Rapportage Standaardbepalingen

Peilingsonderzoek mondelinge taalvaardigheid

2017

Roelien Linthorst

Bas Hemker

Irma Koerhuis

Remco Feskens

Jesse Koops

Cito

juni 2017

Inleiding

Stichting Cito heeft in opdracht van het College voor Toetsen en Examens (CvTE) drie standaardbepalingsbijeenkomsten uitgevoerd in het kader van het Peilingsonderzoek Mondelinge Taalvaardigheid. Het betreft een standaardbepaling luistervaardigheid, een standaardbepaling op een spreektaak (monoloog) en een standaardbepaling op een gesprekstaak (polyloog), voor niveau 1F en niveau 2F. In dit rapport doen we verslag van deze standaardbepalingen.

De standaardbepaling **luistervaardigheid** is bedoeld om een indicatieve cesuur te stellen met het oog op het onderdeel luistervaardigheid uit het Peilingsonderzoek. Deze cesuur zal door de Inspectie gebruikt worden om voor de zomervakantie de deelnemende scholen te kunnen rapporteren over de voorlopige vastgestelde vaardigheidsindicatie van hun leerlingen. In de periode 2016-2018 ontwikkelt Stichting Cito in opdracht van het CvTE een ankerset luistervaardigheid die het Referentiekader Luisteren representeert. Zodra de standaardbepaling van deze ankerset plaatsgevonden heeft, kan de indicatieve cesuur die nu op het Peilingsonderzoek gezet is, omgezet worden in een definitieve standaard. Hiervoor wordt in het voorjaar van 2018 de luistertoets uit het Peilingsonderzoek nogmaals afgenomen, gelijktijdig met het definitieve materiaal voor de ankerset.

De standaardbepalingen **spreken en gesprekken** leveren ook indicatieve cesuren op. In tegenstelling tot luistervaardigheid worden deze cesuren echter niet omgezet in definitieve cesuren, aangezien er geen ankersets spreken en gesprekken ontwikkeld zullen worden. Ook gaat het bij spreken en gesprekken om specifieke taakgerichte cesuren. Leerlingen hebben in dit Peilingsonderzoek om begrijpelijke redenen slechts één spreektaak en één gesprekstaak uitgevoerd, waardoor de standaarden niet representatief kunnen zijn voor de beheersing van de eisen die het Referentiekader stelt aan spreekvaardigheid of gespreksvaardigheid in het algemeen. Hier moet rekening mee gehouden worden bij de interpretatie van de gezette standaarden.

De standaardbepalingen hebben plaatsgevonden op 29 mei, 2 juni en 9 juni 2017. Alle standaardbepalingen zijn uitgevoerd door hetzelfde expertpanel. In paragraaf 1 beschrijven we de samenstelling van dit expertpanel. Vervolgens beschrijven we in paragraaf 2, 3 en 4 de opzet en de uitkomsten van de drie standaardbepalingen en de implicaties van deze uitkomsten. De gehanteerde methode voor de standaardbepaling luisteren, de 3DC-methode, is gebruikt voor standaardsettingen bij eerdere Referentiesets. De wijze van standaardsetting voor spreken en gesprekken is echter nieuw. Vandaar dat de gehanteerde procedure bij spreken en gesprekken uitgebreider beschreven wordt.

1. Het expertpanel standaardbepalingen

Het panel dat de standaarden heeft gezet, bestond in totaal uit achttien personen. Vijf van hen zijn taalexperts die ook zitting hebben in het expertpanel voor de Referentiesets Mondelinge Taalvaardigheid. Elf personen zijn werkzaam als leerkracht of directeur in het reguliere basisonderwijs en twee personen zijn werkzaam als leerkracht in het SBO. De mix van leerkrachten PO en SO/SBO is gemaakt met het oog op de standaardbepalingen Peilingsonderzoek Mondelinge Taalvaardigheid SO/SBO die in 2018 zullen plaatsvinden en waarbij de gestelde cesuren overeen moeten komen met de cesuren van dit project. Om dat te kunnen bewerkstelligen wordt gepoogd de groep standaardsetters in de twee jaren voor een groot deel gelijk te houden.

De deelnemende leerkrachten zijn werkzaam in alle regio's van Nederland, zijn afkomstig van stads- en plattelandsscholen met verschillende denominaties en hebben een leservaring die varieert van 4 jaar tot 38 jaar. Zoals in Tabel 1 te zien is, waren bij de standaardbepaling luisteren uiteindelijk 14 panelleden aanwezig, bij de standaardbepaling spreken 15 panelleden en bij de standaardbepaling

gesprekken ook 15 panelleden. Daarmee is een brede inbreng vanuit verschillende invalshoeken zoveel mogelijk geborgd.

Tabel 1 Overzicht deelnemers standaardbepalingen

deelnemer	achtergrond	luisteren	spreken	gesprekken
1	basisonderwijs	x	x	x
2	basisonderwijs	x	x	x
3	expertpanel	x	x	x
4	expertpanel	x	x	
5	expertpanel		x	x
6	basisonderwijs	x	x	x
7	basisonderwijs		(verhinderd door ziekte)	
8	basisonderwijs	x	x	x
9	SBO	x	x	x
10	basisonderwijs	x	x	x
11	basisonderwijs	x	x	x
12	expertpanel	x		
13	SBO	x	x	x
14	expertpanel		x	x
15	basisonderwijs	x		x
16	expertpanel	x	x	x
17	basisonderwijs	x	x	x
18	basisonderwijs		x	x
Totaal		14	15	15

2. Standaardbepaling Luistervaardigheid

2.1 De luistertoets

Voor de standaardbepaling luistervaardigheid is gebruikgemaakt van een luistertoets die door Bureau ICE ontwikkeld is voor het Peilingsonderzoek. Deze luistertoets bestaat uit 30 opgaven die de onderdelen *luisteren naar radio, tv en internet* en *luisteren naar instructies* uit het Referentiekader Taal meten. De toets bevat opgaven die ontwikkeld zijn voor niveau 1F en 2F en is digitaal en individueel afgenomen. De toets is geschaald volgens een IRT-methode.

