Taalscreening bij aanvang van het leerplichtonderwijs

Wetenschappelijk Eindrapport

Auteurs: Jozefien Loman, Goedele Vandommele, Sofie Vandoninck en Eva Koch
Academisch promotor: Kris Van den Branden

Vlaanderen
is onderwijs & vorming

KU LEUVEN CTO
CENTRUM VOOR TAAL & ONDERWIJS

TAALSCREENING BIJ AANVANG VAN HET LEERPLICHTONDERWIJS

Wetenschappelijk Eindrapport

Auteurs

Fien Loman, Goedele Vandommele, Sofie Vandoninck en Eva Koch

Academisch promotor

Kris Van den Branden

INHOUD 0

VOORWOORD 7

INLEIDING 8

1 ▪ Achtergrond 9

1.1 Doel en probleemstelling 9
1.2 Onderzoeksvragen 10
1.3 Fasen in het onderzoek 11
1.4 Tijdspad 11

2 ▪ Overzicht onderzoeksstappen 12

2.1 Ontwikkelen referentiekader, toetsmatrijs en toetstaken, vertrekkend vanuit de bestaande itembank 12
2.2 Literatuurstudie en expertenbevraging 13
2.3 Vooronderzoek 13
2.4 Pilootonderzoek 13
2.5 Steekproeftrekking kalibratie-onderzoek 14
2.6 Kalibratie-onderzoek 14
2.7 Cesuurbepaling 15
2.8 Definitief instrument 16

HOOFDSTUK 1: TOETSCONSTRUCT 17

1 ▪ Focus op luistervaardigheid 18

2 ▪ Referentiekader 19

2.1 Doelstellingen 19
2.2 Toetsmatrijs 19
2.3 Variatie in complexiteit 22
2.4 Afstemming taalgebruik op de doelgroep 23

3 ▪ Typetaken in KOALA 25

3.1 Doe- en zoekopdrachten op papier of tablet (zoek-opdrachten) 25
3.2 Meerkeuze-opdrachten op papier 27
3.3 Gerichte opdrachten met gestandaardiseerde observatie (doe-opdrachten) 28
3.4 Bredere Observatie en (taal)Stimulering: inbedding van KOALA 30
3.5 Beeldvorming door de leraar tijdens de klaswerking 31
3.6 Gerichte beeldvorming door afname van KOALA 32
3.7 Verbreding en verdieping van de beeldvorming door lerarenteam 32

HOOFDSTUK 2: LITERATUURSTUDIE EN EXPERTENBEVRAGING 33

1 ▪ Methodologie 35

1.1 Methodologie van de expertenbevraging 35

1.2 Methodologie van de literatuurstudie 37
1.3 Rapportage 37

2 ▪ Resultaten 38

2.1 Vier criteria voor het taalscreeningsinstrument voor kleuters 38
2.2 Belang van aansluiting bij de visie op kleuteronderwijs en op evalueren binnen kleuteronderwijs 39
2.3 Brede beeldvorming en inbedding van de taalscreening 40
2.4 Afnamecondities 41
2.5 Het screeningsinstrument 42
2.6 Interpretatie van de resultaten van de screening 44
2.7 Implementatie en gebruik van een screeningsinstrument 44
2.8 Informatie voor de gebruiker 45

HOOFDSTUK 3: VOORONDERZOEK 47

1 ▪ Werkwijze 49

1.1 Participanten 49
1.2 Taken in het vooronderzoek 49

2 ▪ Bevindingen 50

HOOFDSTUK 4: PILOOTONDERZOEK 52

1 ▪ Selectie van de deelnemers 54

2 ▪ Afnamemodaliteiten 55

3 ▪ Samenstelling van overlappende clusters 56

4 ▪ Taken in het pilootonderzoek 58

4.1 Context 58
4.2 Soorten taken in het pilootonderzoek 58
4.3 Aantal taken en items 59

5 ▪ Bevindingen uit kwalitatieve observaties 60

5.1 Afnamecondities 60
5.2 Afkijken 60
5.3 Afnamemodaliteit: op papier versus digitaal 61
5.4 Instructies voor toetsafnemers 62
5.5 Haalbaarheid van typetaken 62

6 ▪ Bevindingen uit de kwantitatieve analyse 67

6.1 Databestand 67
6.2 Betrouwbaarheid van het screeninginstrument als geheel 67
6.3 Betrouwbaarheid van de metingen op het niveau van de kleuters 68
6.4 De verhouding tussen de moeilijkheidsgraad van de toets en de vaardigheid van de kleuters (Wright Map) 69

6.5 Moeilijkheidsgraad van de verschillende items 71
6.6 Vaardigheid van de kleuters 74

7 ▪ Conclusies Pilootonderzoek 77

7.1 Betrouwbaarheid van de taalscreening als geheel 77
7.2 Bereik van de taalscreening 77
7.3 Betrouwbaarheid van de items 78
7.4 Herwerking 78

HOOFDSTUK 5: KALIBRATIEONDERZOEK 79

1 ▪ Steekproef Kalibratieonderzoek 80

1.1 Methode: gestratificeerde steekproef 80
1.2 Vooropgestelde steekproef 82
1.3 Deelnemende scholen 83
1.4 Leerlingen in de steekproef 85

2 ▪ Dataverzameling in coronatijden 85

2.1 Respons 85
2.2 Toetsassistenten 85
2.3 Opleiding van toetsafnemers en toetsassistenten 86
2.4 Verdeling toetspakketten 86
2.5 Uitval van scholen, toetsafnemers en kleuters 87

