M8 vallen de waarden wat lager uit, maar de correlatie is wel hoger dan met semantische vaardigheden. Verder is er een redelijke tot hoge samenhang tussen de deelvaardigheden taalverzorging en de vaardigheid technisch lezen, en dan met name Leestempo en (in iets mindere mate) DMT zoals verwacht.
De correlaties met de andere vakgebieden zijn middelmatig hoog. Dat is ook volgens verwachting: de vaardigheden van taalverzorging hebben een eigen structuur, die tot op grote hoogte bepaald wordt door kennis en het kunnen toepassen van regels en conventies. De andere vaardigheden zoals begrijpend lezen en woordenschat zijn meer semantisch van aard en daarmee duidelijk te onderscheiden van de deelvaardigheden van taalverzorging.
De correlatie van de taken Spelling als onderdeel van taalverzorging met de toets Spelling van de tweede generatie is zeer hoog. Verder is Cito ook bezig met een correlatie-onderzoek met betrekking tot de spellingtoetsen van de derde generatie. De eerste cijfers uit dit onderzoek laten voor de toetsen Spelling niet-werkwoorden als onderdeel van Taalverzorging en de toetsen Spelling 3.0 (M6) een correlatie zien van 0.88. Een uitgebreide beschrijving zal in de nog te verschijnen wetenschappelijke verantwoording van de toetsen Spelling 3.0 voor groep 7 en 8 gepubliceerd worden. Samenvattend kan dus gesteld worden dat de

samenhangen van de toetsen Taalverzorging met andere toetsscores conform de verwachtingen zijn. De data geven aan dat er gemeten wordt wat men beoogt te meten, namelijk deelvaardigheden van taalverzorging.

6.2.3.2 Samenhangen tussen deelvaardigheden

Naast een samenhang met andere taaltoetsen, verwachten we ook een zekere samenhang tussen de deelvaardigheden van Taalverzorging. Onze verwachting is dat de samenhangen tussen de deelvaardigheden matig zijn. Als uitgangspunt bij het ontwikkelen van de toetsen Taalverzorging voor groep 6 tot en met 8 is immers de indeling in vier deelgebieden aangehouden, met ieder hun eigen itembank. We beschouwen deze deelgebieden als vier unidimensionale deelvaardigheden die tezamen de vaardigheid ‘taalverzorging’ vormen. Taalverzorging vatten we dus in groep 6 tot en met 8 niet op als één unidimensionale vaardigheid. Andere domeinen zoals Rekenen hanteren een ander uitgangspunt omdat de samenhang tussen de deelvaardigheden veel groter is. Met name in de bovenbouw treffen we bij Rekenen doorgaans correlaties aan van ruim boven 0,90, reden genoeg om Rekenen op te vatten als één unidimensionale vaardigheid. Zoals eerder besproken zijn er een aantal redenen waarom we taalverzorging niet opvatten als een unidimensionale vaardigheid. De deelvaardigheden van taalverzorging worden vooralsnog geïsoleerd aangeboden in het onderwijs. Ze liggen didactisch gezien soms ver van elkaar verwijderd en zijn niet in alle gevallen ondersteunend ten opzichte van elkaar. De sterkste samenhangen verwachten we voor de beide spellingvaardigheden niet-werkwoorden en werkwoorden. Het blijven immers beide aspecten van spellingvaardigheid.

Tenslotte verwachten we dat de matige samenhang tussen de onderdelen zal toenemen van afnamemoment medio groep 6 naar medio groep 8. In groep 6 worden de deelgebieden nog voornamelijk geïsoleerd aangeboden in het onderwijs. Vanaf eind groep 7 is er sprake van een steeds meer geïntegreerd aanbod van het taalverzorgingsonderwijs. De leerlingen gaan de verschillende deelvaardigheden steeds meer integraal toepassen. We verwachten dat deze ontwikkeling enigszins zichtbaar zal zijn in de correlaties.

Tabel 6.5 Correlaties tussen de deelvaardigheden van Taalverzorging onderling*

M6	**GR**	**IP**	
Interpunctie	0,48		
Spelling niet-werkwoorden	0,47	0,57	
E6	**GR**	**IP**	
Interpunctie	0,56		
Spelling niet-werkwoorden	0,56	0,62	
M7	**GR**	**IP**	**NWW**
Interpunctie	0,67		
Spelling niet-werkwoorden	0,63	0,70	
Spelling Werkwoorden	0,49	0,56	0,83
E7	**GR**	**IP**	**NWW**
Interpunctie	0,67		
Spelling niet-werkwoorden	0,67	0,75	
Spelling Werkwoorden	0,61	0,63	0,78
M8	**GR**	**IP**	**NWW**
Interpunctie	0,66		
Spelling niet-werkwoorden	0,61	0,64	
Spelling Werkwoorden	0,69	0,64	0,77

*Deze correlaties zijn gecorrigeerd voor attenuatie

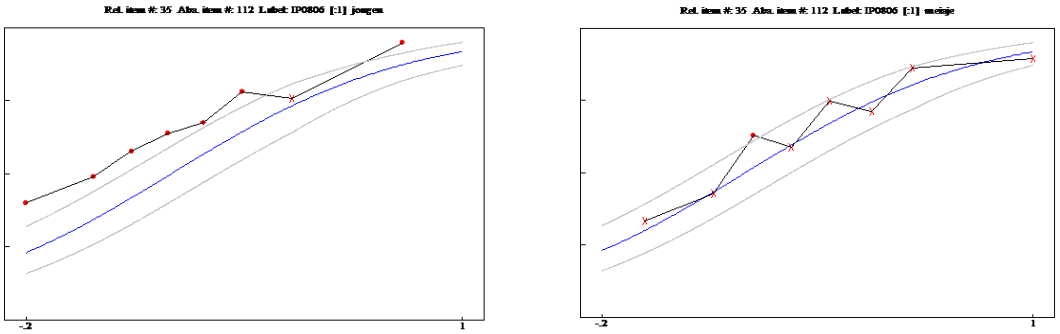
Uit de tabellen blijkt dat de correlatie tussen enerzijds spelling niet-werkwoorden en anderzijds spelling werkwoorden inderdaad heel hoog is. De correlaties van spelling niet-werkwoorden liggen allemaal ruimschoots boven de 0,75 met een uitschieter (0,83) bij M7 waar de waarde 0,83 is. Die hoge samenhang is geheel naar verwachting, het gaat hier immers om dezelfde technische vaardigheid waarbij het woordbeeld herkend moet worden. De samenhang tussen interpunctie en spelling niet-werkwoorden is aanvankelijk erg laag en neemt sterk toe vanaf M6 (0,57) naar E7 (0,75) maar zwakt daarna weer af (op M8). De correlatie tussen interpunctie en grammatica neemt ook duidelijk toe vanaf M6 (0,48) naar M8 (0,66). Hetzelfde geldt voor grammatica en spelling werkwoorden: van M7 (0,49) naar M8 (0,69).
Dit is geheel volgens verwachting: de deelvaardigheden van taalverzorging zijn in groep 6 aparte vaardigheden. Het blijkt echter dat de samenhang toeneemt doordat de deelvaardigheden steeds beter geïntegreerd in het leerproces worden naarmate dat proces vordert. Taalverzorging neigt zich langzamerhand te ontwikkelen tot een meer unidimensionale vaardigheid.
Samenvattend kan dus gesteld worden dat de correlaties tussen de deelvaardigheden bij de toetsen Taalverzorging conform de verwachtingen zijn. Ze bevestigen ten aanzien van het begrip taalvaardigheid dat er (nog) geen sprake is van een echt unidimensionale vaardigheid.

6.3 Itembias

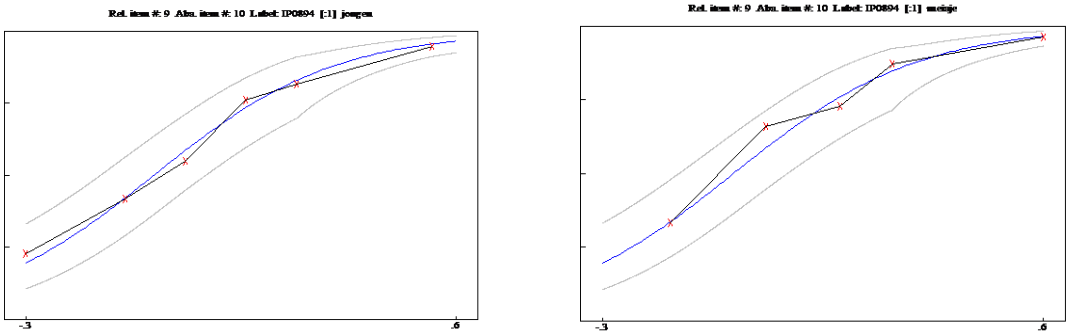
Er is onderzoek uitgevoerd naar differentieel itemfunctioneren (*Differential Item Functioning,* DIF, ook wel itembias genoemd) met betrekking tot afnamemoment en sekse. Voor alle toetsopgaven zijn geobserveerde

en verwachte scores voor zowel jongens als meisjes op de verschillende afnamemomenten in verschillende scoregroepen berekend. Vervolgens is hier een S-statistiek voor berekend, analoog aan hoe dit gebeurt tijdens de kalibratie (zie hoofdstuk 4).
Het onderzoek met betrekking tot DIF naar sekse per afnamemoment liet bij slechts één item significante verschillen zien op het 1%-niveau. In figuur 6.1a is voor dit item Interpunctie M7/E7de toetsing grafisch weergegeven. Verder hebben we ook representatieve items grafisch weergegeven.

Figuur 6.1a S-toets voor een opgave met DIF uit de M7/E7 toets Interpunctie



Figuur 6.1b S-toets voor een representatieve opgave uit de M6/E6 toets zonder DIF



6.4 Verschillen tussen relevante subgroepen

In het kader van het normering- en valideringsonderzoek zijn er twee achtergrondkenmerken waarvoor we graag aandacht vragen. Op de eerste plaats betreft dat de vaardigheidsgroei die met behulp van de toetsen M6/E6, M7/E7 en M8 is vast te stellen via de afnamemomenten M6, E6, M7, E7 en M8. Informatie over de vaardigheidsgroei is al eerder gepresenteerd in de vorm van gemiddelden in tabel 2.1. Op de tweede plaats is ten aanzien van het verschil tussen jongens en meisjes bekend dat bij ‘talige vaardigheden’ en toetsen meisjes over het algemeen in het voordeel zijn (Inspectie van het onderwijs, 2011). Specifiek voor spelling scoren meisjes doorgaans hoger dan jongens op spellingtoetsen (vergelijk bijvoorbeeld van Til, van Weerden, Hemker & Keune, 2014). In tabel 6.6a is te zien dat deze verwachtingen omtrent sekseverschillen opnieuw worden bevestigd.

Tabel 6.6a Effectgroottes voor verschillen tussen jongens en meisjes voor de toetsen Taalverzorging groep 6 tot en met 8 voor alle deelgebieden en alle jaargroepen

	M6	E6	M7	E7	M8
IP	0,412	0,369	0,479	0,451	0,322
NW	0,199	0,137	0,282	0,412	0,274
GR	0,264	0,324	0,309	0,346	0,226
WW			0,282	0,301	0,396

In de tabel zijn de verschillen tussen jongens en meisjes per afnamemoment uitgedrukt in effectgroottes. Positieve waarden weerspiegelen een verschil in gemiddelde ten voordele van meisjes. De effectgroottes variëren van 0,137 (bij Spelling niet-werkwoorden op afnamemoment E6, een verwaarloosbaar effect dus) tot 0,479 (bij Interpunctie op afnamemoment M7, een middelmatig effect). Over het algemeen gaat het, in overeenstemming met de verwachtingen, over kleine tot middelmatige verschillen in het voordeel van meisjes.

In tabel 6.6b en 6.6c is voor jongens en meisjes afzonderlijk de gemiddelde vaardigheidsgroei tussen alle afnamemomenten in termen van effectgroottes in beeld gebracht. Het is duidelijk dat er over het algemeen sprake is van een lichte groei. De effectgroottes zijn niet groter 0,631 (voor meisjes) en 0,573 (voor jongens). De effecten variëren van verwaarloosbaar tot middelmatig. In sommige gevallen is het verschil dermate klein dat het voor een van beide geslachten resulteert in een negatieve waarde (vergelijk Grammatica tussen E7 en M8: verwaarloosbare effecten voor meisjes -0,045 en voor jongens +0,087. Ook Interpunctie tussen E6 en M7: verwaarloosbare effecten voor meisjes +0,015 en voor jongens -0,036). Daarbij valt te constateren dat de effecten voor jongens en meisjes in ongeveer dezelfde intervallen plaatsvinden. Jongens en meisjes laten dus niet op verschillende momenten relatief grote perioden van groei, dan wel stilstand zien; het patroon in de effecten is opvallend consistent.
Dit alles sluit aan bij onze verwachtingen dat er voor zowel jongens als meisjes sprake van een bescheiden groei. Om die reden is ook gekozen voor slechts één toets per leerjaar, waarbij we ervan uitgaan dat het volstaat om slechts een maal per leerjaar de vaardigheidsgroei te registreren en te volgen.