2.2 Procedure standaardbepaling

Uitgangspunt bij alle standaardbepalingen in dit project vormden de eisen die het Referentiekader Taal stelt aan de mondelinge taalvaardigheid van leerlingen. Om die reden is er bij de start van de standaardbepaling luisteren uitgebreid stilgestaan bij de beschrijvingen die het Referentiekader geeft voor luistervaardigheid 1F en 2F. De panelleden bekeken deze beschrijvingen individueel en noteerden vervolgens de hoofdlijnen. Deze hoofdlijnen werden daarna plenair besproken. Ze werd ervoor gezorgd dat de onderscheidende kenmerken van niveau 1F en 2F voor alle panelleden duidelijk waren en dat de panelleden vanuit hetzelfde inhoudelijke vertrekpunt de standaarden konden zetten.

Bij de daadwerkelijke standaardbepaling luisteren is aan de beoordelaars gevraagd een inhoudelijk oordeel te geven over de hoeveelheid goed gemaakte opgaven die nodig is om een bepaald referentieniveauniveau te halen. De methode die hierbij gebruikt werd, is de 3DC-methode (Data Driven Direct Consensus methode, zie Keuning et al, 2017). De 3DC-methode deelt een toets op in

een aantal clusters, in dit geval 3 clusters die varieerden van 9 tot 11 opgaven per cluster. Tijdens de standaardbepalingsprocedure werd aan beoordelaars gevraagd om per cluster opgaven aan te geven welke score leerlingen zouden moeten behalen als zij zich precies op de grens van het desbetreffende referentieniveau bevinden. Aan de beoordelaars stelden we dus de vraag: “Hoeveel opgaven zou een leerling op dit cluster goed moeten maken als zijn/haar vaardigheid zich precies op de grens 1F of 2F van het referentieniveau bevindt?”

De bepaling van de standaard 1F en de standaard 2F gebeurde telkens in twee rondes. In de eerste ronde bepaalde ieder panellid individueel de score van de zogenaamde ‘grensleerling’ per cluster. Nadat alle panelleden hun oordeel gegeven hadden, werden die oordelen in een figuur weergegeven en aan de deelnemers getoond. Over de uitkomsten werd vervolgens gediscussieerd. De discussie werd gestart door aan een aantal beoordelaars te vragen om hun oordeel toe te lichten. De andere deelnemers konden hierop reageren. Na de discussie startte de tweede ronde. In de tweede beoordelingsronde werd de beoordelaars opnieuw gevraagd om in de figuur per cluster de grensscore te markeren. De beoordelaars kregen hierdoor de gelegenheid om hun grensscore bij te stellen als ze nieuwe inzichten uit de discussie wilden integreren in hun oordeel.

2.3 Uitkomsten standaardbepaling

In Tabel 2 worden de uitkomsten van de standaardbepalingen van de 3 clusters weergegeven. Door de grensscores van de 3 clusters op te tellen, krijgen we de standaarden voor de hele luistertoets.

Tabel 2 Grensscores voor clusters luistervaardigheid

	1F			**2F**		
	cluster 1	cluster 2	cluster 3	cluster 1	cluster 2	cluster 3
grensscores (gemiddelde)	5	6	4	8	9	7
maximum	6	7	5	8	9	8
minimum	4	4	3	7	8	7

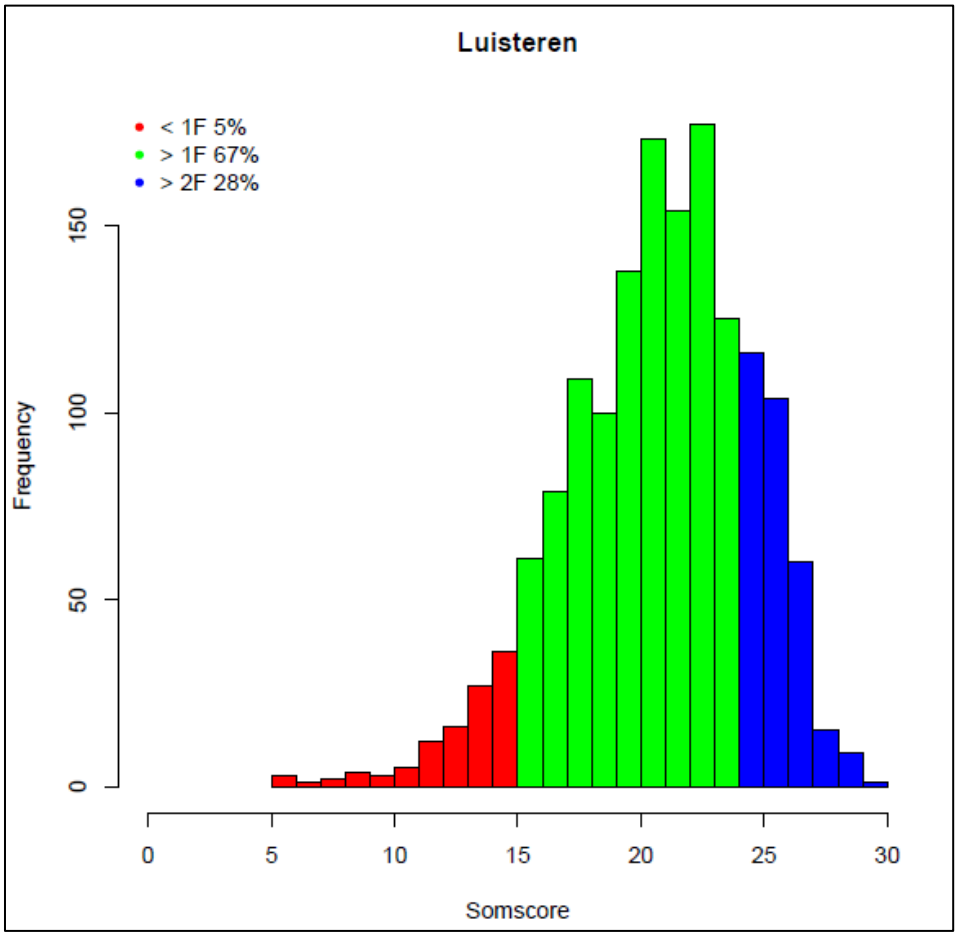
De standaardbepaling leverde zodoende de volgende grenzen op voor luistervaardigheid:

De standaard 1F Luisteren is geplaatst bij **15** van de 30 opgaven goed.

De standaard 2F Luisteren is geplaatst bij **24** van de 30 opgaven goed.

In Figuur 1 zijn de implicaties van de geplaatste standaarden weergegeven. 5% van de leerlingen die deelnamen aan het Peilingsonderzoek behaalt niveau 1F niet. 67 % van de leerlingen heeft een luistervaardigheid op niveau 1F, maar behaalt niveau 2F nog niet. 28 % van de leerlingen ten slotte heeft een luistervaardigheid op niveau 2F.

Figuur 1 Implicaties standaarden Luisteren



3. Standaardbepaling spreektaak

3.1 De spreektaak

De standaard spreekvaardigheid is gezet op een spreekopdracht die door Bureau ICE ontwikkeld is. De leerlingen moesten een monoloog houden in de vorm van een vlog die drie tot vijf minuten duurde. In de vlog deden ze verslag van een bijzondere gebeurtenis: een paar dagen geleden was in hun buurt de straatprijs van een bekende loterij gevallen. Iedereen uit die straat die meedeed aan de loterij had grote geldprijzen gewonnen. Ze moesten met de vlog hun volgers vermaken en ze informeren over wat ze hadden meegemaakt en wat ze daarvan vonden.
De vlog-opdracht bevat elementen die ontwikkeld zijn om spreekproductie op niveau 1F en 2F uit te lokken. De vlogs zijn vastgelegd op video en de filmfragmenten zijn beoordeeld door getrainde beoordelaars. De oordelen zijn geschaald volgens een IRT-methode.

3.2 Procedure standaardbepaling

Voor de standaardbepalingen spreken is gebruikgemaakt van een leerlinggerichte methode in plaats van de itemgerichte methode bij luisteren. Dat is gedaan omdat bij deze spreektaak en het gebruikte beoordelingsmodel de leerlingrespons een eenheid vormt en niet op te delen is in afzonderlijk te interpreteren opgaven. Dit betekent dat de standaard niet bepaald niet wordt door de evaluatie van items maar door de evaluatie van (voorbeelden van) spreekproductie (video-opnamen van vlogs) van

leerlingen. Hiermee kan achterhaald worden welke vaardigheid net voldoende is voor niveau 1F en welke vaardigheid net genoeg is om niveau 2F te behalen.

voorbereiding
Met het oog op de schaling van de leerlingprestaties spreekvaardigheid is de spreekproductie van de leerlingen geëvalueerd met behulp van een vastliggend beoordelingsmodel[1] waar een scoringsmodel op is toegepast. Met het scoringsmodel wordt bepaald hoe de leerlingrespons tot een beoordeling van de prestatie van de leerling leidt. Voor alle leerlingen was op deze manier al een score vastgelegd. Deze scores vormen de basis voor de standaardbepaling. De mogelijk scores lopen van 0 tot en met 17 punten.

Voor de standaardbepalingen moesten videofragmenten geselecteerd worden die representatief zijn voor de punten op de gehanteerde beoordelingsschaal. Het aantal mogelijk voorbeelden verschilde per scorepunt. Zo was het niet mogelijk om een voorbeeld voor score 0 te selecteren aangezien geen enkele leerling die score had gekregen. De minste voorbeelden waren te kiezen bij score 1 (2 voorbeelden), de meeste bij score 13 (285 voorbeelden). Daar waar er een keuzemogelijkheid was, is er bij selectie van fragmenten met vier factoren rekening gehouden: de beoordelaar, het beoordelingspatroon, de school en de leerlingkenmerken.

De factor *beoordelaar* is relevant omdat de score een reflectie moet zijn van de leerlingprestatie en niet van de beoordelaar. Hierbij is gelet op de ervarenheid en de strengheid van de beoordelaar. De beoordelaar zal relatief veel leerlingen gezien moeten hebben: dat komt hier neer op minstens 75 leerlingen van 7 verschillende scholen. Een beoordelaar moet daarnaast niet te streng of te mild zijn. Bij een te strenge beoordelaar ligt de score lager dan de prestatie rechtvaardigt, en bij een milde beoordeling is deze juist te hoog. Het was niet mogelijk om geheel onafhankelijk de strengheid van de beoordelaars te evalueren aangezien iedere prestatie op het moment van de selectie van de voorbeelden slechts door één beoordelaar was beoordeeld. Om toch enig zicht te hebben op de strengheid is de gemiddelde score van de beoordelaars als indicatie van de strengheid genomen[2].