3 ▪ Instrument voor het Kalibratie-onderzoek 88

4 ▪ Afnameprocedure 88

4.1 Overlappende clusters 88
4.2 Beoogd aantal prestaties 90
4.3 Codering van de kleuters 91
4.4 Afname 91
4.5 Standaardisatie via toetsassistenten 91
4.6 Documenten per cluster 93
4.7 Papieren afname versus digitale afname 100

HOOFDSTUK 6: RESULTATEN KALIBRATIEONDERZOEK 101

1 ▪ Kwantitatief: Statistische Analyses 102

1.1 Samenstelling van het databestand 102
1.2 Dataset van betrouwbare toetsitems 103
1.3 Betrouwbaarheid van het screeningsinstrument als geheel 103
1.4 Betrouwbaarheid van de metingen op het niveau van de kleuters 104
1.5 De verhouding tussen de moeilijkheidsgraad van de screening en de vaardigheid van de kleuters (Wright Map) 105
1.6 Moeilijkheidsgraad van de verschillende items 108
1.7 Relatie tussen moeilijkheidsgraad van de items en eigenschappen van de items 110

1.8 Vaardigheid van de kleuters (n=1955) 111
1.9 Itembias : Differential item functioning (dif-analysis) 115
1.10 Relatie tussen de vaardigheid van kleuters en hun persoonskenmerken 121
1.11 Inschatting taalvaardigheid door de leraar 129
1.12 Voorspellers van luistervaardigheid op school- en kindniveau (multilevel analyses) 139

2 ▪ Kwalitatief: analyse van de verzamelde feedback 144

2.1 Kanalen voor het verzamelen van de feedback 144
2.2 Samenvatting van de feedback 145
2.3 Samenvattende conclusies uit het kalibratieonderzoek 155

HOOFDSTUK 7: CESUREN BIJ KOALA 157

1 ▪ Cesuurbepaling met cesuurcommissie 158

2 ▪ Procedure 160

2.1 Contact en communicatie met de cesuurcommissie 160
2.2 Selectie van items voor de cesuurbepaling 160
2.3 Twee cesuren, drie groepen 162

3 ▪ Cesuurbepaling Dag 1 165

3.1 Deel 1: In groep 165
3.2 Deel 2: Individueel 167

4 ▪ Cesuurbepaling Dag 2 168

4.1 Deel 1: terugkoppeling na ronde 1 168
4.2 Deel 2: Tweede ronde 176

5 ▪ Resultaat (Finale Cesuur) 181

5.1 Presentatie van de cesuren na de derde ronde 181
5.2 Presentatie van de impactdata op kindniveau na de derde ronde 182
5.3 Presentatie van de impactdata op schoolniveau (OKI-categorie van de school) na de derde ronde 183
5.4 Presentatie van de impactdata voor OKI-indicatoren na de derde ronde 183

6 ▪ Conclusie 185

HOOFDSTUK 8: GEVOLGEN VAN DE VERSCHILLENDE ONDERZOEKSSTAPPEN VOOR TOETSONTWIKKELING 186

1 ▪ Vertrekpunt: de SALTO-toetsbatterij 188

1.1 Screening van de psychometrische informatie op itemniveau 188
1.2 Inhoudelijke screening van de items 189

2 ▪ Hergebruikte SALTO-taken 190

2.1 Feedback op geselecteerde SALTO-taken 190

2.2 Vaststellingen uit het vooronderzoek 191
2.3 Conclusie 192
2.4 Aanpassingen toetsconstruct, toetsmatrijs en typetaken 193
2.5 Herwerking van SALTO-taken 193
2.6 Nieuwe taken, geschikt voor kleuters 194
2.7 Taken verwijderen 194
2.8 Conclusie 194

3 ▪ Tweede test-en ontwikkelronde: pilootonderzoek 196

3.1 Vaststellingen uit het pilootonderzoek 196
3.2 Herwerking na het pilootonderzoek 197
3.3 Concrete voorbeelden van herwerking 198

4 ▪ Instrument voor Kalibratie-Onderzoek 200

4.1 Definitie van taken en items 200
4.2 Typetaken in het kalibratieonderzoek 200
4.3 Aantal taken en items 200
4.4 Papieren en digitale versie 201

5 ▪ Definitief instrument 202

HOOFDSTUK 9: DEFINITIEVE INSTRUMENT 203

1 ▪ Screening Items kalibratie-onderzoek 204

1.1 Psychometrische screening 204
1.2 Inhoudelijke screening 208
1.3 Vormelijke screening 209

2 ▪ Selectie Voor KOALA 211

2.1 Deel A 211
2.2 Deel B 212

3 ▪ Controle van de selectie 214

3.1 Psychometrische criteria 215
3.2 Inhoudelijke criteria 218
3.3 Vormelijke criteria 218

4 ▪ Aangepaste versie voor kleuters met grote ondersteuningsnoden 219

4.1 Plaatsen van het afbreekpunt 219
4.2 Extra toetsitems voor Deel B* 220

5 ▪ Controle van de selectie voor de aangepaste versie 222

5.1 Psychometrische criteria 222
5.2 Inhoudelijke criteria 224
5.3 Vormelijke criteria 225

6 ▪ Besluit 226

HOOFDSTUK 10: BELEIDSAANBEVELINGEN 227

LITERATUURLIJST 230

BIJLAGEN 237

Bijlage 1: Scenario voor gesprekken met experten 238

Bijlage 2: Overzicht aanpassingen items Pilootonderzoek 240

BIJLAGE 3: ANONIEM SCHOOLFEEDBACKRAPPORT 249

Bijlage 4: Vergelijking steekproeftrekking taalscreening versus reële set voor dataverzameling 276

Bijlage 5 : Leden van de cesuurcommissie 284

VOORWOORD

Het onderzoek naar de taalscreening werd uitgevoerd in opdracht van het Vlaams Ministerie van Onderwijs en Vorming.