Tabel 6.6b Effectgroottes voor groei meisjes voor de toetsen Taalverzorging groep 6 tot en met 8 voor alle deelgebieden en tussen alle afnamemomenten

	M6-E6	E6-M7	M7-E7	E7-M8
IP	0,463	0,015	0,182	0,245
NW	0,394	0,316	0,304	0,040
GR	0,631	0,122	0,425	-0,045
WW			0,430	0,257

Tabel 6.6c Effectgroottes voor groei jongens voor de toetsen Taalverzorging groep 6 tot en met 8 voor alle deelgebieden en alle jaargroepen

	M6-E6	E6-M7	M7-E7	E7-M8
IP	0,384	-0,036	0,188	0,347
NW	0,401	0,154	0,152	0,210
GR	0,573	0,167	0,326	0,087
WW			0,376	0,159

De resultaten van deze analyses geven aan dat de toetsen Taalverzorging voor groep 6 tot en met 8 prima passen binnen de reeks van toetsen in het Cito Volgsysteem primair en speciaal onderwijs die bedoeld zijn om de vaardigheden van de verschillende taalverzorgingsaspecten in kaart te brengen en te volgen.

7 Samenvatting

In dit samenvattende hoofdstuk geven we kort weer wat in de voorafgaande hoofdstukken is besproken. De LVS-toetsen Taalverzorging voor groep 6 tot en met 8 uit het Cito Volgsysteem primair en speciaal onderwijs vormen een hulpmiddel om het vierde domein uit het Referentiekader Taal en Rekenen (Expertgroep Doorlopende Leerlijnen Taal en rekenen, 2009a) te kunnen meten: het domein Begrippenlijst en taalverzorging. De toetsen in het toetspakket Taalverzorging van het Cito Volgsysteem primair en speciaal onderwijs kunnen worden gebruikt om vast te stellen hoe goed een leerling in het primair en speciaal onderwijs de juiste spelling- en interpunctieregels kan toepassen en hoe de grammaticale kennis van de leerling zich in de loop van de jaren ontwikkelt.

We beschreven in hoofdstuk 2 dat de inhoud van de toetsen aansluit bij het referentiekader Nederlandse taal, de kerndoelen Nederlandse taal, de Leerstoflijnen begrippenlijst en taalverzorging en recente publicaties over taalverzorging. Deze bronnen vormden een adequate basis voor de domeinbeschrijving van de toetsen Taalverzorging. In de domeinbeschrijving legden we uit welke aspecten en principes een rol spelen bij het verzorgen van taal en beschreven we de ontwikkeling van de vier deelvaardigheden. Daarnaast beschreven we de opgavenbanken die gebruikt worden voor de toetsen van het Cito Volgsysteem voor primair en speciaal onderwijs en lichtten we toe dat de deelvaardigheden van taalverzorging, te weten spelling niet-werkwoorden, spelling werkwoorden, interpunctie en grammatica ieder afzonderlijk kunnen worden opgevat als een unidimensionaal continuüm. Verder werd in hoofdstuk 2 het gehanteerde meetmodel beschreven, dat gebaseerd is op de itemresponstheorie.
In aansluiting op deze theoretische uitgangspunten is in hoofdstuk 3 de domeinbeschrijving voor de toetsen Taalverzorging verder uitgewerkt en verantwoord. De constructie van de opgaven is afgeleid van de beschreven domeinindeling en de definitieve selectie van opgaven in de toetsen is gebaseerd op een gewenste verdeling van opgaven binnen en over de verschillende taalverzorgingscategorieën. Ook is in dit hoofdstuk verslag gedaan van de itemconstructie, de opzet van de proeftoetsingen en de normerings-onderzoeken en de samenstelling van de definitieve toetsen. Ten slotte bevat hoofdstuk 3 een beschrijving van enkele psychometrische kenmerken.
In hoofdstuk 4 rapporteerden we over de kalibratie en normering. We beschreven de opzet van en de gevolgde stappen bij de kalibratie en de toetsing van het gehanteerde IRT-model. Uit de resultaten van de S-toetsen op het niveau van de individuele toetsitems, de analyses in termen van R1c en de zogenoemde constante ‘c’ trokken we de conclusie dat de kalibratie geslaagd is. Dit betekent dat de toetsitems succesvol konden worden geschaald en dat het functioneren van leerlingen op de toetsen terug te voeren is op vier unidimensionale concepten: vaardigheid spelling niet-werkwoorden, spelling werkwoorden, interpunctie en grammatica. Ook is verantwoord hoe de dataverzamelingsdesigns voor de normeringsonderzoeken zijn opgezet. Vervolgens werd aangetoond dat de normeringssteekproeven op basis van de variabelen regio, urbanisatiegraad, schooltype en sekse een bevredigende afspiegeling vormen van de populatie. We betoogden dat de gekozen aanpak de best mogelijke garantie vormt voor een adequate normering. In de laatste paragraaf van hoofdstuk 4 presenteerden we de normeringsresultaten en gaven we aan met welke schaalscores de grenzen van de niveau-indelingen samenvallen.
In hoofdstuk 5 stond de betrouwbaarheid van de toets centraal. Wanneer de toetsen Taalverzorging als één geheel wordt gezien (dat is bijvoorbeeld het geval als het referentieniveau wordt bepaald), is de betrouwbaarheid van de volledige toets Taalverzorging met 0,90 als goed te interpreteren. Per deelgebied variëren de betrouwbaarheidscoëfficiënten. De betrouwbaarheid op alle normeringsmomenten voor het deelgebied interpunctie is goed (variërend van 0,81 tot 0,86), voor spelling niet-werkwoorden is de betrouwbaarheid voldoende (variërend van 0,76 tot 0,79), voor grammatica goed (variërend van 0,81 tot 0,86) en voor spelling werkwoorden is alleen het normeringsmoment M8 voldoende (0,70). Voor M7 en E7 is de betrouwbaarheid niet voldoende gebleken en voldoet de deeltoets niet aan de COTAN-norm.
De verklaring hiervoor hebben we uitvoerig beschreven in hoofdstuk 5. In verband met de lage betrouwbaarheid van de toetsen spelling werkwoorden M7/E7 raden we leerkrachten in groep 7 aan bij de interpretatie van de behaalde vaardigheidsscores voorzichtigheid te betrachten. Als er belangrijke

beslissingen verbonden moeten worden aan de scores voor de deeltoets spelling werkwoorden M7/E7, adviseren wij leerkrachten gebruik te maken van de toetsen Cito Spelling werkwoorden voor groep 7 (2e generatie) en de in 2017 te verschijnen toetsen Cito Spelling werkwoorden 3.0 voor groep 7.

Verder zijn in dit hoofdstuk betrouwbaarheidstabellen opgenomen die de betekenis van de meetnauwkeurigheid voor de beslissingen die met de toetsen genomen worden, laten zien. Daarnaast gaven we inzicht in de lokale betrouwbaarheid van de deelvaardigheden op de verschillende meetmomenten: de meetfout blijkt het kleinst te zijn in de lagere en gemiddelde vaardigheidsregionen. Over het algemeen is er sprake van een kleine meetfout voor de range waarin de vaardigheidsverdelingen op het betreffende afnamemoment grotendeels liggen.
Ook hebben we de lokale betrouwbaarheid rondom de gerapporteerde referentieniveaus inzichtelijk gemaakt. De meetfout is klein voor de range waarin de cesuren liggen die de referentieniveaus representeren.
In het laatste hoofdstuk, hoofdstuk 6, stelden we de inhoudsvaliditeit en de begripsvaliditeit van de toetsen aan de orde. De *inhoudsvaliditeit* werd aangetoond door te verwijzen naar de gehanteerde uitgangspunten en bronnen, de analyse van lesmethoden en domeinbeschrijving, de inhoudelijke verantwoording van de taalverzorgingscategorieën, constructieprocedures en itemselectie op basis van empirisch onderzoek (zie hierboven). Een eerste belangrijke aanwijzing voor de *begripsvaliditeit* is te vinden in het unidimensionale karakter van de deeltoetsen, zoals dat in hoofdstuk 4 is aangetoond. De bevindingen met betrekking tot het welslagen van de kalibraties laten zien dat per deeltoets alle opgaven terug te voeren zijn op dezelfde latente trek (vaardigheid). Daarmee wordt voldaan aan een, in onze ogen fundamentele voorwaarde voor begripsvaliditeit.
Een belangrijke aanwijzing voor de convergente en discriminerende validiteit is af te leiden uit de intercorrelaties tussen de toetsen Taalverzorging en andere toetsen uit het Cito Volgsysteem primair onderwijs. Uit deze gegevens blijkt dat de scores op de toetsen Taalverzorging sterk samenhangen met scores op meer technische taalvaardigheidsonderdelen, zoals spelling en technisch lezen (leestempo), en minder met scores op andere, meer semantische onderdelen van taalvaardigheid, zoals begrijpend lezen.
We hebben verder gekeken naar de samenhang tussen de verschillende deelvaardigheden van Taalverzorging. Op basis van de correlaties tussen de deeltoetsen konden we laten zien dat onze beslissing om taalverzorging niet op te vatten als een unidimensioneel concept correct is geweest. Aanvankelijk (in groep 6) is de samenhang matig maar het is gebleken dat de samenhang toeneemt doordat de deelvaardigheden in het leerproces steeds beter geïntegreerd raken naarmate dat proces vordert. Taalverzorging neigt zich langzamerhand te ontwikkelen tot een meer unidimensionele vaardigheid, al lijkt het laatste meetmoment voor deze reeks toetsen (M8) nog net te vroeg te komen om van een geïntegreerde unidimensionale vaardigheid te kunnen spreken als basis voor de toets(en).
Uit de resultaten van de kalibratieanalyses is al af te leiden dat de kwaliteit van de items hoog is. Dit wordt bevestigd door de ‘klassieke’ itemparameters. DIF-onderzoek toont daarnaast aan dat er bij slechts één item (in de toets M7/E7) sprake is van differentieel functioneren met betrekking tot sekse. De gemiddelde scores op Taalverzorging naar sekse laten zien dat meisjes als groep iets hoger scoren. Deze bevinding sluit aan bij het gegeven dat meisjes bij talige toetsonderdelen over het algemeen enigszins in het voordeel zijn. De verschillen zijn echter klein.
Als laatste werden verschillen tussen relevante subgroepen (naar leeftijd, sekse en leerlinggewicht) gepresenteerd. De resultaten bleken aan te sluiten bij de verwachtingen die op grond van theoretische inzichten en eerder onderzoek konden worden geformuleerd en vormen daarmee extra ondersteuning voor de validiteit van de toetsen.
Op basis van deze analyses, die licht werpen op diverse aspecten van validiteit, kunnen we concluderen dat de LVS-toetsen Taalverzorging naast inhoudsvalide ook begripsvalide instrumenten zijn om de vier deelvaardigheden die samen taalverzorging vormen, te beschrijven en te volgen.

8 Literatuur

Beek, A. van der & Paus, H. (2011). *Leerstoflijnen begrippenlijst en taalverzorging beschreven. Uitwerking van het referentiekader Nederlandse taal voor het domein begrippenlijst en taalverzorging op de basisschool.* Enschede: SLO.

Bloom, H.S., Bos, J.M., & Lee, S. (1999). Using cluster random assignment to measure program impacts. *Evaluation Review, 23*, 445-469.

Bloom, H.S., Richburg-Hayes, L., & Black, A.R. (2007). Using Covariates to Improve Precision for Studies that Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis, 29*, 30-59.

Bon, W.H.J. van (1993). *Spellingproblemen: Theorie en praktijk*. Rotterdam: Lemniscaat.

Bonset, H., & M. Hoogeveen (2009). *Spelling in het basisonderwijs. Een inventarisatie van empirisch onderzoek*. Enschede: SLO

Boxtel, H. van & B.T. Hemker (2009). *Wetenschappelijke verantwoording van de Intelligentietest Eindtoets Basisonderwijs*. Arnhem: Cito.

Cito (2014). *Cito Volgsysteem primair en speciaal onderwijs. Spelling 3.0 Groep 4*. Arnhem: Cito.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Erlbaum.

Eggen, T.J.H.M. (1993). Itemresponstheorie en onvolledige gegevens. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk*. (pp. 239-284). Arnhem: Cito.

Engelen, R.J.H. en Eggen, T.J.H.M., (1993). Equivaleren. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk*. (pp. 309-348). Arnhem: Cito.

Evers, A., Lucassen, W., Meijer, R. & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests.* Amsterdam: NIP/COTAN.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008a). *Over de drempels met taal en rekenen. Hoofdrapport van de Expertgroep Doorlopende Leerlijnen Taal en Rekenen*. Enschede: Expertgroep doorlopende leerlijnen Taal en Rekenen.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008b). *Over de drempels met taal. De niveaus voor de taalvaardigheid*. Enschede: Expertgroep doorlopende leerlijnen Taal en Rekenen.

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

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2009b). *Een nadere beschouwing. Over de drempels met taal en rekenen*. Enschede: Expertgroep doorlopende leerlijnen Taal en Rekenen.