De factor *beoordelingspatroon* is gebruikt omdat de totaalscore op verschillende manieren tot stand kan komen. Zo kan een leerling bijvoorbeeld een totaalscore van 4 punten halen door deze te scoren op één onderdeel óf door op verschillende onderdelen één punt te scoren. Sommige manieren zijn echter waarschijnlijker dan andere. Daarom is er bij de selectie van voorbeelden gekozen voor een samenstelling aan observaties die relatief vaak voorkomt bij die gegeven score. Ook is, om opnieuw rekening te houden met de factor beoordelaar, bij selectie van fragmenten gekeken of de prestatie van de leerlingen in de gekozen voorbeeldfragmenten inderdaad opliep bij toename van scores.

De factoren *school* en *leerling* zijn meegenomen om ervoor te zorgen dat de geselecteerde voorbeelden niet allemaal van dezelfde school of paar scholen komen, of dat de voorbeelden alleen afkomstig zijn van jongens of alleen van meisjes. Het filteren van de factor school kon op basis van de onderzoeksgegevens. Voor de leerlingkenmerken moesten eerst de opnamen bekeken worden. Tijdens het bekijken van de opnamen is ook naar de kwaliteit van de opname gekeken.

---

[1] Het is in feite een observatie-coderings-model: hierbij wordt een specifiek geobserveerde respons gecodeerd. Het wordt vaak een beoordelingsmodel genoemd, maar de beoordeling volgt echter pas nadat er een waardering op basis van de codering toegepast wordt.

[2] Aangenomen is dat de leerlingen willekeurig aan de beoordelaars zijn toegewezen en dat de vaardigheid van de leerlingen niet samenhangt met de koppeling aan een beoordelaar. In dat geval zal de score van de gemiddelde vaardigheid van de leerlingen bij een beoordelaar vrijwel gelijk zal zijn. Hierdoor is variatie in gemiddelde scores vooral aan de beoordelaar te wijten.

De standaardbepaling spreken startte net als bij luistervaardigheid met het doornemen van de referentieniveaus 1F en 2F. Deze referentieniveaus zijn eerst bekeken, individueel op hooflijnen genoteerd en vervolgens besproken. Zo waren de onderscheidende kenmerken van niveau 1F en 2F voor alle panelleden duidelijk.

De daadwerkelijke standaardbepaling bestond uit drie rondes (zie Figuur 2). In een *eerste globale ronde* van de standaardbepaling kregen de deelnemers van het expertpanel een aantal filmpjes te zien van leerlingen die de spreektaak uitgevoerd hebben. Deze filmpjes - die vrijwel de gehele vaardigheidsschaal besloegen - liepen op in beheersingsniveau (plaats op de vaardigheidsschaal spreken), van het laagste behaalde vaardigheidsniveau van de schaal tot het hoogst behaalde vaardigheidsniveau. De deelnemers gaven individueel aan bij welk filmpje volgens hen de grens lag voor beheersing van 1F. Hiermee stelde het expertpanel een eerste globale cesuur.

Figuur 2 Voorbeeld rondes standaardbepaling Spreken

STANDAARDBEPALING 1F

Nummer:
Naam expert:

1. Globale ronde

Punt op schaal	**3**	**5**	**7**	**9**	**11**	**13**	**15**
Filmpje	10790101	10230104	10870132	20470105	11080102	20030101	11470106
Eerste leerling uit reeks die niveau 1F net haalt							

2. Detailronde

Punt	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Filmpje	11440114	10570103	11410106	11360101	10530108	11360116	11090112	10870101	20550103	10350103	20150103	20120105	11440102
reeks A													
reeks B													
reeks C													
reeks D													
reeks E													
reeks F													
reeks G													
Ronde 1													
Eerste leerling 1F													
Ronde 2													
Eerste leerling 1F													

Hierna volgde een *tweede gedetailleerde ronde* waarin ingezoomd werd op de globale cesuur die in de eerste ronde gesteld was. Deelnemers kregen nu enkele filmpjes te zien van leerlingen met een score vlak onder, op en vlak boven de globale cesuur van de eerste ronde. Ze stelden opnieuw individueel de cesuur vast voor deze serie filmpjes. Deze ronde werd besproken en de experts kregen in een derde en laatste ronde de kans om hun oordeel nog bij te stellen op basis van deze discussie. Voor niveau 2F werd de hele procedure herhaald. Door te werken met een globale en een gedetailleerde ronde werd de standaard bepaald op basis van zo veel mogelijk leerlingprestaties, waarbij tegelijkertijd een zo nauwkeurig mogelijke standaard gesteld werd. Dit levert een optimale verdeling van benodigde tijd en nauwkeurigheid op, zonder de beoordelaars van tevoren te sturen naar een verwacht niveau.

3.3 Uitkomsten standaardbepaling

De standaardbepaling voor Spreken 1F leverde na de derde ronde een gemiddelde grensscore op van 7,27 van de 17 mogelijke scorepunten, met een standaardfout van het gemiddelde van 0,23. Op basis van dit gemiddelde en deze standaardfout verwachten we met ongeveer 90% zekerheid dat het *werkelijke gemiddelde* tussen de 7 en 8 in ligt. In alle drie de rondes zat de gemiddelde waarde tussen de 7 en de 8.

De standaardbepaling voor Spreken 2F leverde na de derde ronde een gemiddelde grensscore op van 11,73 van de 17 mogelijke scorepunten, met een standaardfout van het gemiddelde van 0,18. Op basis van deze gegevens verwachten we met tussen en 90% en 95% zekerheid dat het werkelijke gemiddelde tussen de 11 en 12 in ligt.

De standaardbepaling leverde zodoende de volgende grenzen op bij deze taak:

De standaard 1F Spreken is geplaatst bij **8** van de 17 scorepunten.