Het onderzoek naar en de ontwikkeling van KOALA kwam tot stand dankzij de hulp en ondersteuning van vele mensen.

Allereerst danken wij de vele scholen en kleuters die aan de verschillende onderzoeks- en ontwikkelfasen hebben deelgenomen.

Een aantal van onze CTO-collega's hielp bij het ontwikkelen van eerste versies van de taken: Saartje Gobyn en Ellen Smits.

We konden voor de toetsafnames rekenen op de hulp van vele toetsassistenten: pedagogisch begeleiders, CTO-medewerkers en studenten van Arteveldehogeschool en UCLL. Sofie Robyns, Katrien Van Rysseghem en Cindy Leman leverden een belangrijke bijdrage als vrijwillig medewerkers.

Jobstudenten stonden mee in voor de verwerking van de data. Voor de praktische organisatie werden we ondersteund door de secretariaatsmedewerker van het Centrum voor Taal en Onderwijs en jobstudente Margot Saeyens.

We danken de stuurgroep en de medewerkers van het Departement Onderwijs voor hun begeleiding.

Onze resonansgroep gaf erg gewaardeerde suggesties en feedback gedurende het onderzoek en de ontwikkeling van de toets. Volgende mensen willen we daarvoor van harte danken: Sofie Robyns, Nadia Dewaele, Marlies Algoet, Carolien Frijns, Ellen Smits, Alida Pierards, Kristien Coussement, Karen Van Renterghem, Katrijn Denies, Sven De Maeyer, Bart Masquillier, Frederik Vanackere, Liv Camps, Lotje De Spiegeleer en Karen Dehaen.

Dank ook aan Tim van Stubio Tibo die de vele tekeningen bij KOALA maakte en Bezig Bieke die de lay-out verzorgde. Simon (Studio Monk) en Vincent en Kenneth (Tomatojuice) maakten de filmpjes.

Het onderzoeks- en ontwikkelteam,

Leuven, juni 2021

INLEIDING

1 ▪ Achtergrond

1.1 Doel en probleemstelling

Dit onderzoeksproject gaat na in welke mate de itembank die ten grondslag lag aan de ontwikkeling van het Screeningsinstrument Aanvang Lager Onderwijs Taalvaardigheid (SALTO) op een valide en betrouwbare manier kan herbestemd worden om de taalvaardigheid Nederlands van kleuters aan het begin van de derde kleuterklas van het Vlaams kleuteronderwijs te screenen en daarbij risicoleerlingen te detecteren.

Leerlingen die het lager onderwijs instromen met een onvoldoende schoolse taalvaardigheid Nederlands, lopen een verhoogd risico op leerachterstand. Uit onderzoek (OECD, 2004; peiling Nederlands luisteren, 2019) blijkt dat leerlingen met een lage SES-status en bovendien als thuistaal niet het Nederlands hebben, meer kans hebben om zwakker te presteren op taken die schoolse taalvaardigheid in het Nederlands vragen, en als gevolg daarvan vaker minder goed presteren in het onderwijs. Het is dus van belang om die leerlingen tijdig op te sporen en zo nodig extra stimulansen te geven om hun taalvaardigheid Nederlands te ontwikkelen.

Momenteel biedt het Departement Onderwijs en Vorming van de Vlaamse Gemeenschap reeds een screeningsinstrument voor taalvaardigheid Nederlands aan dat aan het begin van het eerste leerjaar wordt afgenomen (SALTO)[1]. SALTO werd ontwikkeld, gevalideerd en van cesuren voorzien door het Centrum voor Taal en Onderwijs (Faculteit Letteren KU Leuven) op basis van een grootschalig kalibratie-onderzoek, en wordt op dit moment nog gebruikt door Nederlandstalige basisscholen van het Vlaams en Brussels Hoofdstedelijk Gewest om de taalvaardigheid Nederlands van leerlingen aan het begin van het eerste leerjaar te screenen. Een vroegere meting, bijvoorbeeld aan het begin van de derde kleuterklas, zou het mogelijk maken om sneller risicoleerlingen te detecteren en bij hen tekorten weg te werken die de kans op succes in het eerste leerjaar, en in het verder lager onderwijs, hypothekeren. De vervroeging van de leerplicht tot 5 jaar kan gerichte interventies na deze vroegere screening faciliteren.

Het screenen van 5-jarige kleuters is echter niet vanzelfsprekend. Zoals de studie naar de wenselijkheid en haalbaarheid van centrale taaltoetsen (Colpin e.a., 2006) uitwees, verhoogt bij het screenen van kinderen jonger dan 6 jaar het risico op onbetrouwbare metingen. Kleuters kunnen zich minder lang concentreren als leerlingen van het lager onderwijs, zijn wispelturiger in hun prestaties, sneller afgeleid, en sterk onderhevig aan de omstandigheden waarin de screening wordt afgenomen. Om die factoren beter in kaart te brengen, voeren we ook een (beperkte)

[1] https://onderwijs.vlaanderen.be/nl/screeningsinstrument-aanvang-lager-onderwijs-taalvaardigheid

literatuurstudie uit en informeren we ons via gesprekken met experten. Daarnaast is het van groot belang dat een grootschalig kalibratie-onderzoek vaststelt of de afnames van bepaalde toetsitems tot voldoende betrouwbare metingen kunnen leiden, en onder welke omstandigheden dat het best het geval is. Een stuurgroep en resonansgroep (met een brede vertegenwoordiging van experten taal en toetsonwikkeling, beleidsmakers, onderwijsverstrekkers, lerarenopleiders en leraren kleuteronderwijs) volgen het onderzoeksproject mee op.