Gijsel, M., Scheltinga, F., Druenen, M. van & Verhoeven, L. (2011a). *Protocol Leesproblemen en Dyslexie voor groep 3*. Nijmegen: Expertisecentrum Nederlands.

Gijsel, M., Scheltinga, F., Druenen, M. van & Verhoeven, L. (2011b). *Protocol Leesproblemen en Dyslexie voor groep 4*. Nijmegen: Expertisecentrum Nederlands.

Glas, C.A.W. & N.D. Verhelst (1993). Een overzicht van itemresponsmodellen. In: T.J.H.M. Eggen & P.F.

Sanders (red.). *Psychometrie in de praktijk*. (pp. 179-238). Arnhem: Cito.

Hedges, L.V., & Hedberg, E.C. (2007). Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis, 29*, 60-87.

Hemker, B.T., J. Kordes & J.J. van Weerden (2011). *Peiling van de rekenvaardigheid en de taalvaardigheid in jaargroep 8 en jaargroep 4 in 2010 - Jaarlijks Peilingsonderzoek van het Onderwijsniveau*. Arnhem: Cito.

Huizenga, H. (2010). *Taal & didactiek. Spelling* (4e herziene druk). Groningen: Noordhoff Uitgevers.

Keuning, J. (2011). *Normeren op schoolniveau met Cito dataretour*. Arnhem: Cito.

Keuning, J., Boxtel, H. van, Lansink, N., Visser, J., Weekers, A. & Engelen, R. (2015). *Actualiteit en kwaliteit van normen. Een werkwijze voor het normeren van een leerlingvolgsysteem.* Arnhem: Cito.

Kleijnen, R. (1997). *Strategieën van zwakke lezers en spellers in het voortgezet onderwijs. Dissertatie Vrije Universiteit*. Lisse: Swets en Zeitlinger.

Kleijnen, R. (2004). *Hardnekkige spellingfouten. Een taalkundige analyse*. Lisse: Harcourt Book publishers.

Kuhlemeier, H., Til, A. van, Hemker, B., Klijn, W. de & Feenstra, H. (2013). *Balans van de schrijfvaardigheid in het basis- en speciaal basisonderwijs 2. Uitkomsten van de peiling in 2009 in groep 5, groep 8 en de eindgroep van het SBO.* PPON-reeks nummer 53. Arnhem: Cito.

Kuiken, F. & Droge, S. (2010). *Woordenlijst Amsterdamse Kinderen*. Amsterdam: Universiteit van Amsterdam.

Ministerie van Onderwijs, Cultuur en Wetenschappen (2006). *Kerndoelen primair onderwijs.* Den Haag: MinOCW.

Nederlandse Taalunie (2009). *Technische handleiding. Regels voor de officiële spelling van het Nederlands*. Geraadpleegd op 19 juni 2014 via http://taalunieversum.org/inhoud/spelling-meerhulpmiddelen/-technische-handleiding.

Pilliner, A. (1969). *Estimation of number of grades to be awarded in an examination by consideration of its reliability coefficient*. Edinburgh: The Godfrey Thomson Unit for Educational Research.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Schijf, G.M. (2009). *Lees- en spellingvaardigheden van brugklassers* (proefschrift). Amsterdam: SCOKohnstamm Instituut, Universiteit van Amsterdam.

Schochet, P. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics, 33*, 62-87.

Schryver, J. de & A. Neijt (2005). *Handboek Spelling* (5e herziene druk). Mechelen: Wolters Plantyn.

Snijders, T.A.B. & Bosker, R.J. (1993). *Standard errors and sample sizes for two-level research*. Journal of Educational Statistics, 18, 237-260.

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid: de ontwikkeling van een domeingericht meetinstrument.* Enschede: Universiteit Twente.

Bijlagen

Bijlage 1 Referentieniveaus Begrippenlijst en Taalverzorging

4. Begrippenlijst en Taalverzorging

4.1. Begrippenlijst

Om te spreken over taal en taalverschijnselen is een beperkt aantal begrippen noodzakelijk. De meeste daarvan zijn aan het einde van het basisonderwijs wel aan de orde geweest (1F). Kennis van deze begrippen bevordert het gesprek binnen en buiten het taalonderwijs over taal en taalverschijnselen: het gaat erom dat docenten (en leerlingen) bepaalde verschijnselen kunnen benoemen in contextrijke taalsituaties. Dat wil zeggen dat docenten deze termen moeten kunnen gebruiken in hun onderwijs in de vaardigheidsdomeinen.

Regels

Regel voor overeenkomst in getal (onderwerp-persoonsvorm; referent-verwijswoord) en geslacht (referent-verwijswoord).

4.2. Taalverzorging

De vereiste kwaliteit van productief taalgebruik (spreken, schrijven) wordt steeds aangeduid bij de kenmerken van de taakuitvoering in die domeinen.
In dit domein van taalverzorging gaat het alleen om kennis van regels en begrippen die ten dienste staan van correct taalgebruik. Bij de niveaubepaling is steeds uitgegaan van volledige beheersing, dat wil zeggen, vrijwel automatische beheersing en bij uitzondering terugvallend op regelkennis in taalproductie, zoals in de domeinen schrijven en spreken beschreven.

Tabel 1: Niveaubeschrijvingen Begrippen Taal

	1F	2F
Leestekens	Dubbele punt, punt, komma, puntkomma, uitroepteken, vraagteken, aanhalingsteken.	Trema, accent.
Woordsoorten	Zelfstandig naamwoord, werkwoord (klankvast, klankveranderend (zwak, sterk)), bijvoeglijk naamwoord.	
Grammaticale kennis	Onderwerp, lijdend voorwerp, hoofdzin, bijzin, gezegde, persoonsvorm.	Lijdende en bedrijvende vorm, vragende vorm.
Tekstkennis	Standpunt, argument, feit, mening, tekstsoort en gesprekvormen, paragraaf.	Aanduidingen voor tekstsoorten en genres (ook: aanduidingen voor gespreksvormen), hoofdgedachte (van tekst), tekstthema. Metatalige vormen: Woorden, zinnen en tekstfragmenten die informatie geven over de rest van de tekst (zoals signaalwoorden, prospectieve en retrospectieve tekstelementen in inleiding, samenvattende zin aan slot).
Stilistiek en semantiek	Betekenis, symbool, synoniem, context, letterlijk, figuurlijk, uitdrukking, spreekwoord, gezegde, moedertaal, tweede taal, vreemde taal, standaardtaal, dialect, meertalig, formeel en informeel taalgebruik, leenwoord.	Homoniem, homofoon, vakjargon, stilistische adequaatheid (publiekgericht), presentatiekenmerken (van mondelinge en schriftelijke tekst).
Morfologie	Woordvorm, woorddeel, samengesteld, voorvoegsel, achtervoegsel, lettergreep. Getal (meervoud/enkelvoud), tijd (tegenwoordig, verleden, voltooid, onvoltooid). Verkleinwoord, verschijningsvormen werkwoord (stam, infinitief, bijvoeglijk naamwoord).	
Opmaak	Bladzijde, woord, zin, hoofdletter, uitspraak, titel, hoofdstuk, regel, lettertype, alinea, kopje.	
Klanken	Articulatie, klemtoon, intonatie, spreekpauze.	

Grammaticale begrippen voor werkwoordspelling

1. Werkwoord;
2. Tijd van het werkwoord (tegenwoordig en verleden, onvoltooid en voltooid);
3. Getal: meervoud, enkelvoud;
4. Eerste, tweede en derde persoon;
5. Persoonsvorm;
6. Voltooid deelwoord;
7. Stam van het werkwoord;
8. Hele werkwoord (infinitief);
9. Onderwerp;
10. Zwakke en sterke werkwoorden;
11. Werkwoordelijk gezegde.

Regelkennis en toepassing in oefentaken gaat aan die beheersing vooraf. De niveaus geven een eindpunt aan: het verwerven van de regels tot een vrijwel automatische beheersing vergt veel leertijd. Het geleerde moet voortdurend in onderhoud zijn. Dat kan betekenen dat van tijd tot tijd nieuwe instructie en oefening gegeven moeten worden (opfrissen) en dat er zorgvuldig feedback gegeven dient te worden op schrijf- en spreekproduct- en door alle bij het onderwijs betrokkenen, docenten Nederlands en docenten van andere vakken.

4.3. Niveaubeschrijvingen

4.3.1. Spelling

Categorieën

Deze paragraaf bevat de categorieën van spellingsproblemen en -regels. De basis voor de spelling is kennis van de beschaafde uitspraak van het Nederlands ('klankzuiver'):

1. Klankzuivere woorden (wil, dier, maat, daar, moet, wesp, kalf etc.): woorden die in een standaard Nederlandse uitspraak geen alternatieve spelling toelaten.
2. Klankambigue woorden: woorden die indien de klank gevolgd wordt fout gespeld zullen worden. Het gaat om algemene regels en dialectische bijzonderheden. Het zijn fouten die in de ene regio vaker zullen voorkomen dan in een andere: bodum (bodem), enugu (enige), flakbij (vlakbij), prijsen (prijzen), prongeluk (per ongeluk), srijf (schrijf), teminste; tuminste, tuminstu (tenminste), trugbetalen (terugbetalen).
3. Spelambigue woorden zoals mouwen (mauwen), klijn (klein), dagt (dacht), antwoort (antwoord), direkt (direct). Het zijn woorden die op twee manieren gespeld kunnen worden, omdat de klank geen uitsluitsel geeft. Twee lettertekens representeren één klank (au/ou, d/t, ei/ij, ch/g, c/k).

Regels voor lettergreepgrenzen

4. Regels voor verdubbeling en verenkeling op lettergeepgrenzen: ontsmetting, nummer, verstoppen, liggen, lopen, oversteken, haren.
5. Afbreekregels (ge-trokken; getrok-ken, get-rokken, getrokk-en), als een samenspel van morfologische en spellingregels.

Regels voor woordgrenzen

6. Aaneen- en losschrijven van woorden (autoweg, kwijtraakte, voor altijd).

Morfologische spelling:

7. Regel van gelijkvormigheid bij assimilatie: *zakdoek in plaats van zaddoek*.
8. Meervoudsvorming
   - 8.1. –s na medeklinker, -a, -o, -u, -y, -e: (a) fuchsia's, (b) cafés, (c) garages, meisjes
   - 8.2. –en (a) zonder en (b) met verdubbeling): *latten (zelfstandig naamwoord), laten (werkwoord).*
9. Vorming van bijvoeglijk naamwoord
   - 9.1. --e (bij zelfstandig naamwoord in enkelvoud als meervoud), met mogelijk toepassing van andere regels (verenkeling/verdubbeling op lettergreepgrenzen). Ook bij bijvoeglijke naamwoorden afgeleid van werkwoorden
   - 9.2. Stoffelijke bijvoeglijke naamwoorden op –en: *gouden, zilveren* (zowel bij zelfstandig naamwoord in enkelvoud als meervoud).
10. Vorming van verkleinwoord
    - 10.1. Basis+dimunitief
    - 10.2. Uitzondering op verenkelings/verdubbelingsregel: verkleinwoord na open klinker: *chocolaatje, cafeetje, parapluutje.*
11. Schrijfwijze van achtervoegsels (-heid, -lijk).
12. 's en –s: 's nachts, 's Nachts (begin van een zin).
13. Meervouds –n bij zelfstandig en bijvoeglijk gebruikte verwijzingen naar personen/niet personen: *alle, vele, weinige, maar ook allen, weinigen, velen etc.*

Regels voor de werkwoordspelling

14. Persoonsvorm
    - 14.1. tegenwoordige tijd van werkwoorden met stam op –d
      - 14.1.1. enkelvoud: word(t)
        - eerste persoon stellend en vragend (ik word/word ik)
        - tweede persoon stellend en vragend (jij wordt/word jij)
        - derde persoon enkelvoud stellend en vragend (hij wordt/wordt hij)
          - wordt je broer, wordt jou de toegang ontzegd
          - derde persoon, enkelvoud stellend en vragend bij werkwoorden met prefix (kans op verwarring met woordbeeld van voltooid deelwoord): hij beoordeelt (niet: beoordeeld)
      - 14.1.2. meervoud: worden, laten
    - 14.2. verleden tijd van zwakke werkwoorden met stam op –d of –t: (morfologische regel leidt tot verdubbeling van d/t, hoewel fonetisch niet nodig) *antwoordde*
    - 14.3. verleden tijd van sterke werkwoorden met stam op –d of –t
      - enkelvoud: *werd, liet*
      - meervoud: *werden.*
15. Infinitief
    'Gewone' werkwoorden met stam op –d of –t: *worden, laten*
    - 15.1. Werkwoorden met stam op –d en –t die in de verleden tijd dd/tt krijgen: *vergoeden, verplichten* (verwisseling woordbeelden)
    - 15.2. Als 15.1, in bijvoeglijke bepalingen, in een omgeving met verleden tijd *('de te verlichten straten waren niet afgesloten').*
16. Voltooid deelwoord
    - 16.1. (per prefix), met kans op verwarring met woordbeeld persoonsvorm
      - op –d: gebeurd, beoordeeld
      - op –d: na een 'valse' f (stam op v): *geverfd*
      - op -d, na een 'valse' s (stam op z): *verhuisd*
    - 16.2. op –den of –ten: *geladen, gelaten*
      - in de omgeving van meervoud (de geladen wagens)
      - in de omgeving van enkelvoud *(de geladen wagen)*
    - 16.3. op –d of –t, gebruikt als bijvoeglijk naamwoord: *geparkeerde, geraakte, beschutte*
      - in de omgeving van enkelvoud/meervoud (de beschutte tuin/tuinen (bijvoeglijk naamwoord buigt niet met getal mee)
      - in de omgeving van tegenwoordige/verdelen tijd: hij zag/zij ziet verlichte straten.