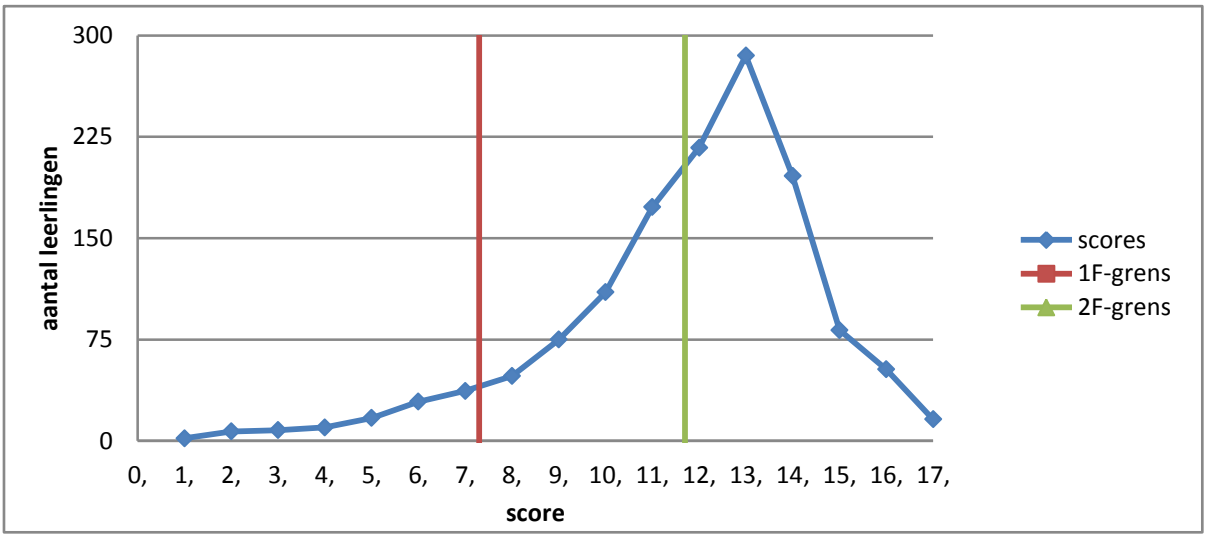
De standaard 2F Spreken is geplaatst bij **12** van de 17 scorepunten.

Voor de groep van 1365 beoordeelde leerlingen betekent dit dat 8% van de leerlingen onder niveau 1F presteert (37 leerlingen), en 92% van de leerlingen minstens op niveau 1F scoort. Bij 2F zien we dat 62% van de leerlingen dat niveau haalt op deze taak. Dat betekent dat 30% van de leerlingen 1F haalt, maar geen 2F. De gegevens per niveau en per score zijn te vinden in Tabel 3. Deze vormen de basis van Figuur 3 waarin de scoreverdeling is gegeven en Figuur 4 waarin de cumulatieve scoreverdeling gegeven is.

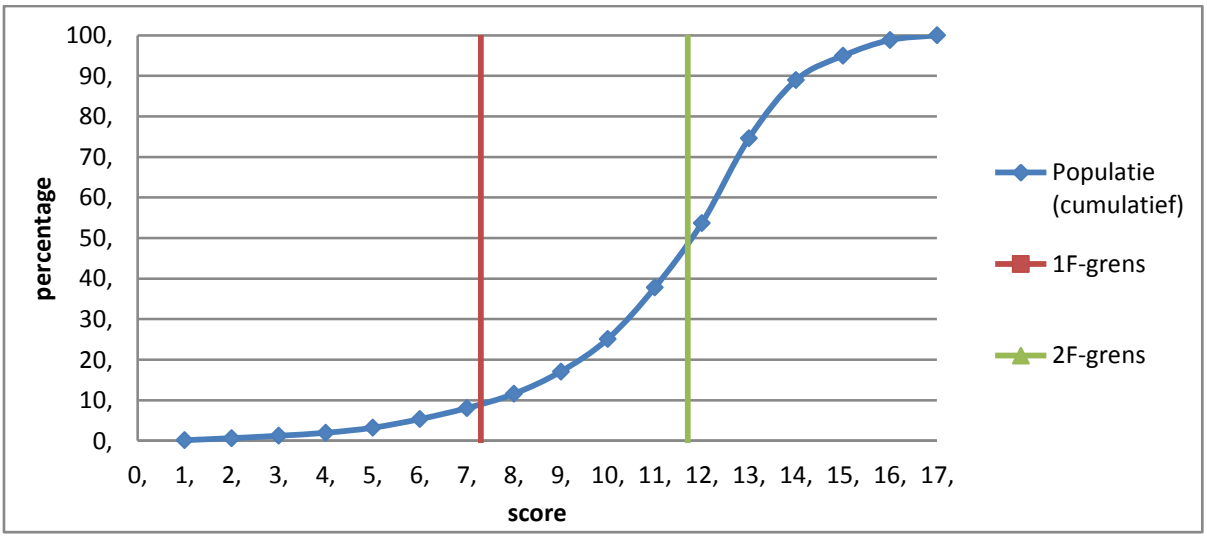
Tabel 3 Verdelingen onderzoeksgroep leerlingen per niveau en score Spreken

Niveau	onder 1F								1F				2F					
% per niveau	8								30				62					
Score	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
aantal leerlingen	0	2	7	8	10	17	29	37	48	75	110	173	217	285	196	82	53	16
% per score	0	0	1	1	1	1	2	3	4	5	8	13	16	21	14	6	4	1
cumulatief %	0	0	1	1	2	3	5	8	12	17	25	38	54	75	89	95	99	100

Figuur 3 Frequentieverdeling van de scores Spreken



Figuur 4 Cumulatieve verdeling percentages bij scores Spreken



3.4 Discussie standaardbepaling

Een kanttekening die gemaakt moet worden bij de standaarden voor spreken is dat de standaarden hier gebaseerd zijn op één taak, met één beoordelingsschema en één scoringsprocedure. Een generalisatie naar 'het spreken in het basisonderwijs' in het algemeen is daarmee zoals eerder gezegd niet mogelijk.