De diagnostische kracht van een gestandaardiseerd screeningsinstrument is beperkt (Colpin e.a. 2006). Voor een volwaardige diagnose van problemen op het vlak van taalvaardigheid dient de screening te worden opgevolgd door een interactieve, doorgedreven en gecontextualiseerde probleemanalyse. Met andere woorden, indien wordt vastgesteld dat een leerling onder een van de cesuren scoort, moet deze informatie samengelegd worden met observaties rond het (talig) functioneren (mondeling en schriftelijk, receptief en productief) van de leerling en dient in een eerste fase een verdere diagnose van de taalcompetentie Nederlands van deze leerling te worden uitgevoerd. Dit gebeurt bij voorkeur op basis van een voldoende aantal observaties tijdens een variatie aan activiteiten die deel uitmaken van de reguliere klasdag. Zulk een diagnose kan ervoor zorgen dat de extra taalondersteuning nadien ingevuld wordt op maat van de leerling. We nemen deze aandachtspunten mee bij het ontwikkelen van het instrument en de bijhorende handleiding.

Het screeningsinstrument KOALA kan vanaf schooljaar 2021-2022 worden ingezet bij het begin van de derde kleuterklas. Die taalscreening zal toelaten risicokleuters op te sporen en om (1) hen na de screening breder te observeren in functie van meer gedetailleerde analyse van hun taalleerbehoeften en (2) hen extra ondersteuning en taalstimulering te bieden bij de ontwikkeling van hun taalvaardigheid Nederlands.

1.2 Onderzoeksvragen

Dit onderzoek wil een antwoord geven op volgende vragen:

- RQ1 Is het mogelijk de taalvaardigheid van 5-jarige kleuters (dus bij de start van het leerplichtonderwijs) op een betrouwbare en valide manier te meten?
- RQ2 Vormt het bestaande screeningsinstrument voor het eerste leerjaar (SALTO) daarvoor een goede basis?
- RQ3 Welke aanpassingen zijn nodig om gebruik bij 5-jarige kleuters mogelijk te maken?

1.3 Fasen in het onderzoek

In de ontwikkeling van een gestandaardiseerd screeningsinstrument (voor taalvaardigheid) dat voorzien wordt van cesuren, kunnen we de volgende onderdelen onderscheiden (Alderson, Clapham, & Wall, 1995; Manual A.L.T.E. , 2011):

- ontwikkelen referentiekader en toetsmatrijs;
- ontwikkeling eerste versie toetstaken;
- pilootonderzoek + bijsturing van toetstaken;
- steekproeftrekking, contactname scholen en voorbereiding dataverzameling;
- instructiefase voor toetsafnemers en communicatie met de scholen;
- dataverzameling voor kalibratie-onderzoek;
- data-analyse;
- cesuurbepaling;
- afwerking instrument, handleiding;
- communicatie met afnemend veld en verspreiding.

Aangezien in dit onderzoek een bestaande itembank wordt ingezet voor een nieuwe doelgroep, kunnen we voor de invulling van de eerste drie onderdelen grotendeels vertrekken van de inzichten uit het SALTO-onderzoek, mits een focus op de kenmerken en de context van de nieuwe doelgroep.

1.4 Tijdspad

2 ▪ Overzicht onderzoeksstappen

Hieronder geven we een samenvatting van gezette onderzoeksstappen. In de volgende hoofdstukken worden die stappen meer in detail toegelicht.

2.1 Ontwikkelen referentiekader, toetsmatrijs en toetstaken, vertrekkend vanuit de bestaande itembank

Dit onderzoek is gebaseerd op het referentiekader, de toetsmatrijs en de batterij toetstaken die werden ontwikkeld voor de ontwikkeling van SALTO. De huidige ontwikkelingsdoelen taal (Nederlands) en het *Referentiekader vroege tweedetaalverwerving* (Nederlandse Taalunie, 2001) zijn daarbij richtinggevend.

We vertrokken van de toetsmatrijs en batterij van 21 toetstaken die voor SALTO werd ontwikkeld, en de bijbehorende parameters van complexiteit. Inhoudelijk sluiten de toetstaken van SALTO redelijk goed aan bij de doelstellingen van het huidige onderzoek. Deze toetstaken zijn immers (1) geënt op de interesses van 5 à 6-jarige kleuters en (2) afgestemd op de ontwikkelingsdoelen taal voor het kleuteronderwijs (gesprekspartner, tekstsoort, verwerkingsniveau). De 21 toetstaken variëren qua complexiteit. Alle toetstaken van SALTO werden door de ontwikkelaars en door de resonansgroep kritisch bekeken met het oog op de nieuwe doelstelling en doelgroep: zowel de psychometrische waarden (fit, moeilijkheidsgraad) uit het vorige kalibratie-onderzoek als de inhoud en vormgeving werden gescreend. De nodige aanpassingen (bv. taalgebruik of aansluiting met de context van een kleuterklas) werden gedaan.

Omdat voor de uiteindelijke screening waarschijnlijk een 8-tal taken zouden volstaan (analoog met SALTO), bevatte de batterij voor het kalibratie-onderzoek, bewust te veel taken. Dit schept de mogelijkheid om slecht fittende items of te moeilijke items te verwijderen. Om de bestaande batterij aan te vullen, werden nieuwe taken ontwikkeld. Deze nieuw ontwikkelde taken hebben het voordeel dat de aansluiting met het taalgebruik en context van de kleuterklas van bij aanvang werd meegenomen.