Overige regels

17. Schrijfwijze van tussenklanken –s en –e(n).
18. Gebruik van trema en koppelteken.

4.3.2. Leestekens

1. Hoofdletters en punten bij zinsmarkering.
2. Vraagtekens, uitroeptekens en aanhalingstekens.
3. Hoofdletters bij eigennaam en directe rede.
4. Komma's, dubbele punt.

4.4. Moeilijkheid

De moeilijkheid van spelling is op twee manieren te ordenen. Er zijn empirische gegevens over wat leerlingen einde BO kunnen (PPON) en toetsgegevens van brugklasleerlingen*. Dat levert een overzicht van itemmoeilijkheden op, zoals gepresenteerd in het eerste rapport van de Expertgroep (2008). Spellingsproblemen kunnen ook in grotere klassen worden ondergebracht, zoals Schijf (2009) laat zien. Naast een zekere logische opeenvolging van klassen van problemen, speelt ook de frequentie waarin het te spellen woord verschijnt een rol. De 'stomme e' bijvoorbeeld, in 'stomme' wordt in het algemeen pas beheerst na groep 4, maar zeer frequente woorden met een stomme *'e'* worden al in groep 3 goed gespeld.
Als ordening voor de spellingsproblemen gebruiken we een indeling in vijf klassen. Deze indeling wordt gebruikt bij het diagnosticeren van spellingvaardigheid.

1. Alfabetisch: hier gaat het om het volgen van de beschaafde Nederlandse uitspraak: dezelfde klank heeft dezelfde letter. De basiskennis is de klank-tekenkoppeling, ook voor bijvoorbeeld oe, ui. Allofonen (v/f; z/s afwisseling) kunnen hierbij gerekend worden. Eind groep 3 wordt deze categorie beheerst.
2. Orthografisch: hier gaat het om autonome regels over de grens van lettergrepen heen: woorden met sch, ng, nk, aai, ooi, oei, ch(t), -eeuw, -ieuw, -uw, -ee, de ë in ie of ieë, medeklinkerverdubbeling, open lettergrepen, kleefletters behoren tot deze categorie.
3. Morfologisch: alle woorden die gevormd worden door de toevoeging van voor- of achtervoegsels zoals verkleinwoorden (-tje, -pje, -je), meervoudsvorming en achtervoegsels als (-ig, -heid, -teit, -lijk, -aard, -erd, -tie, -iaal/-eaal/-ieel/-ueel, -isch); ook: bijvoeglijk gebruikt voltooid deelwoord. Woorden met 's als meervoud.
   Alle woorden die gevormd worden door samenstellingen (assimilatieverschijnselen: voortdurend).
4. Morfologisch, met gebruikmaking van syntactische kennis: werkwoordsspelling waarin persoon en getal van het onderwerp leidend is voor de spelling (persoonsvorm), de functie van het werkwoord moet worden bepaald (persoonsvorm, infinitief, voltooid deelwoord). Homofonen zijn hier de moeilijkste problemen (verhuisd/verhuist, beleeft/beleefd): kennis van de functie is hier noodzakelijk.
5. Logografisch: vaststaande combinaties, die als zodanig gekend moeten worden (geen regelvorming): /zj/ geschreven als g (garage), open lettergreep /ie/ geschreven als -i-, woorden op -isch, /sj/ geschreven als –ch-, /oo/ geschreven als -au- of -ou-, /s/ geschreven als -c- voor i, ie en e; /ks/ geschreven als –x-, /oe/ geschreven als -ou-, woorden met -aise, -aire, /sj/ geschreven als –ci-, /ie/ geschreven als -y-, leenwoorden (team, jam, tram). Woorden met een trema, woorden voorafgegaan door 's.

In schema: zie tabel 2 op pagina 20.

**G.M. Schijf (2009). Lees- en spellingsvaardigheden van brugklassers, diss, Universiteit van Amsterdam*

Bijlage 2 Moeilijkheid van opgaven per jaargroep en taak in de toetsen Taalverzorging voor groep 6, 7 en 8

	M6			E6				
Volgnr	**P-waarde**	**RIT**	**info**	**P-waarde**	**RIT**	**info**	**dsc**	**beta**
1.1	0,849	0,485	2,427	0,902	0,456	1,717	5	-0,227
1.2	0,776	0,410	1,356	0,831	0,400	1,102	3	-0,231
1.3	0,774	0,411	1,364	0,830	0,401	1,110	3	-0,226
1.4	0,734	0,552	3,471	0,817	0,536	2,703	5	-0,034
1.5	0,612	0,453	1,796	0,691	0,458	1,609	3	0,077
1.6	0,588	0,455	1,828	0,670	0,462	1,662	3	0,117
1.7	0,640	0,450	1,749	0,716	0,452	1,541	3	0,030
1.8	0,704	0,561	3,670	0,793	0,549	2,925	5	0,008
1.9	0,631	0,451	1,765	0,708	0,454	1,563	3	0,045
1.10	0,638	0,450	1,753	0,714	0,452	1,547	3	0,034
1.11	0,520	0,456	1,878	0,606	0,468	1,777	3	0,227
1.12	0,583	0,456	1,835	0,665	0,463	1,673	3	0,126
1.13	0,773	0,481	2,227	0,839	0,466	1,737	4	-0,139
1.14	0,500	0,366	0,914	0,565	0,380	0,892	2	0,258
1.15	0,606	0,454	1,804	0,686	0,459	1,622	3	0,087
1.16	0,662	0,517	2,752	0,749	0,513	2,320	4	0,038
1.17	0,628	0,360	0,858	0,686	0,366	0,789	2	-0,027
1.18	0,694	0,510	2,629	0,776	0,503	2,168	4	-0,010
1.19	0,738	0,425	1,493	0,800	0,419	1,242	3	-0,151
1.20	0,716	0,558	3,595	0,802	0,545	2,839	5	-0,008
2.1	0,884	0,438	2,930	0,920	0,422	2,107	6	-0,314
2.2	0,720	0,381	1,633	0,771	0,387	1,423	3	-0,207
2.3	0,584	0,520	4,635	0,674	0,535	4,112	5	0,058
2.4	0,689	0,455	2,870	0,755	0,461	2,455	4	-0,093
2.5	0,734	0,377	1,587	0,783	0,382	1,374	3	-0,232
2.6	0,710	0,449	2,768	0,773	0,454	2,342	4	-0,123
2.7	0,448	0,400	1,972	0,518	0,425	1,954	3	0,226
2.8	0,746	0,490	3,740	0,811	0,489	3,016	5	-0,133
2.9	0,774	0,478	3,482	0,833	0,475	2,762	5	-0,172
2.10	0,442	0,399	1,968	0,513	0,425	1,955	3	0,234
2.11	0,775	0,425	2,386	0,827	0,425	1,951	4	-0,224
2.12	0,634	0,400	1,861	0,694	0,412	1,685	3	-0,059
2.13	0,690	0,389	1,727	0,745	0,397	1,524	3	-0,153
2.14	0,658	0,396	1,809	0,716	0,406	1,620	3	-0,099
2.15	0,762	0,431	2,474	0,816	0,432	2,037	4	-0,202
2.16	0,518	0,406	1,990	0,587	0,426	1,902	3	0,119
2.17	0,707	0,450	2,784	0,771	0,455	2,360	4	-0,119
2.18	0,742	0,492	3,769	0,808	0,491	3,046	5	-0,128
2.19	0,739	0,440	2,610	0,798	0,442	2,175	4	-0,167

2.20	0,854	0,469	3,477	0,899	0,455	2,553	6	-0,258
3.1	0,669	0,496	4,343	0,774	0,532	3,238	5	-0,180
3.2	0,644	0,502	4,477	0,756	0,540	3,397	5	-0,151
3.3	0,599	0,454	3,213	0,708	0,497	2,623	4	-0,119
3.4	0,556	0,556	6,367	0,697	0,603	4,999	6	-0,050
3.5	0,538	0,391	1,998	0,637	0,436	1,780	3	-0,055
3.6	0,537	0,391	1,998	0,636	0,436	1,782	3	-0,054
3.7	0,563	0,512	4,759	0,692	0,559	3,836	5	-0,064
3.8	0,504	0,391	2,008	0,606	0,439	1,833	3	-0,003
3.9	0,639	0,503	4,503	0,751	0,542	3,429	5	-0,145
3.10	0,597	0,554	6,230	0,729	0,595	4,732	6	-0,089
3.11	0,656	0,500	4,415	0,765	0,536	3,322	5	-0,165
3.12	0,567	0,457	3,277	0,681	0,503	2,739	4	-0,079
3.13	0,563	0,556	6,348	0,703	0,601	4,955	6	-0,057
3.14	0,568	0,457	3,275	0,682	0,503	2,734	4	-0,080
3.15	0,606	0,553	6,191	0,735	0,593	4,670	6	-0,098
3.16	0,494	0,310	0,946	0,569	0,349	0,905	2	0,014
3.17	0,557	0,390	1,984	0,653	0,434	1,747	3	-0,083
3.18	0,524	0,391	2,004	0,624	0,438	1,803	3	-0,034
3.19	0,628	0,505	4,550	0,743	0,545	3,492	5	-0,134
3.20	0,659	0,444	3,028	0,756	0,481	2,375	4	-0,195

Bijlage 3 Categorieënoverzichten

Overzicht interpunctiecategorieën in toetsen Taalverzorging groep 6, 7 en 8

Omschrijving Categorie	Afnamemoment		
Zinseindetekens:			
• Hoofdletter als markering zinsgrens	M6/E6	M7/E7	M8
• Punt als markering zinsgrens	M6/E6	M7/E7	M8
• Hoofdletter bij directe rede		M7/E7	M8
• Vraagteken	M6/E6	M7/E7	M8
• Uitroepteken		M7/E7	M8
Zinsgeleders & markeerders van een citaat:			
• Dubbele punt bij opsomming			M8
• Dubbele punt bij conclusie		M7/E7	M8
• Dubbele punt bij directe rede		M7/E7	
• Komma bij opsomming	M6/E6	M7/E7	M8
• Komma tussen twee persoonsvormen			M8
• Aanhalingstekens (bij opening en sluiting van directe rede)		M7/E7	M8

Spellingscategorieën (niet-werkwoorden) in de toetsen Taalverzorging groep 6, 7 en 8

Cat.	Omschrijving			
14	woorden met (-)ei(-) of (-)ij(-)	M6E6	M7E7	
15	woorden eindigend op -d	M6E6		
17	woorden met -au(-), -auw, -ou(-) of -ouw	M6E6		
18	woorden met -ch(t)	M6E6		
19	woorden op -eeuw, -ieuw, -uw	M6E6		
20	woorden met open lettergreep	M6E6	M7E7	M8
21	woorden met gesloten lettergreep	M6E6	M7E7	M8
22	verandering van -f in -v- en -s in -z- bij vervoeging of meervoudsvorming	M6E6		
23	woorden met -em, -elen, -enen of -eren	M6E6		
24	woorden op -ig(e) en -lijk(e)	M6E6		
25	woorden waarin /ie/ geschreven wordt als i	M6E6	M7E7	M8
26	woorden waarin /s/ geschreven wordt als c	M6E6	M7E7	M8
27	woorden waarin /k/ geschreven wordt als c	M6E6	M7E7	M8
28	woorden beginnend met 's of eindigend op 's		M7E7	M8
29	woorden met -tie(-) waarin t klinkt als ts		M7E7	
30	woorden met -heid of -teit		M7E7	
31	leenwoorden waarin /zj/ geschreven wordt als g(e)		M7E7	
32	leenwoorden waarin /sj/ geschreven wordt als ch (nieuw)		M7E7	
33	woorden met -b(-)		M7E7	
34	woorden met (-)y(-)		M7E7	M8
35	woorden met een trema			M8
36	woorden met een koppelteken			M8
37	samenstellingen met tussen -e(n)- en tussen -s-			M8
38	woorden met of zonder een hoofdletter		M7E7	
39	Franse leenwoorden			M8
40	Engelse leenwoorden			M8
41	woorden waarin /t/ geschreven wordt als th			M8
42	woorden met -isch(e)		M7E7	
43	woorden waarin /ks/ geschreven wordt als x			M8
44	verkleinwoorden -aatje, -eetje, -ootje, -uutje en met de uitgang -nkje		M7E7	M8
45	woorden met assimilatieverschijnselen (nieuw)		M7E7	M8
46	woorden op -iaal, -ieel, -ueel, -eaal			M8