Ook valt er op te merken dat de standaardbepaling ondanks een zorgvuldige selectie door een enkele verzameling van experts is uitgevoerd. Een panel in een andere samenstelling was mogelijkerwijs tot een ander oordeel gekomen. Daarom hebben we een evaluatie van de oordelen van de panelleden ten opzichte van de beoordelaars uitgevoerd. In het beoordelingsschema dat de beoordelaars gehanteerd hebben, is een vraag opgenomen naar een globaal oordeel over de uitvoering van de spreektaak door de leerling, waarbij de beoordelaars de beoordeelde leerling op een vijfpuntschaal

(zwak-matig-voldoende-goed-zeer goed) moesten beoordelen met de referentieniveaus is het achterhoofd.

Als we aannemen dat een globaal oordeel 'voldoende' niveau 1F aanduidt en het oordeel 'goed' 2F, dan is het mogelijk aan te geven bij welke score de verdeling naar wel of niet 1F en wel of niet 2F de hoogste samenhang tussen beoordelaars en de experts van de standaardbepaling zou opleveren. Bij vergelijking van oordelen op deze manier ligt de grens voor het niveau 1F zoals deze uit de standaardbepaling komt rond het optimale beslispunt. Er is hier dus sprake van een overeenkomstig oordeel. Voor niveau 2F ligt de bepaalde grens *nabij* het optimale beslispunt. Dit optimale beslispunt zou hier één punt zou hoger liggen. Het is goed om bevestigd te zien dat de standaarden ook om en nabij het 'optimale beslispunt' zitten voor de grenzen onvoldoende/voldoende en voldoende/goed.

4. Standaardbepaling gesprekken

4.1 De gesprekstaak

De standaard gesprekken is gezet op een gespreksopdracht die door Bureau ICE ontwikkeld is. De gespreksopdracht betrof een groepsgesprek met drie leerlingen die ongeveer tien minuten duurde. In dit gesprek moesten de leerlingen plannen maken om geld in te zamelen voor een goed doel. Het doel van het gesprek was dat de leerlingen in gezamenlijk overleg tot een onderbouwd idee kwamen. De gespreksopdracht bevat elementen die ontwikkeld zijn om spreekproductie op niveau 1F en 2F uit te lokken. De gesprekken zijn vastgelegd en de filmfragmenten zijn beoordeeld door getrainde beoordelaars. De oordelen zijn geschaald volgens een IRT-methode.

4.2 Procedure standaardbepaling

De procedure voor de standaardbepalingen bij gesprekken heeft veel overeenkomsten met de procedure die gebruikt is bij spreken. Er is, om dezelfde redenen als bij spreken, gebruikgemaakt van een leerlinggerichte methode. Dit betekent dat de standaard niet bepaald wordt door de evaluatie van items maar van de evaluatie van (voorbeelden van) leerlingproductie. Hiermee kan achterhaald worden welke vaardigheid net voldoende is voor niveau 1F en welke vaardigheid net genoeg is om niveau 2F te behalen.

Voorafgaand aan de standaardbepaling is – net als bij spreken- de gespreksproductie van de leerlingen al geëvalueerd met behulp van een vastliggend beoordelingsmodel[3] waar een scoringsmodel op toe is gepast. Met het scoringsmodel wordt bepaald hoe de leerlingrespons tot een beoordeling van de prestatie van de leerling leidt. Voor alle leerlingen was op deze manier al een score vastgelegd. Deze scores vormen de basis voor de standaardbepaling.

Om representatieve voorbeeldfragmenten te selecteren zijn leerlingprestaties geclusterd per scorepunt. De mogelijk scores lopen van 0 tot en met 10 punten. Het aantal mogelijk voorbeelden verschilde per scorepunt. De minste voorbeelden zijn er van leerlingresponsen die tot een score 0 leiden (19 leerlingen); de meeste zijn er van responsen die tot een score 8 of 9 leiden (in beide gevallen 259 leerlingen). Bij het selecteren van voorbeelden van responsen is met dezelfde vier factoren rekening gehouden als bij spreken. Dat zijn de factoren: *beoordelaar*, *beoordelingspatroon*, de *school* en de *leerlingkenmerken*. De redenen hiervoor zijn besproken bij de beschrijving van de

[3] Het is in feite een observatie-coderings-model: hierbij wordt een specifiek geobserveerde respons gecodeerd. Het wordt vaak wel een beoordelingsmodel genoemd. De beoordeling volgt echter pas nadat er een waardering op basis van de codering toegepast wordt.

procedure bij spreken. Bij gesprekken is er nog een vijfde factor meegenomen. Dat is de samenstelling van de groep leerlingen die het gesprek voerde.
De gesprekken zijn gevoerd in drietallen. Leerlingen met een zelfde score, bijvoorbeeld 7, kunnen in verschillende contexten geobserveerd zijn. De score 7 kan binnen een drietal de hoogste, de laagste, de middelste of dezelfde score zijn. Dat laatste punt geeft ook aan dat scores van gespreksdrietallen heterogeen kunnen zijn (scores van de leerlingen liggen uit elkaar), of juist homogeen (in het uiterste geval: alle leerlingen hebben dezelfde score). Bij de selectie is er met deze variatie rekening gehouden door responsen in verschillende contexten te kiezen.

Wederom startte de standaardbepaling met het doornemen van de referentieniveaus 1F en 2F. Deze referentieniveaus zijn eerst bekeken, individueel op hooflijnen genoteerd en vervolgens besproken. Zo waren de onderscheidende kenmerken van niveau 1F en 2F voor alle panelleden duidelijk.