2.2 Literatuurstudie en expertenbevraging

Een screeningsinstrument ontwikkelen dat op een betrouwbare en valide manier taalvaardigheid van kleuters in kaart te brengt, brengt heel wat uitdagingen met zich mee. Via een literatuurstudie en gesprekken met experten brachten we de huidige wetenschappelijke inzichten rond deze uitdagingen in kaart. De literatuurstudie en expertenbevraging gebeurden in verschillende fasen tijdens het onderzoek. De inzichten werden immers steeds verfijnd en verder afgetoetst naar gelang de noden en uitdagingen van het onderzoek.

2.3 Vooronderzoek

De aangepaste taken uit de SALTO-batterij en de nieuwe taken werden uitgetest bij 5-jarige kleuters om na te gaan of ze voldoende afgestemd zijn op de leefwereld en het ontwikkelingsproces van de doelgroep, en of instructies, tekeningen, antwoorden eenduidig worden begrepen. Een individuele afname met interactie over de antwoordkeuze van een kleuter biedt heel wat inzichten in de manier waarop de instructies (beter) kunnen worden geformuleerd, de manier waarop kleuters afbeeldingen interpreteren…

We namen ook op 1 school de opdrachten af bij groepjes van 5-jarige kleuters. Deze afname gaf o.a. informatie over de afnamecondities voor het pilootonderzoek.

2.4 Pilootonderzoek

Het pilootonderzoek vormde een belangrijke stap in het onderzoek waarbij de toetstaken voor het eerst aangeboden werden aan een groep die een jaar jonger is dan de SALTO-doelgroep. De 30 toetstaken werden daarom in overlappende sets van taken verdeeld. Elke set werd aan een 30-tal kleuters van 5 jaar aangeboden.

Dit pilootonderzoek verschafte informatie over de moeilijkheidsgraad van de items, de duidelijkheid van de instructies en de haalbaarheid van toetsing bij jonge kinderen. De items werden ook voorgelegd aan de leraren van de betrokken kleuters en aan een resonansgroep (die bestaat uit stakeholders, zoals pedagogische begeleiders en lerarenopleiders).

2.5 Steekproeftrekking kalibratie-onderzoek

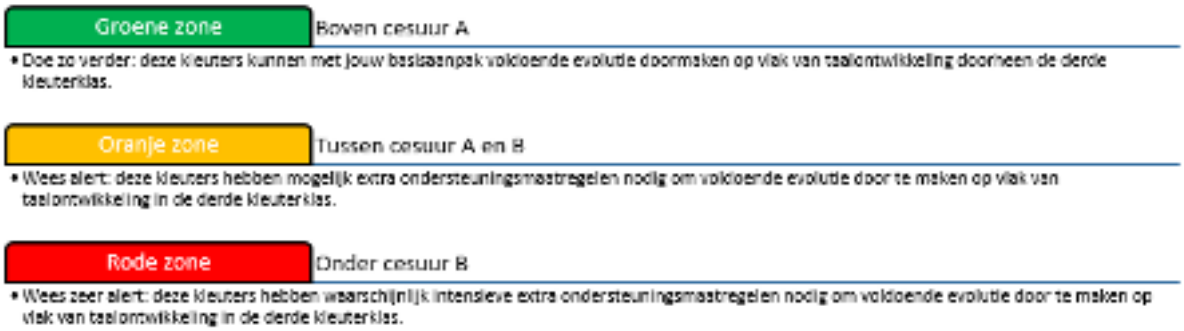
Voor de steekproef selecteerden we kleuters uit een populatie van kleuterscholen met minstens 25% SES-doelgroepleerlingen. De steekproeftrekking hanteerde een gestratificeerd steekproefdesign, met onderwijsnet en provincie als gebruikte stratificatievariabelen. Dit resulteerde in een steekproef van ongeveer 2000 kleuters, met een proportionele verdeling over de netten en een voldoende vertegenwoordiging van Nederlandstalige kleuterscholen uit het Brussels Hoofdstedelijk Gewest.

2.6 Kalibratie-onderzoek

Via een grootschalig kalibratie-onderzoek werd vastgesteld of de afnames van bepaalde toetsitems tot voldoende betrouwbare metingen kan leiden, en onder welke omstandigheden dat het best het geval is. De data-analyse van het kalibratieonderzoek maakte het mogelijk slecht fittende items te verwijderen. De resultaten van de dataverzameling werden geanalyseerd via een multidimensionele IRT-gebaseerde Raschanalyse. IRT-gebaseerde analyse laat toe een objectieve moeilijkheidsgraad van de toetstaken en -items te bepalen, onafhankelijk van de vaardigheid van de groep getoetste leerlingen, door het kalibreren van de toetsitems volgens een latente schaal die zowel de moeilijkheidsgraad van het toetsitem weergeeft als de vaardigheid van de leerling nodig om het toetsitem goed op te lossen. Er werd, naast betrouwbaarheids- en validiteitsanalyses, ook een biasonderzoek uitgevoerd om te bepalen of bepaalde toetsitems bepaalde kleuters (bv. met een bepaalde gezinstaal of van een bepaalde achtergrond) bevoordelen of benadelen. Daarnaast gingen we na welke achtergrondkenmerken op kind- en schoolniveau de luistervaardigheid van een kleuter konden voorspellen.