Overzicht grammaticacategorieën in toetsen Taalverzorging groep 6, 7 en 8

Omschrijving Categorie	Afnamemoment		
Woordbenoeming en benoeming werkwoorden:			
• Zelfstandig naamwoord	M6/E6	M7/E7	M8
• Bijvoeglijk naamwoord	M6/E6	M7/E7	M8
• Lidwoord	M6/E6	M7/E7	M8
• Infinitief/hele werkwoord	M6/E6	M7/E7	M8
• Voltooid deelwoord		M7/E7	M8
• Tijd van het werkwoord (voltooide & onvoltooide tijd, tegenwoordige & verleden tijd)			M8
Zinsvormen en zinsontleding:			
• Hoofd- en bijzin			M8
• Onderwerp	M6/E6	M7/E7	M8
• Persoonsvorm	M6/E6	M7/E7	M8
• Lijdend voorwerp			M8
• Gezegde (werkwoordelijk)			M8

Spellingscategorieën (werkwoorden) in de toetsen Taalverzorging groep 6, 7 en 8

Cat.	omschrijving			Voorbeelden	Toets M7/E7	Toets M8
1	tijd van nu	1.a	tijd van nu (ott): -t achter stam van zwak ww dat in o.v.t. de uitgang -de(n) krijgt	jij tekent	M7E7	
		1.b	tijd van nu (ott): wel of geen -t achter een stam op -d	ik vind, jij onthoudt, hij verbindt, hij redt	M7E7	M8
		1.c	tijd van nu (ott): bij inversie pv-ond: wel of geen -t achter een stam op -d (vraag of gebiedende wijs)	bind ik? word jij? houdt u? schudt hij?		M8
		1.d	tijd van nu (ott): homofone gevallen: o.t.t.	het gebeurt, hij verdeelt		M8
2	tijd van toen	2.a	tijd van toen (ovt): zwak ww dat in o.v.t. de uitgang -te(n) of -de(n) krijgt	ik bakte, zij tekende, wij hoopten	M7E7	M8
		2.b	tijd van toen (ovt): verdubbeling d of t bij zwak ww met stam op -d of -t	ik raadde, jij stootte, wij landden	M7E7	M8
		2.c	tijd van toen (ovt): geen -t bij sterk ww dat in 2e en 3e persoon eindigt op -d	jij werd, zij hield, hij zond, zij stond	M7E7	M8
		2.d	tijd van toen (ovt): uitgang -sde(n) of -fde(n) bij zwak ww met stam op -z of -v	ik prijsde, hij beefde, jullie verfden	M7E7	
		2.e	tijd van toen (ovt): geen verdubbeling medeklinker bij sterk werkwoord	wij liepen, zij stonden	M7E7	
3	voltooid deelwoord	3.a	voltooid deelwoord: keuze voor eind-d of eind-t bij zwakke werkwoorden met een stam die niet eindigt op -d of -t of -v en -z	geblust, gewerkt	M7E7	M8
		3.b	voltooid deelwoord: homofone gevallen: volt.dw.	is beoordeeld, is verbrand	M7E7	M8
		3.c	zwakke werkwoorden met stam op v en z of d en t	geprijsd; afgemeld	M7E7	M8
4	(on)voltooid deelwoord (bijvoegelijk gebruikt)	4.a	wel of geen -n aan het eind; d of t; woorden met d en t onvoltooid deelwoord bijvoeglijk gebruikt	de gekookte eieren, het gebraden vlees; gegrild/ gegrilt vlees; trillende/ trillendde stem	M7E7	M8
5	infinitief	5.a	infinitief: homofone gevallen	praten/praatten, wieden/ wiedden	M7E7	M8

Bijlage 4 Voorbeeldopgaven

Interpunctie

Voorbeelden van de verschillende opgaventypen. De antwoordsleutel is aangegeven met een asterisk (*).

1 Opgaven met één stamzin waarover een vraag wordt gesteld. De antwoordalternatieven zijn onderdelen van deze zin.

de verkoopster in de kledingwinkel zegt die spijkerbroeken zijn in de aanbieding

Welk onderstreept woord moet met een hoofdletter worden geschreven?

A zegt
B* die
C spijkerbroeken
D aanbieding

2 Opgaven met twee of meer stamzinnen waarover een vraag wordt gesteld. De antwoordalternatieven zijn vier variaties van deze stamzinnen.

In welke zin staat de komma (,) op de juiste plaats?

A je kunt bij chocolade, vaak kiezen tussen puur melk en witte chocolade
B je kunt bij chocolade vaak kiezen tussen, puur melk en witte chocolade
C* je kunt bij chocolade vaak kiezen tussen puur, melk en witte chocolade
D je kunt bij chocolade vaak kiezen tussen puur melk en, witte chocolade

3 Opgaven met meerdere stamzinnen waarover een vraag wordt gesteld. De antwoordalternatieven zijn mogelijke leestekens op een bevraagde plek. De bevraagde plek wordt gedefinieerd door een woord uit de stamzin.

wat zou jij doen als je lesrooster voor de zoveelste keer zou **veranderen** ik ben al heel vaak boos geworden maar dat helpt niet

Welk leesteken moet er worden gezet achter **veranderen**?

A een punt (**.**)
B* een vraagteken (**?**)
C een uitroepteken (**!**)
D een dubbele punt (**:**)

4 Opgaven waarin het aanwezig zijn van een directe rede bevraagd wordt. De antwoordalternatieven zijn een weergave van een gesprekssituatie.

In welke zin moet het onderstreepte stukje tussen aanhalingstekens openen en sluiten ("...") worden gezet?

A Marissa vraagt of wij ook allemaal naar het museum gaan
B Marissa vraagt wie er allemaal naar het museum gaan
C* Marissa vraagt wie gaan er allemaal naar het museum
D Marissa vraagt zich af wie er allemaal naar het museum gaan

Niet-Werkwoorden & Werkwoorden

In welke zin zijn de dikgedrukte woorden **allebei goed** gespeld?

A Een **achtbaan** vind ik wel een beetje **griezeleg**.
B* Een **achtbaan** vind ik wel een beetje **griezelig**.
C Een **agtbaan** vind ik wel een beetje **griezeleg**.
D Een **agtbaan** vind ik wel een beetje **griezelig**.

In welke zin zijn de dikgedrukte woorden **allebei goed** gespeld?

A Sarah **maakd** haar kleren erg vuil omdat ze een kuil **graafd**.
B Sarah **maakd** haar kleren erg vuil omdat ze een kuil **graaft**.
C Sarah **maakt** haar kleren erg vuil omdat ze een kuil **graafd**.
D* Sarah **maakt** haar kleren erg vuil omdat ze een kuil **graaft**.

Grammatica

1 Opgaven met vier zinnen waarin steeds een woord is onderstreept waarover een vraag wordt gesteld.

In welke zin is de persoonsvorm onderstreept?

A De mail is drie dagen geleden al gestuurd.
B Huilend rende het verdwaalde meisje weg.
C Je wil altijd de baas spelen over mij!
D* Zingend stond Larissa onder de douche.

2 Opgaven met vier zinnen waarin hetzelfde woord is onderstreept en waarover een vraag wordt gesteld. De antwoordalternatieven zijn zinnen met dit (stam)woord.

In welke zin is kopen de persoonsvorm?

A Gaan jullie morgen die hippe jurk kopen?
B Het kopen van alle schoolspullen kostte veel tijd.
C Mieke is gek op kleren kopen.
D* Zij kopen elk jaar een nieuwe auto.

3 Opgaven met een stamzin waarover een vraag wordt gesteld. De antwoordalternatieven zijn gedeeltelijke variaties van deze zin.

Het vliegtuig uit Parijs zal over een uur landen.

Wat is het gezegde in deze zin?

A Het vliegtuig uit Parijs
B Het vliegtuig uit Parijs zal
C zal
D* zal landen

4 Opgaven met een stamzin waarover een vraag wordt gesteld.
De antwoordalternatieven zijn onderdelen van deze zin.

Verbaasd trok Naima haar wenkbrauwen op.

Wat is het onderwerp in deze zin?

A Verbaasd
B trok
C* Naima
D haar wenkbrauwen

5 Opgaven met één zin waarover een vraag wordt gesteld.
De antwoordalternatieven zijn variaties van deze zin.

Welk onderstreept deel is de hoofdzin?

A Jamal gaat morgen wandelen, hoewel de weerman regen heeft voorspeld.
B* Jamal gaat morgen wandelen, hoewel de weerman regen heeft voorspeld.
C Jamal gaat morgen wandelen, hoewel de weerman regen heeft voorspeld.
D Jamal gaat morgen wandelen, hoewel de weerman regen heeft voorspeld.

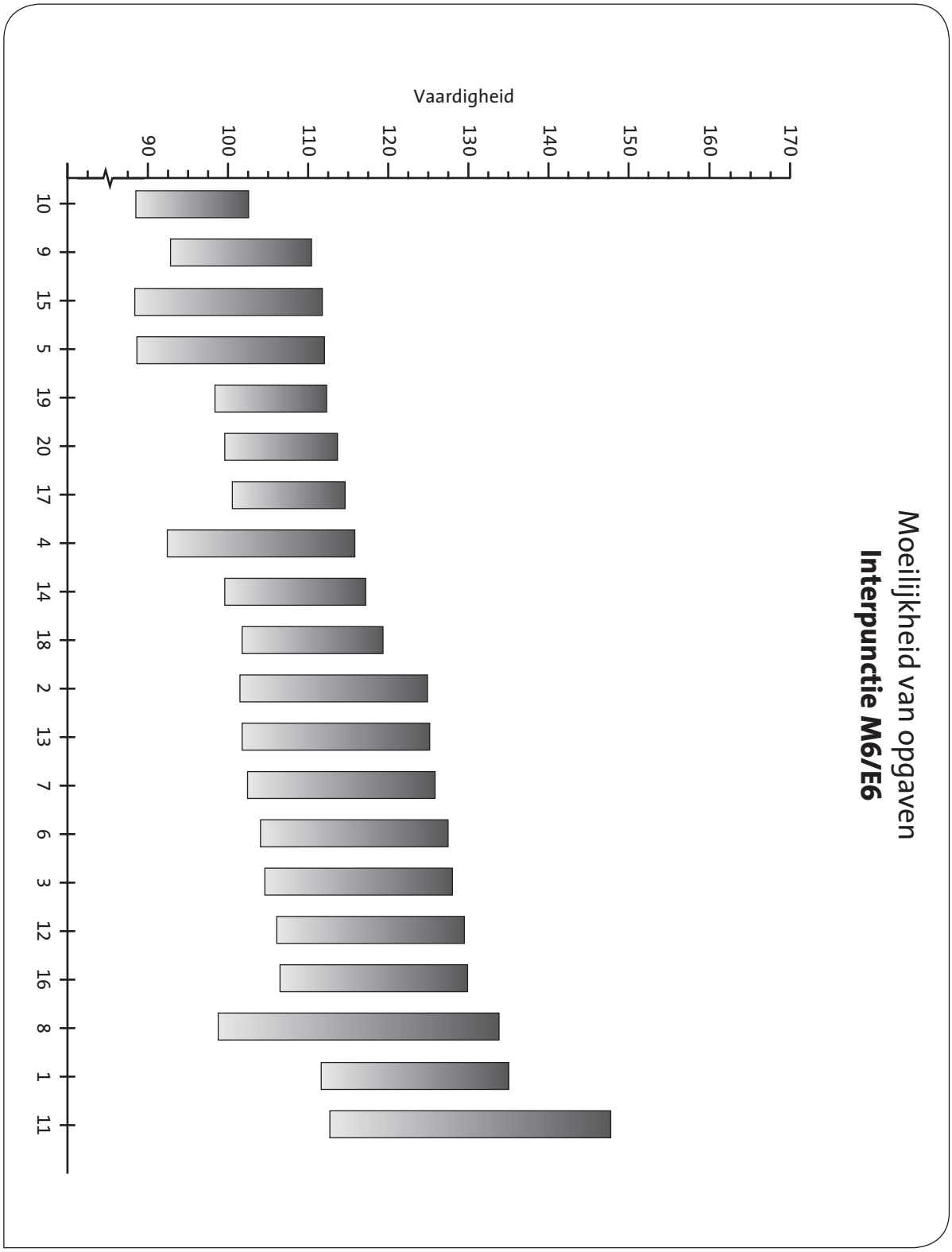
6 Opgaven met één zin waarover een vraag wordt gesteld.
De antwoordalternatieven zijn combinaties van bewerkingen van het gevraagde aspect.

We wachten een kwartiertje, niet langer!

Wat is juist?