In een *eerste globale ronde* van de standaardbepaling kregen de deelnemers van het expertpanel een aantal filmpjes te zien van leerlingen die de gesprekstaak uitgevoerd hebben. Hierbij werd aangegeven welke leerling hier beoordeeld werd (links-midden-rechts; jongen-meisje). Deze filmpjes - die vrijwel de gehele vaardigheidsschaal besloegen - liepen op in beheersingsniveau (plaats op de vaardigheidsschaal gesprekken). De deelnemers gaven individueel aan bij welk filmpje volgens hen de grens lag voor beheersing van 1F. Hiermee stelde het expertpanel een eerste globale cesuur.

Ook de tweede ronde was gelijk aan die bij spreken: een *gedetailleerde ronde* waarin ingezoomd werd op de globale cesuur die in de eerste ronde gesteld was. Deelnemers kregen nu enkele filmpjes te zien van leerlingen met een score vlak onder, op en vlak boven de globale cesuur van de eerste ronde. Ze stelden opnieuw individueel de cesuur vast voor deze serie filmpjes. Deze ronde werd besproken en de experts kregen in een derde en laatste ronde de kans om hun oordeel nog bij te stellen op basis van deze discussie. Voor niveau 2F werd de hele procedure herhaald.

De positieve ervaringen bij de standaardbepaling Spreken bevestigde de voordelen van deze methode. Bij gesprekken was het bepalen van de standaard complexer omdat de filmpjes langer waren dan bij spreken. Ook speelde de interactie tussen de leerlingen een rol, wat een grotere concentratie vergde om te focussen op de beoordeelde leerling. Daarom kregen de deelnemers bij gesprekken de mogelijkheid om nog meer voorbeelden van een scorepunt te bekijken door deze te selecteren uit een extra map filmpjes. Zo konden de experts bij twijfel nog extra observaties doen van leerlingen met eenzelfde scorepunt.

4.3 Uitkomsten standaardbepaling

De standaardbepaling voor gesprekken 1F leverde na de derde ronde een gemiddelde grensscore op van 4,60 van de 10 mogelijke scorepunten, met een standaardfout van het gemiddelde van 0,16. Met een zekerheid van 99% ligt de standaard tussen de 4 en de 5. De standaardbepaling voor Gesprekken 2F leverde na de derde ronde een gemiddelde grensscore op van 7,87 van de 10 mogelijke scorepunten, met een standaardfout van het gemiddelde van 0,09. Met een betrouwbaarheid van tussen de 90-95% kan gesteld worden dat de gemiddelde grensscone tussen 7 en 8 ligt.

De standaardbepaling leverde zodoende de volgende grenzen op bij deze taak:

De standaard 1F Gesprekken is geplaatst bij **5** van de 10 scorepunten.

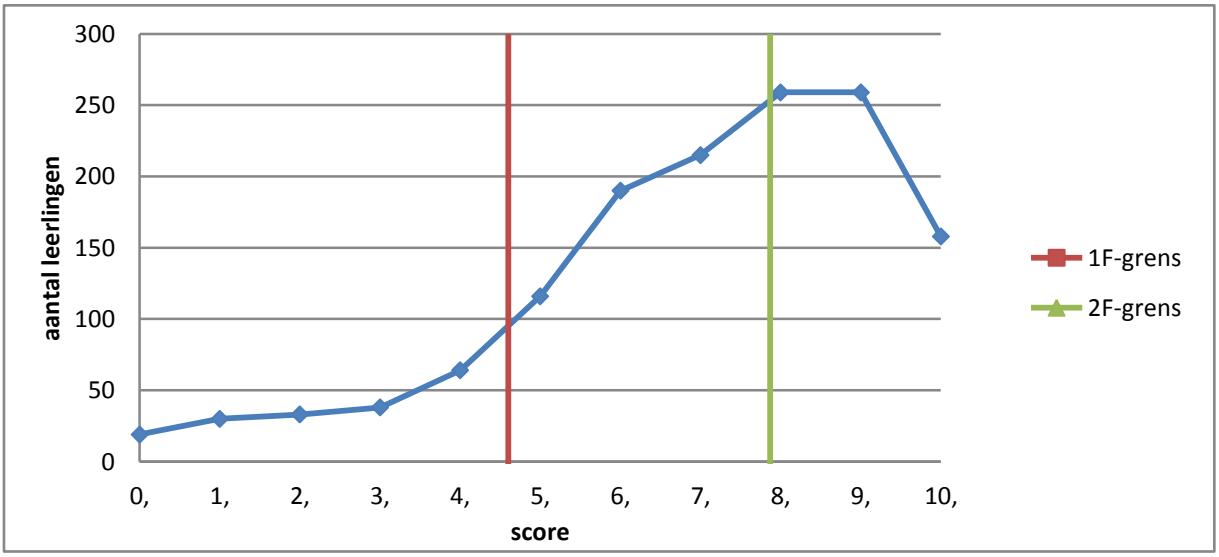
De standaard 2F Spreken is geplaatst bij **8** van de 10 scorepunten.

Voor de groep van 1381 beoordeelde leerlingen betekent dit dat 13% van de leerlingen onder niveau 1F presteert (184 leerlingen) en 84% van de leerlingen minstens op 1F niveau scoort. Bij 2F zien we dat 49% van de leerlingen (676 leerlingen) dat niveau haalt op deze taak. Dat betekent dat 30% van de leerlingen 1F haalt, maar geen 2F. De gegevens per niveau en per score zijn te vinden in Tabel 4. Deze vormen de basis van Figuur 5 waarin de scoreverdeling is gegeven, en Figuur 6 waarin de cumulatieve scoreverdeling gegeven is.