2.7 Cesuurbepaling

Voor de bepaling van de cesuren werd de Bookmarkmethode gehanteerd. Een cesuurcommissie van 19 leden met een brede vertegenwoordiging van het onderwijsveld kwam in verschillende rondes tot twee voorstellen voor cesuren (één die kinderen identificeert die net onder de ondergrens scoren en één die kinderen identificeert die sterk onder de ondergrens scoren). Het is de bedoeling dat op basis van die twee cesuren drie groepen van leerlingen kunnen worden onderscheiden:



Figuur 1: Beschrijving van de groepen en cesuren

2.8 Definitief instrument

Om het definitief instrument te kunnen samenstellen, vertrokken we van de 148 ontwikkelde items. We keken naar de psychometrische analyses en sloten op die manier 5 items met misfit uit. Van de overige 143 items bekeken we DIF, discriminerende waarde en de verdeling over de vaardigheidsniveaus. Daarna bestudeerden we ook inhoudelijke criteria zoals de verdeling over doelstellingen en over typetaken. Ook deden we een vormelijke screening. Op basis van deze selectiecriteria kwamen we tot 33 items voor het definitieve screeningsinstrument. Omwille van praktische redenen werden deze items verdeeld over twee delen - een deel voor individuele afname en een deel voor afname in groep.

Tot slot onderzochten we de mogelijkheid om snel kleuters te detecteren die erg zwak scoren om hen niet het volledige screeningsinstrument aan te bieden, maar snel te leiden naar een versie die beter afgestemd is op hun lagere luistervaardigheid. We bepaalden hiervoor een afbreekpunt en selecteerden items voor een gemakkelijkere versie, waarvan opnieuw alle items tegemoet kwamen aan de vooropgestelde selectiecriteria.

HOOFDSTUK 1: TOETSCONSTRUCT

1 ▪ Focus op luistervaardigheid

Net als bij het reeds bestaande SALTO-instrument voor het eerste leerjaar, is het de bedoeling om de luistervaardigheid Nederlands in kaart te brengen, in dit geval van 5-jarige kleuters.

Schriftelijke vaardigheden zijn uiteraard niet aan de orde, hoewel sommige taken aanleiding geven om ook een (beperkte) blik te werpen op het proces van (ontluikende) geletterdheid van de kleuters en aanleiding geven tot een bredere observatie ervan (zie 3.4 Bredere Observatie en (taal)Stimulering: inbedding van KOALA). Spreekvaardigheid toetsen is bijzonder tijds- en arbeidsintensief: de interacties met individuele kleuters moeten dan achteraf met een uitgebreide scorewijzer betrouwbaar worden beoordeeld. Dat is noch voor dit ontwikkelingsproject, noch voor later gebruik in het kleuteronderwijs haalbaar.

Om deze reden wordt gekozen voor een focus op luistervaardigheid: luistervaardigheid blijkt een zeer goede voorspeller voor schoolsucces in het eerste leerjaar (Colpin et al., 2006), en is ook vanuit praktisch oogpunt haalbaar.

In lijn met de ontwikkelingsdoelen en andere referentiekaders[2] voor taaldoelen in de kleuterklas, staat de vaardigheid van kleuters centraal om

a) mondelinge instructies te begrijpen;
b) korte mondelinge informatieve mededelingen te begrijpen;
c) korte mondelinge narratieve boodschappen (bv. een kort verhaaltje) te begrijpen;
d) mondelinge vragen te begrijpen.

Bij het screenen van deze 4 luistervaardigheidsdoelen wordt heel wat schoolse woordenschat geïntegreerd die kleuters aan het einde van de derde kleuterklas best kennen om goed te kunnen starten in eerste leerjaar.

[2] Zoals *Referentiekader vroege tweedetaalverwerving* (Nederlandse Taalunie, 2001), kerndoelen en tussendoelen (www.slo.nl)

2 ▪ Referentiekader

2.1 Doelstellingen

We baseren ons in de eerste plaats op de ontwikkelingsdoelen kleuteronderwijs en het *Referentiekader vroege tweedetaalverwerving* (Nederlandse Taalunie, 2001) voor het ontwikkelen van het referentiekader voor deze taalscreening. Ook volgen we de ontwikkeling van de nieuwe eindtermen/ontwikkelingsdoelen kleuteronderwijs op. Deze doelenkaders vullen we aan met een complexiteitsmeter om de moeilijkheidsgraad van taken te definiëren.

Het *Referentiekader vroege tweedetaalverwerving* is gepubliceerd in 2001. Uit navraag bij de Taalunie naar de mate waarin het nog up-to-date is en of er plannen voor vernieuwing zijn blijkt het kader nog steeds relevant. Plannen voor vernieuwing zijn er niet omdat er momenteel geen vraag naar update voorligt vanuit het onderwijsveld. De huidige leerplannen sluiten nog steeds zeer nauw aan bij de doelen uit dit referentiekader.

2.2 Toetsmatrijs

Een toetsmatrijs geeft een overzicht van wat je wil toetsen en vormt op die manier een hulpmiddel om valide te toetsen: op die manier helpt de toetsmatrijs een toetsontwikkelaar om concreet weer te geven wat een toets inhoudelijk wil toetsen. Een toetsmatrijs zorgt er dus voor dat de taalscreening qua inhoud representatief is voor wat we bij kleuters bedoelen met 'luistervaardigheid'.

Bij het opstellen van de toetsmatrijs voor deze screening zijn de cruciale vragen: Is de taalvaardigheid van 5-jarige kleuters hoog genoeg om goed functioneren en participeren in het onderwijs (met name in de derde kleuterklas, en later in het eerste leerjaar) mogelijk te maken? Welke kleuters hebben extra taalstimulering nodig wanneer dat niet het geval is?

De toetsmatrijs geeft aan welk soort luistertaken worden opgenomen om deze vragen te kunnen beantwoorden. De taken zijn een combinatie van (aangepaste) items uit SALTO en nieuw ontwikkelde items. Voor de verschillende doelstellingen wordt het teksttype en de handeling weergegeven. In de toetsmatrijs staat ook informatie over de soort opdracht en het aantal toetsitems. Bij het selecteren en ontwikkelen van taken wordt getracht om in de mate van het mogelijke de ontwikkelingsdoelen en het referentiekader te dekken. Aangezien het instrument niet

bedoeld is om een diagnose te stellen op het gebied van alle verschillende doelstellingen hoeven we echter niet per doelstelling een diepgaande meting na te streven.