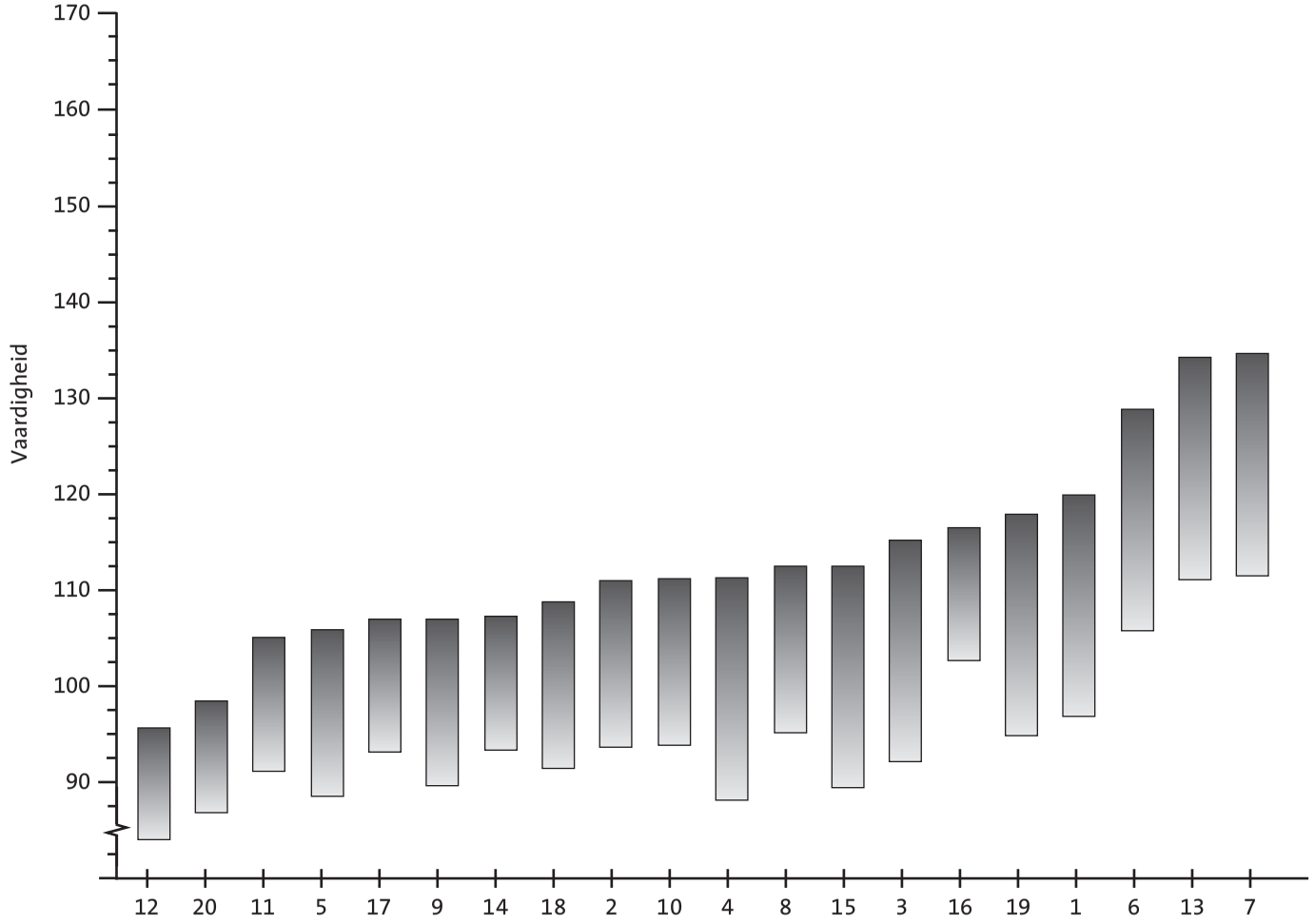
A wachten is enkelvoud en tegenwoordige tijd
B wachten is enkelvoud en verleden tijd
C* wachten is meervoud en tegenwoordige tijd
D wachten is meervoud en verleden tijd

Bijlage 5 P50- en P80-kanspunten van de opgaven in de toetsen voor groep 6, 7 en 8 in relatie tot de gemiddelde vaardigheidsscore voor de afnamemomenten