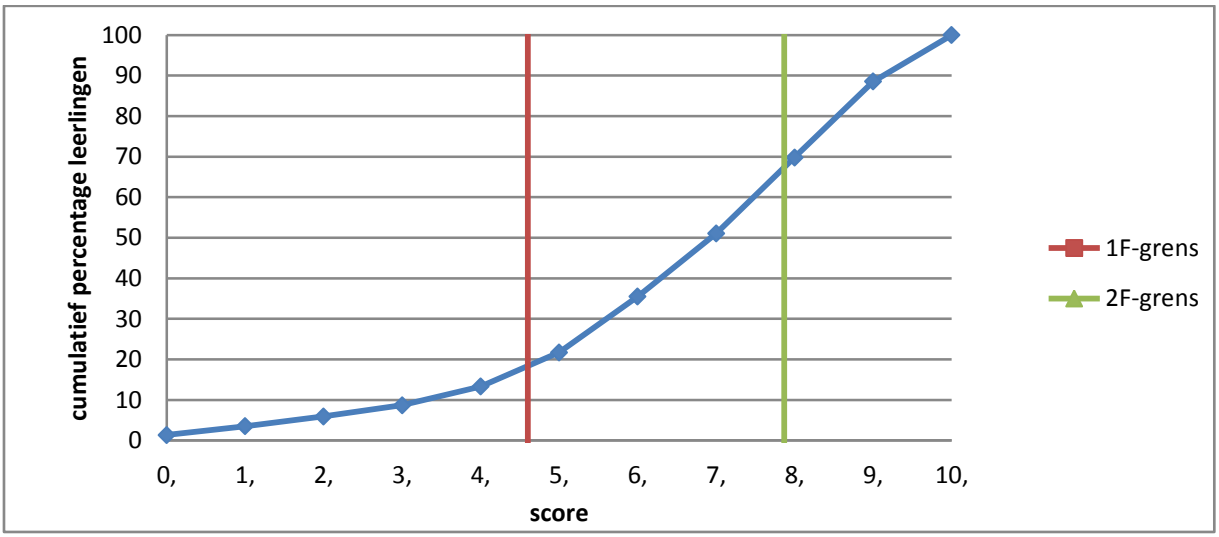
Tabel 4 Verdelingen onderzoeksgroep leerlingen per niveau en score Gesprekken

Niveau	<1F					1F			2F		
% per niveau	13					38			49		
Score	0	1	2	3	4	5	6	7	8	9	10
aantal leerlingen	19	30	33	38	64	116	190	215	259	259	158
% per score	1	2	2	3	5	8	14	16	19	19	11
cumulatief %	1	4	6	9	13	22	35	51	70	89	100

Figuur 5 Frequentieverdeling van de scores Gesprekken



Figuur 6 Cumulatieve verdeling percentages bij scores gesprekken



4.4 Discussie standaardbepaling

Eenzelfde kanttekening als bij spreken is hier ook te maken: het betreft slechts een enkele gesprekstaak, met een beoordelingsschema en een scoringsprocedure. Voor een verdere generalisatie naar het gehele domein van gespreksvaardigheid in het basisonderwijs zouden meer verschillende taken onderzocht moeten worden. De beoordeling van een individu bij gesprekken is daarbij complex vanwege de interactie en dus de invloed van de gespreksgenoten.

Ter validering van de gezette standaard is ook bij gesprekken geëvalueerd hoe de standaarden zich verhouden tot de optimale grens tussen onvoldoende (matig of zwak) en voldoende (al dan niet inclusief hogere oordelen) en tussen voldoende (of lager) en (zeer) goed. Op basis van de globale oordelen zou het *optimale grenspunt* tussen onvoldoende en voldoende moeten liggen tussen 4 en 5 aangezien dat de hoogste kappa oplevert. Dat komt overeen met de standaard voor 1F.
De optimale grens tussen voldoende en goed zou liggen tussen 7 en 8, wat ook weer overeenkomt met de standaard die bepaald is voor 2F. Deze resultaten geven vertrouwen in de hier bepaalde standaarden.

Als afsluiting van de standaardbepaling is de deelnemers aan het expertpanel gevraagd naar hun bevingen over deze procedure. Deelnemers ervaarden de bijeenkomsten als zinvol door de opbouw van theorie (Referentieniveaus scherp krijgen) naar de relatie met eigen praktijk (leerlingen in eigen organisaties of eigen expertise rond dit thema). Daarnaast werd een grote meerwaarde gezien in de uitwisseling van de onderbouwing van de individueel bepaalde scorepunten. Door van globaal naar detail te gaan kon in relatief korte tijd diepgang bereikt worden. Uitwisseling onder leiding van een onafhankelijke gespreksleider en onderzoeksbegeleider werd belangrijk gevonden om objectief te blijven. De deelnemers gaven aan dat de bijeenkomsten voor henzelf en hun organisaties verrijkend zijn geweest en pleitten dan ook voor beschikbaar maken van voorbeeldmaterialen.

Alles meegenomen is er een bruikbare procedure ontwikkeld die ook zeer geschikt lijkt voor een vervolgonderzoek in het SBO en voor andere standaardbepalingen op het gebied van de mondelinge taalvaardigheid.

Literatuur

Keuning, J., Straat, H., and R.C.W. Feskens (2017). The Data-Driven Direct Consensus (3DC) Procedure: A New Approach to Standard Bepaling. In *Standard Bepaling – International State of Research and Practices in the Nordic Countries* (Eds: Sigrid Blömeke and Jan-Eric Gustafsson). Springer series Methodology of Educational Measurement and Assessment, pp 263-278).