In de toetsmatrijs geven we eveneens weer met hoeveel taken we de verschillende doelen in kaart brengen in het kalibratie-onderzoek en in het definitieve instrument.

De 30 taken die hier worden opgelijst zijn meer dan voldoende voor het bereiken van de doelstellingen van het uiteindelijke instrument: 6 à 7 taken bleken te volstaan om in het kader van een taalscreening voldoende valide en betrouwbare informatie over de taalvaardigheid van een 5-jarige kleuter te verkrijgen. Door aanvankelijk meer items en taken te ontwikkelen en uit te proberen dan vereist, bleef er ruimte om na het pilootonderzoek en na het kalibratie-onderzoek inhoudelijk minder goede of misfittende items te kunnen verwijderen.

Omschrijving taaltaken	Aantal taken	
	Kalibratie	Definitief
1. MONDELINGE INSTRUCTIE OF OPDRACHT BEGRIJPEN (EN ADEQUAAT REAGEREN) ***OD 1.4*** *De kleuters kunnen door de kleuteronderwijzer gegeven opdrachten, met betrekking tot activiteiten in de klas of op school, begrijpen*		
Concrete fysieke handeling: 1.1 Instructies voor een concrete fysieke handeling in het hier-en-nu, bestemd voor de leerling of een leeftijdsgenoot, begrijpen	4	11
Mentale handeling (of verbale) handeling: 1.2 Instructies voor een mentale (of talige handeling), bestemd voor de leerling of een leeftijdsgenoot, begrijpen	2	9
2. MONDELINGE VRAAG BEGRIJPEN (EN ADEQUAAT REAGEREN) ***OD 1.2*** *De kleuters kunnen voor hen bestemde vragen in concrete situaties begrijpen*		
Over intenties, voorkeuren: 2.1 Vragen naar intenties, interesses of voorkeuren begrijpen	5	3
Over persoonlijke ervaringen: 2.2 Open vragen over eigen ervaringen en belevingen begrijpen	2	
Over gevoelens van leerlingen of anderen: 2.3 Vragen over zijn gevoelens of van partners in de omgeving begrijpen[3]	2	
Over situaties, handelingen, voorwerpen: 2.4 Vragen over situaties, handelingen of voorwerpen in de concrete omgeving begrijpen	2	
3. EEN GESPROKEN VERHAAL BEGRIJPEN ***OD 1.5*** *De kleuters kunnen een beluisterd verhaal, bestemd voor hun leeftijdsgroep, begrijpen*		
3.1 Voor hem bestemd verhaal volgen en begrijpen	5	4
4. EEN INFORMATIEVE MEDEDELING BEGRIJPEN ***OD 1.3*** *De kleuters kunnen een mondelinge, voor hen bestemde boodschap, ondersteund door beeld en/of geluid, begrijpen*		

[3] Ook in 3 'gesproken verhalen begrijpen' zitten vragen over 'gevoelens'.

Over concrete gebeurtenissen, feiten, hier-en-nu en over concrete gebeurtenissen, feiten, daar- en -toen: *OD 1.1 De kleuters kunnen een mondelinge boodschap, eventueel ondersteund door gebaar, mimiek met betrekking tot een concrete situatie begrijpen* 4.1 Informatieve mededelingen over concrete gebeurtenissen en feiten in het hier-en-nu en het daar-en-toen begrijpen	5	6
Over regels en voorschriften: 4.2 Mededelingen over regels en voorschriften in concrete situaties begrijpen	3	

Tabel 1: Toetsmatrijs met omschrijving taaltaken

2.3 Variatie in complexiteit

We streven naar een selectie of ontwikkeling van opdrachten die variëren qua moeilijkheidsgraad, onderwerp en taalvaardigheidseisen. Uit een analyse van de ontwikkelingsdoelen Nederlands voor het kleuteronderwijs, de eindtermen Nederlands voor het lager onderwijs en het *Referentiekader vroege tweedetaalverwerving (2001)*, leiden we de volgende parameters af die de moeilijkheidsgraad van de opdrachten mee bepalen:

EENVOUDIG → COMPLEX

Onderwerp	concreet	minder concreet	abstract	
Context	hier-en-nu		daar-en-toen	
Perspectief	vraagt geen inleving	beperkte inleving	veel inleving	
Verwerkingsniveau	beschrijvend - één element	beschrijvend - meer elementen	structurerend - één element	structurerend - meer elementen
Visuele ondersteuning	veel		beperkt	
Syntactisch	eenvoudig	minder eenvoudig	complex	
Lexicaal	alledaags	minder alledaags	schools	
Wiskundetaal - kwantitatief	niet aanwezig		wel aanwezig	
Wiskundetaal - ruimtelijk	niet aanwezig		wel aanwezig	

Figuur 2: Beschrijving complexiteit van items - aangepast voor kleuters

Niet alleen taken, maar ook items binnen taken kunnen verschillen qua moeilijkheidsgraad, en kunnen zich op deze continua verschillend situeren. De inhoudelijke context is binnen een taak dezelfde, maar daarbinnen kunnen items wel verschillende deelaspecten van de luisterdoelen meten én/of verschillen op bovenstaande moeilijkheidsparameters.

2.4 Afstemming taalgebruik op de doelgroep

Het taalgebruik in de opdrachten van de items uit SALTO (bijvoorbeeld woordenschat) is gebaseerd op de beschrijving van schooltaal in het onderzoek naar schoolse taalvaardigheidseisen (Schrooten, 1997) dat op het Centrum voor Taal en Onderwijs is uitgevoerd.