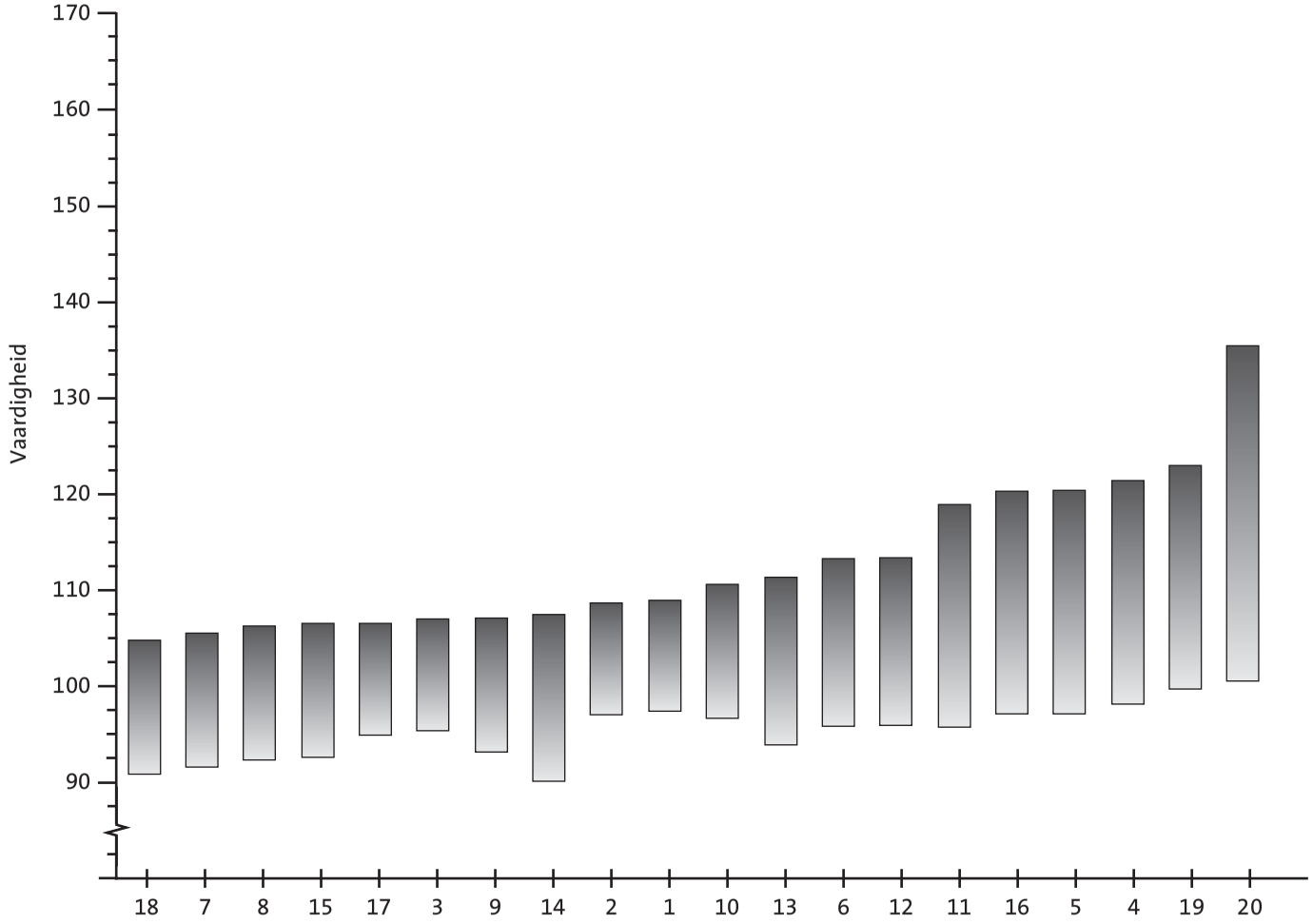
Moeilijkheid van opgaven

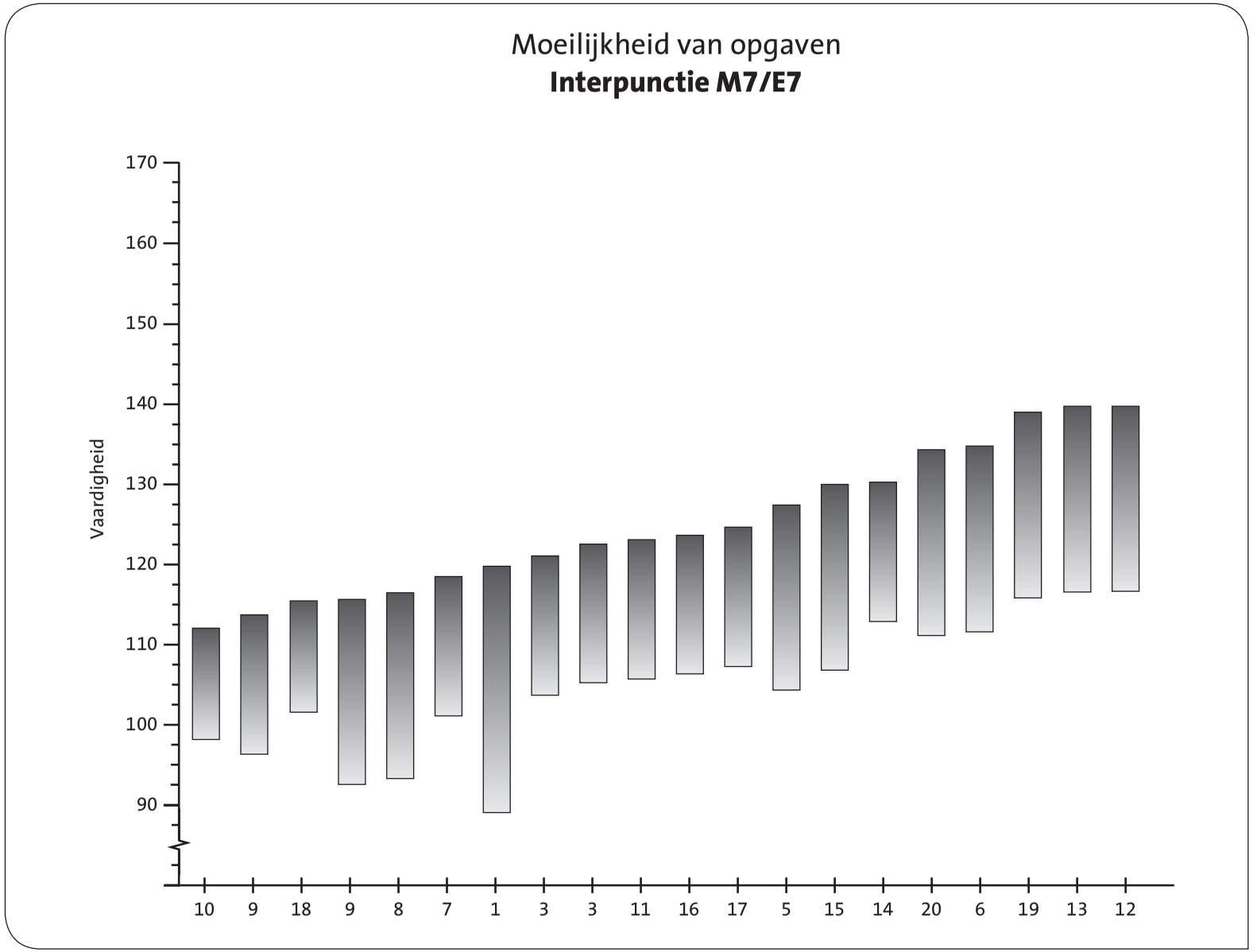
Spelling niet-werkwoorden M6/E6



Moeilijkheid van opgaven

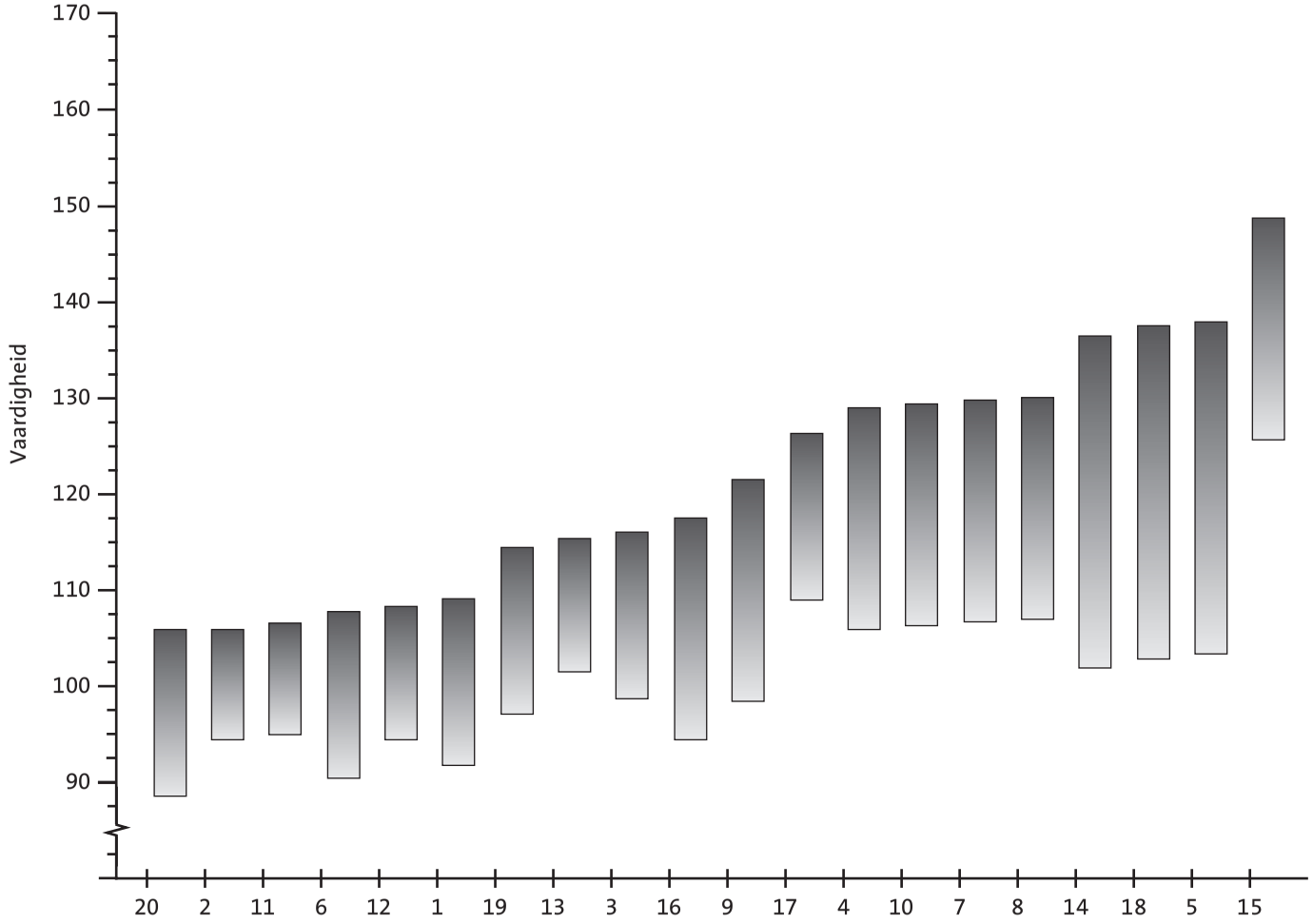
Grammatica M6/E6




Moeilijkheid van opgaven
Interpunctie M7/E7
Vaardigheid
170
160
150
140
130
120
110
100
90
10 9 18 9 8 7 1 3 3 11 16 17 5 15 14 20 6 19 13 12

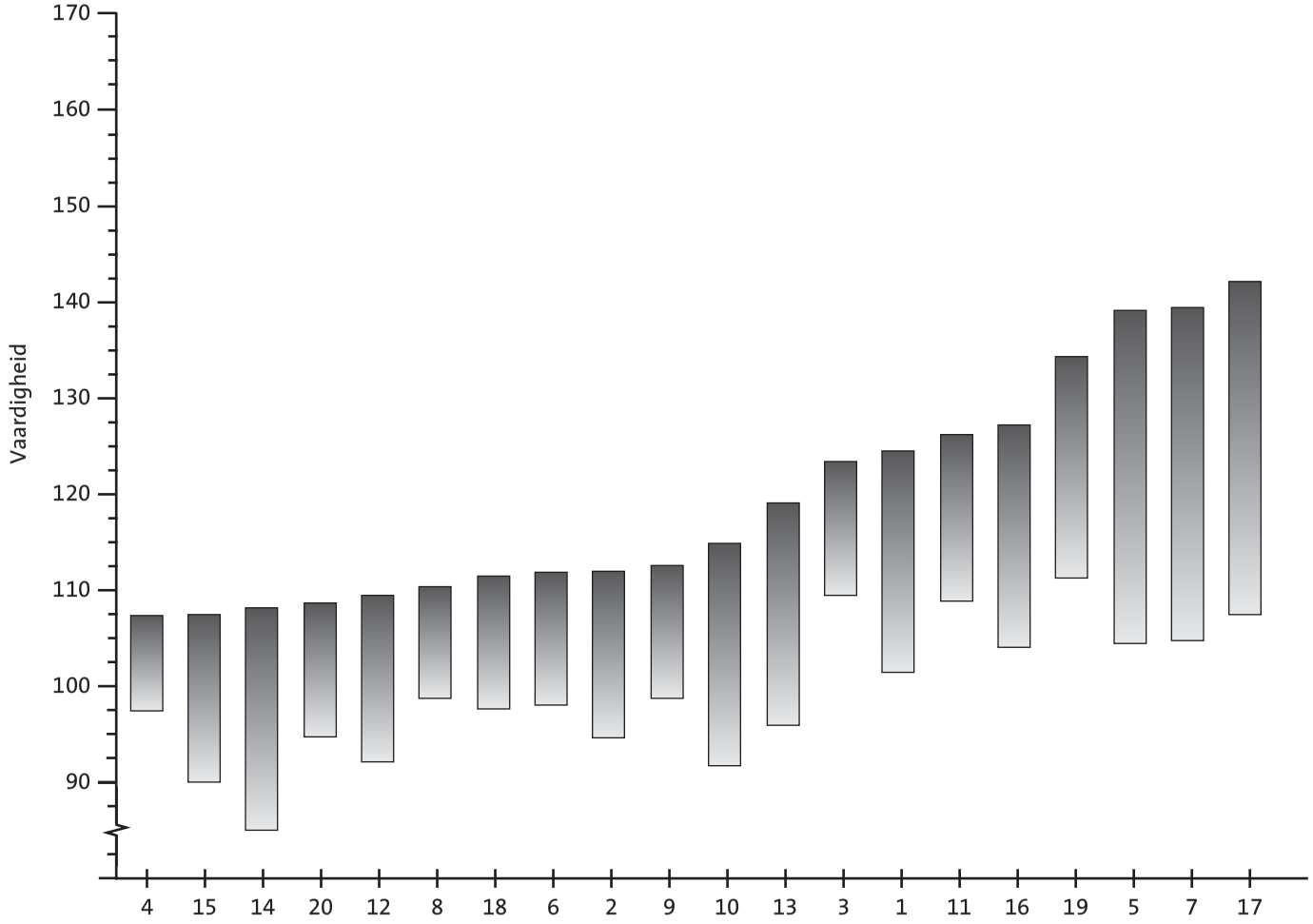
Moeilijkheid van opgaven

Spelling niet-werkwoorden M7/E7



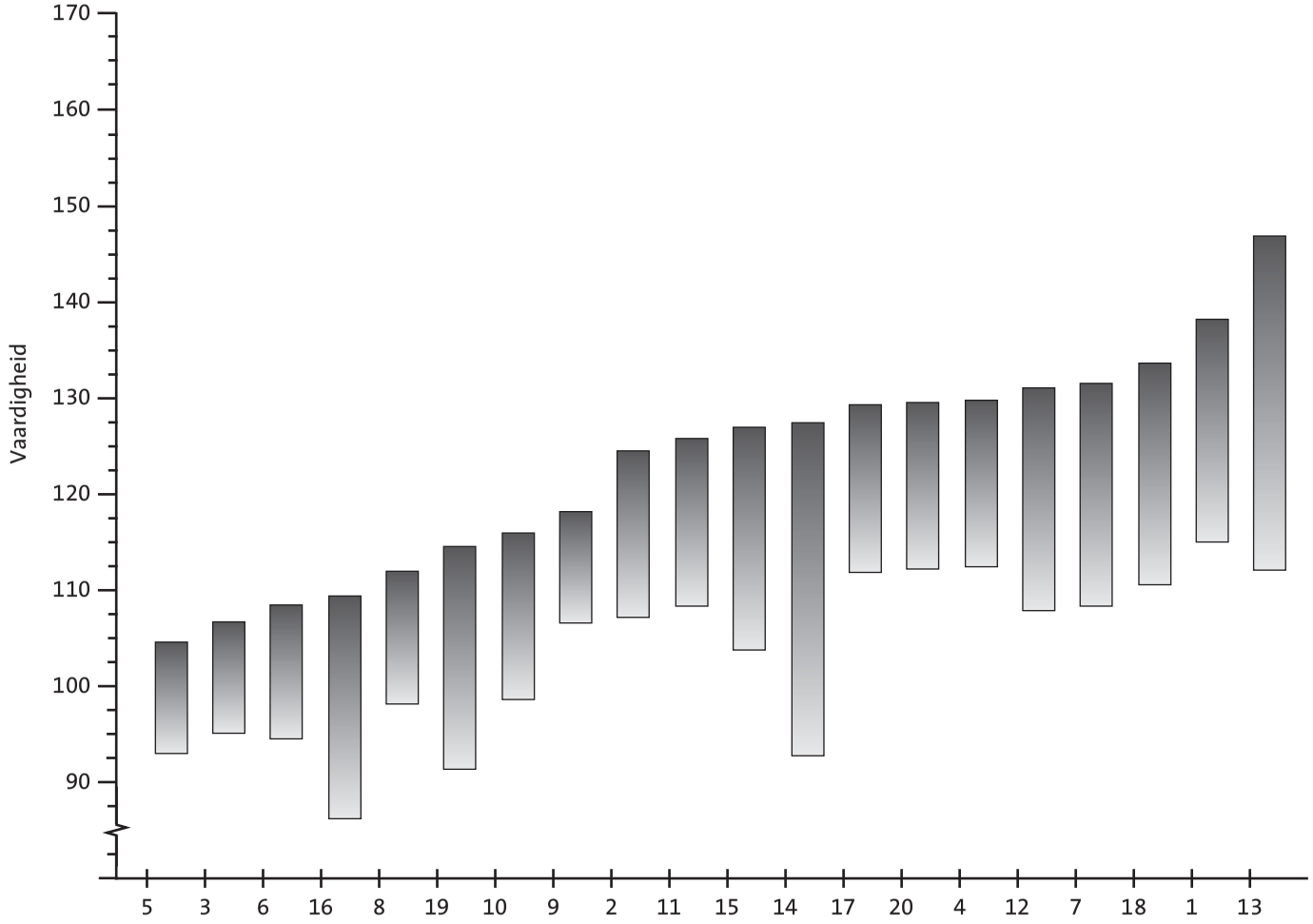
Moeilijkheid van opgaven

Grammatica M7/E7



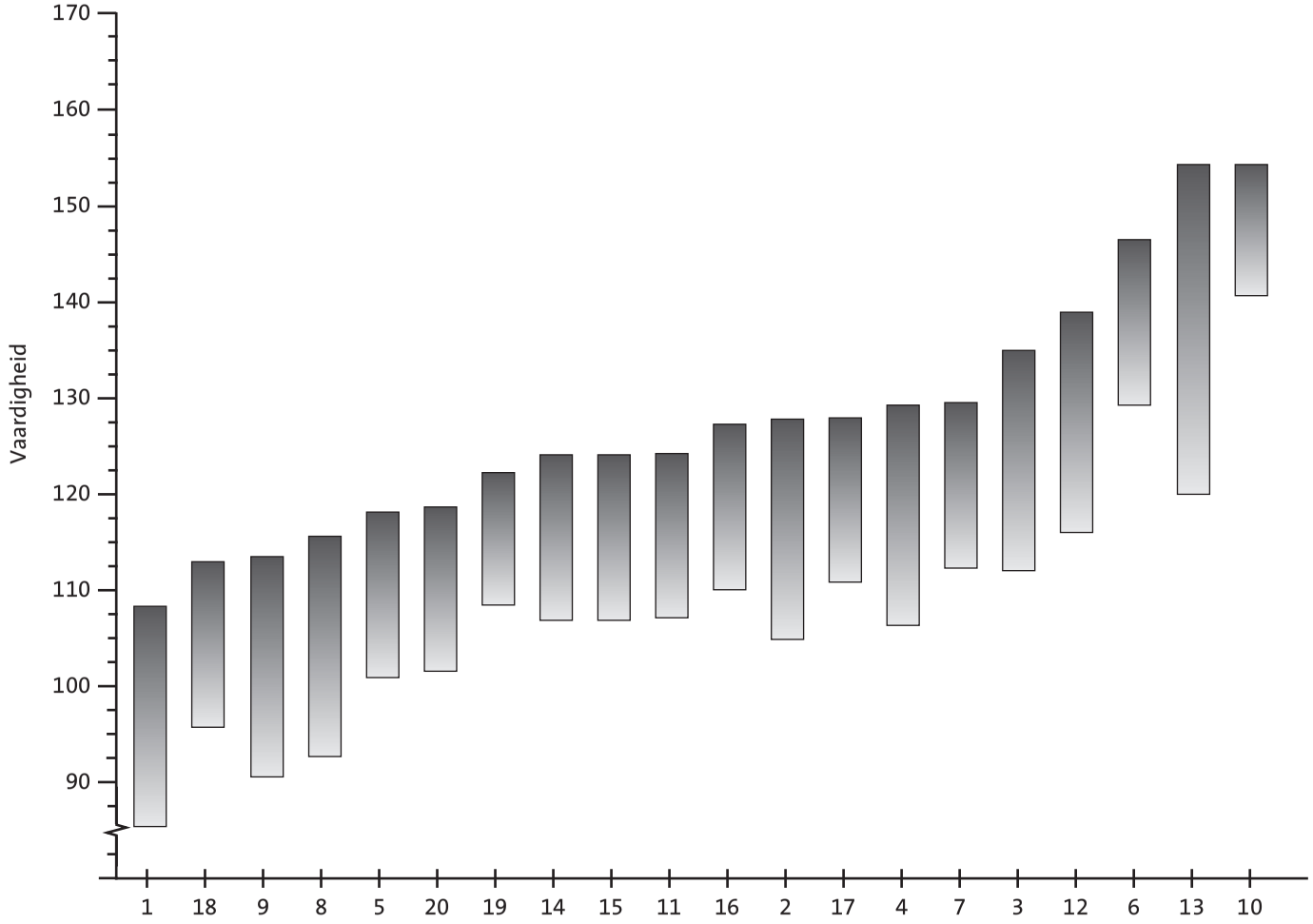
Moeilijkheid van opgaven

Spelling werkwoorden M7/E7



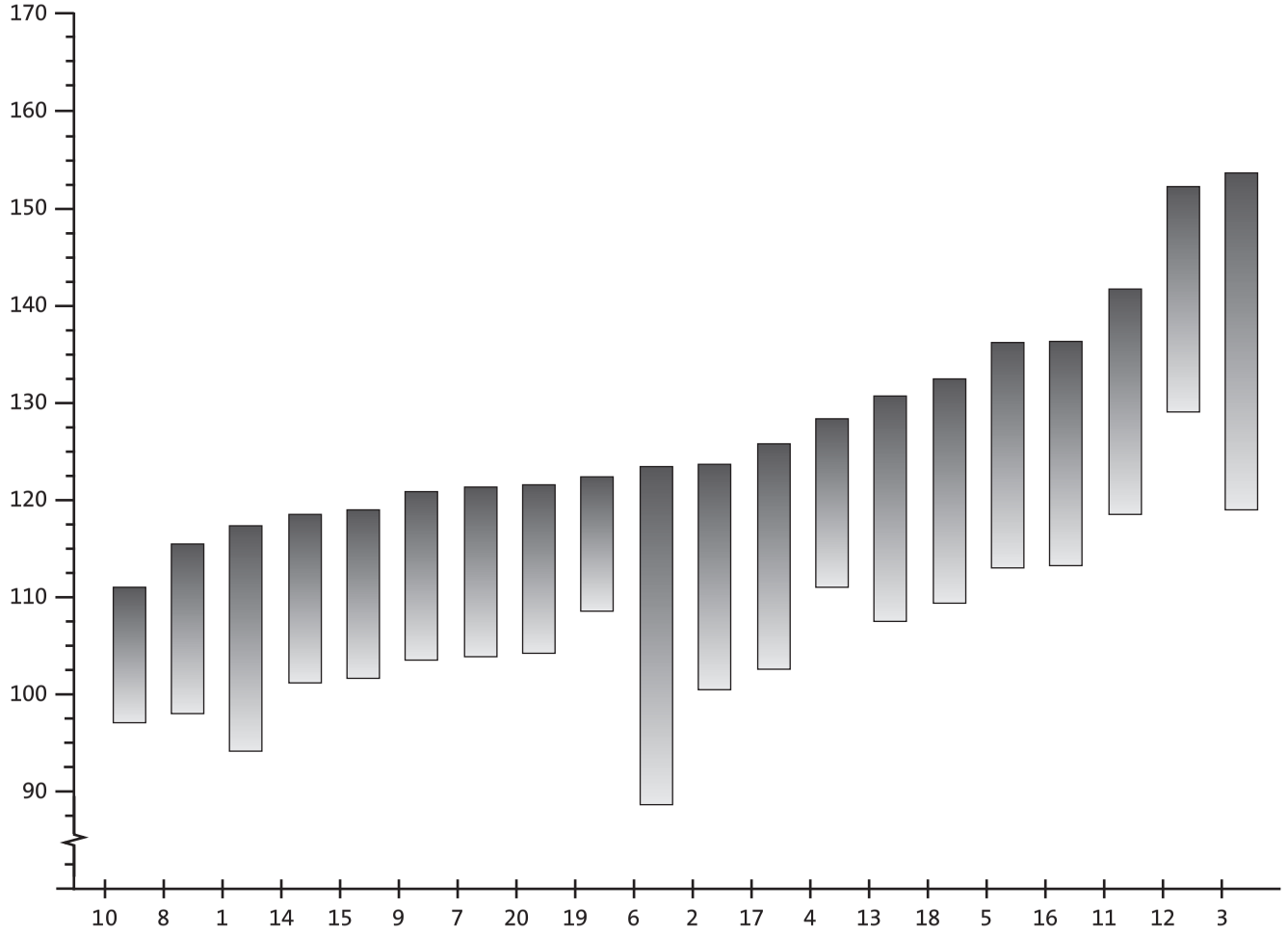
Moeilijkheid van opgaven

Interpunctie M8



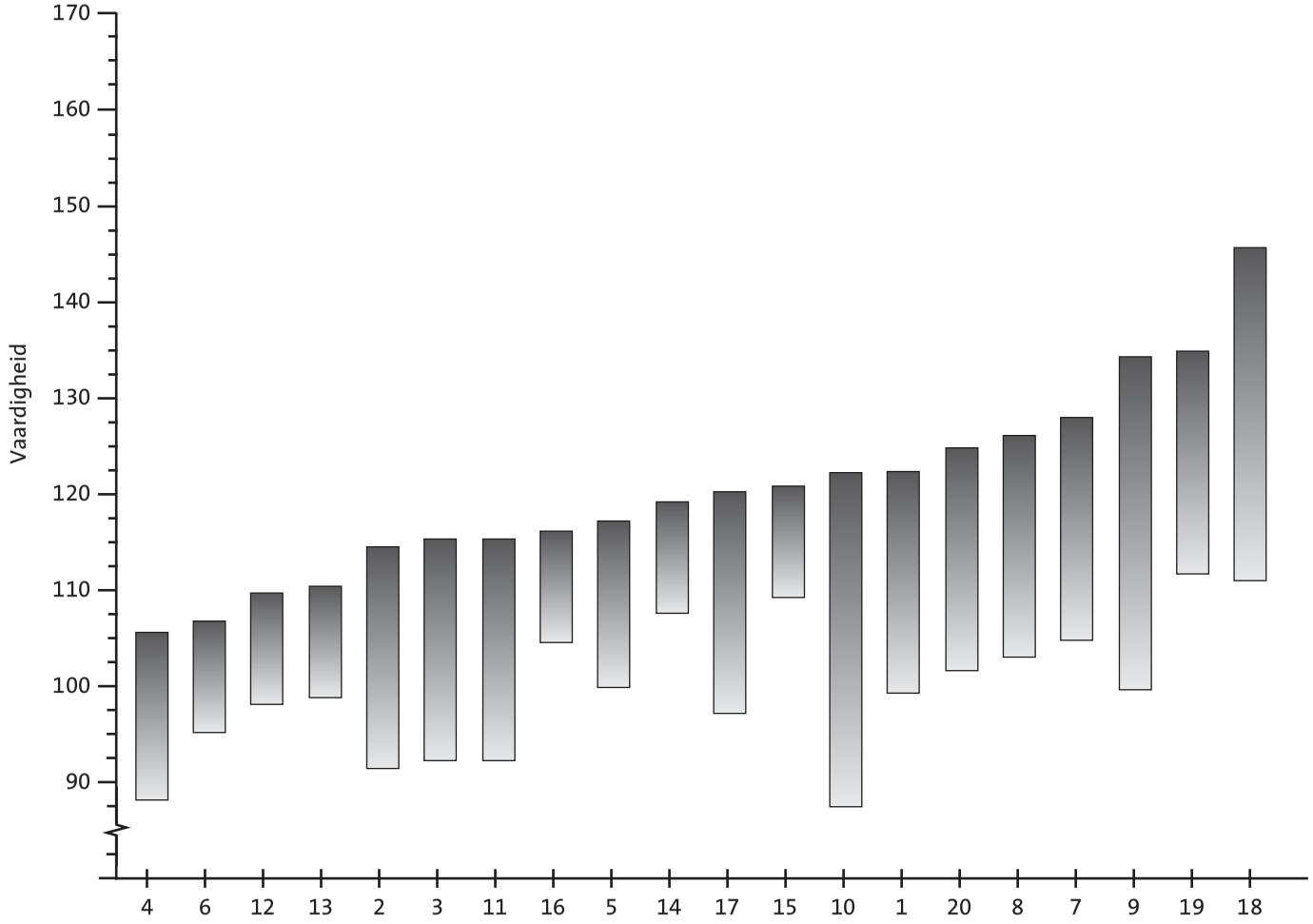
Moeilijkheid van opgaven

Spelling niet-werkwoorden M8



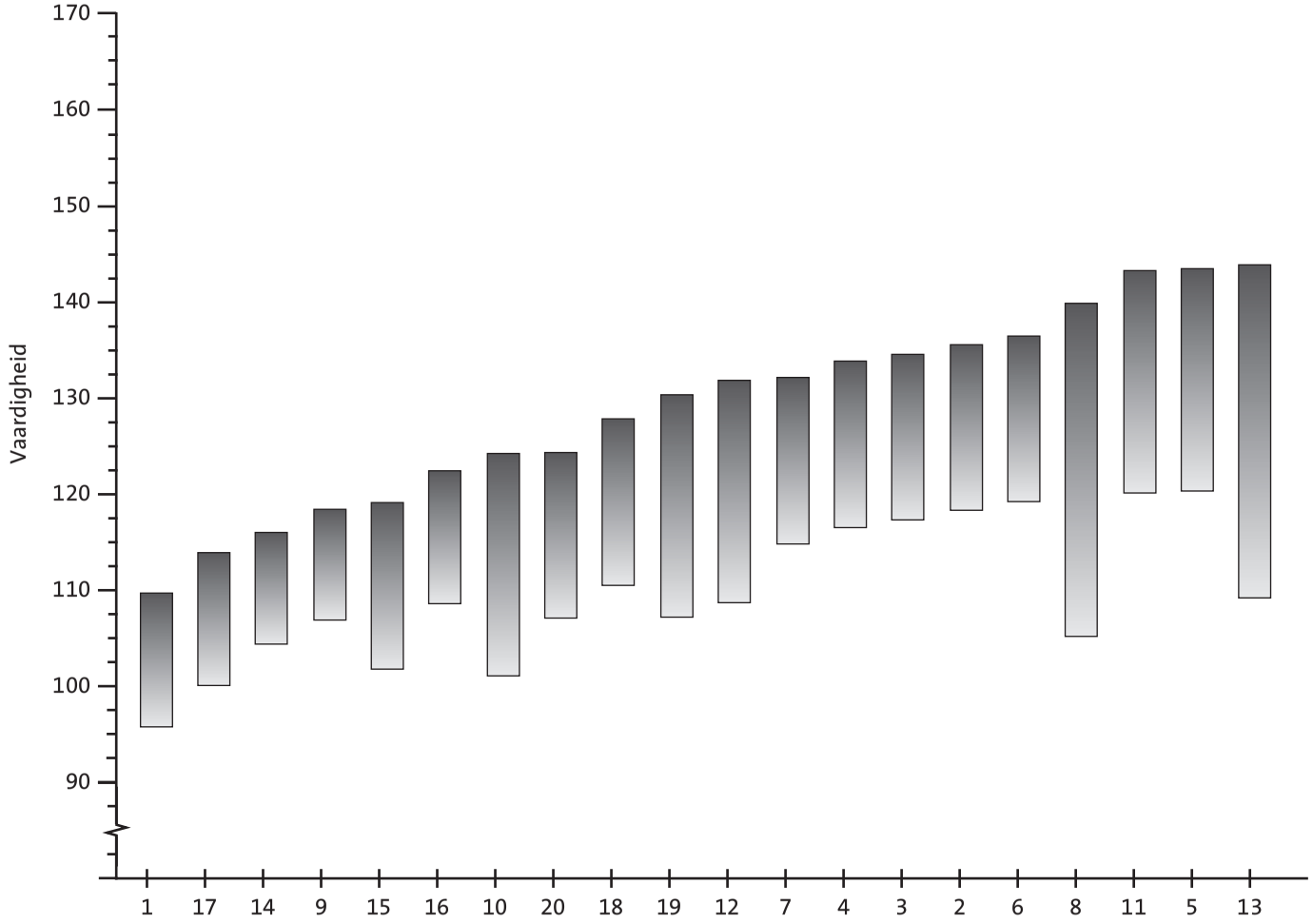
Moeilijkheid van opgaven

Grammatica M8



Moeilijkheid van opgaven

Spelling werkwoorden M8



Cito helpt je inzicht te krijgen in je ontwikkeling en mogelijkheden. Door kennis, vaardigheden en competenties objectief meetbaar te maken en de ontwikkeling er van te volgen, kun je het beste uit jezelf halen, verantwoorde keuzes maken en beter richting geven aan je toekomst. Cito draagt daaraan bij door wereldwijd werk te maken van goed en eerlijk toetsen, vanuit de kernwaarden kundig, toonaangevend, integer, innovatief en betrokken.

Cito
Amsterdamseweg 13
Postbus 1034
6801 MG Arnhem
T (026) 352 11 11
www.cito.nl

Fotografie: Ron Steemers