Omdat er geen recent Vlaamse onderzoek voorhanden is én de opdrachten haalbaar moeten zijn voor 5-jarigen (in plaats van voor 6-jarigen), zorgen we dat het taalgebruik voor KOALA is afgestemd door:

- gebruik te maken van contexten die lijken op de dagelijkse klascontext van een kleuter;
- veel gebruik te maken van dagelijkse taal en eenvoudige schooltaal die in betekenisvolle opdrachten aan bod komt en geduid wordt (de betekenis van woorden en meerwoordverbindingen is met andere woorden (ook) af te leiden uit de context en het niet begrijpen van één woord leidt niet noodzakelijk tot een fout antwoord);
- het gebruik van minder frequente en moeilijkere schoolse woorden te beperken: enkel het begrip van essentiële moeilijkere schoolse woorden (bv. belangrijke woorden in het proces van geletterdheid en gecijferdheid) kan noodzakelijk zijn voor het uitvoeren van bepaalde opdrachten.

3 ▪ Typetaken in KOALA

Om de verschillende doelstellingen uit de toetsmatrijs op een adequate manier te testen bij kleuters, onderscheiden we drie typetaken. In deze paragraaf beschrijven we deze typetaken in hun algemeenheid, met voor elke typetaak een concreet voorbeeld zoals opgenomen in het instrument voor het kalibratie-onderzoek.

3.1 Doe- en zoekopdrachten op papier of tablet (zoek-opdrachten)

3.1.1 Algemene beschrijving van doe-en zoekopdrachten

Doe- en zoekopdrachten op papier of tablet laten toe om instructies of het begrijpen van informatieve mededelingen op een directe manier te testen. De kleuters voeren een handeling uit op papier, of zoeken naar een beschreven voorwerp of persoon. Als antwoord kleuren kleuters iets in, trekken zij een lijn om een voorwerp op de correcte plaats te 'leggen', kleuren zij een voorwerp in een gevraagde kleur, zetten een kruisje op een voorwerp of persoon op een grote prent…

Omdat de antwoordmogelijkheden van doe-opdrachten (kleuren, lijnen trekken) moeilijker te realiseren zijn (technisch en voor kleuters zelf) op tablet, worden de doe-opdrachten ook in de digitale versie op papier afgelegd en het antwoord digitaal ingevoerd op de tablet door de afnemer. Van de zoekopdrachten is een digitaal equivalent, waarbij kleuters de geïdentificeerde persoon of het gevonden voorwerp aantikken.

Dit soort opdrachten zijn niet alleen geschikt voor het testen van bepaalde belangrijke doelstellingen voor kleuters; ze laten eveneens toe om op een snelle manier meerdere kleuters tegelijkertijd te testen. Bovendien bieden de doe- en zoekopdrachten op papier de mogelijkheid om buiten het hier-en-nu te treden en 'verdere' contexten en werelden te introduceren, zoals een herkenbare (maar niet bekende) eetzaal, kleuterklas of speelplaats.

In documenten voor de leraren verwijzen we naar de doe- en zoekopdrachten als 'zoek-opdrachten'.

3.1.2 Voorbeeld van doe- en zoekopdrachten

Voor een concrete doe-opdracht voeren kleuters een handeling uit op papier, zoals in het voorbeeld van 'Verjaardagsfeest'. In dit geval wordt de kleuter gevraagd om een glas te vullen met cola. Als antwoordmogelijkheid moet een kleuter het glas inkleuren tot het vol is.

Voorbeeld van een doe- opdracht op papier: verjaardagsfeest
*Op de tafel staan **vijf glazen cola**. Sommige glazen zijn leeg, andere zijn halfvol. Het **middelste** glas moet **helemaal vol** cola. Kleur het glas tot het helemaal vol is.*


Figuur 3: Voorbeeld van een doe-opdracht op papier: verjaardagsfeest

Bij een zoek-opdracht wordt aan kleuters gevraagd om een persoon of voorwerp te identificeren. In het geval van de Zandtafel gaat het over identificeren van personen die een bepaalde handeling uitvoeren, zoals lachen.

Voorbeeld zoek-opdracht op papier: zandtafel
Nikola heeft een flesje gevuld met water. Hij giet het uit en het water spettert op Anaïs. Het water is koud en daar moet Anaïs mee lachen. Waar is Anaïs? Zet een kruisje op Anaïs.


Figuur 4: Voorbeeld van een zoek-opdracht op papier: zandtafel

3.2 Meerkeuze-opdrachten op papier

3.2.1 Algemene beschrijving van meerkeuzeopdrachten (kies-opdrachten)

Deze meerkeuze-opdrachten vormen de meest typische 'toets'opdrachten. Kleuters luisteren naar een opdracht, verhaal of vraag en duiden hun antwoord aan door een kring te trekken rond één van de vier aangeboden prenten. In de digitale versie tikken de kleuters de juiste prent aan.

Via deze typetaak worden doelstellingen geïntroduceerd waarbij kleuters rekening moeten houden met uitgebreidere talige input en langere stukken informatie. Welgekozen afleiders tussen de antwoordmogelijkheden bieden de mogelijkheid om na te gaan of kleuters in staat zijn om meerdere talig aangeboden elementen te combineren, of gedetailleerde informatie op te pikken. Deze meerkeuze- opdrachten zijn bij uitstek geschikt om fantasiewerelden, of werelden die zich verderaf bevinden van de kleuters te introduceren en worden ook om die reden opgenomen.

In documenten voor de leraren verwijzen we naar de meerkeuzeopdrachten als 'kies-opdrachten'.

3.2.2 Voorbeeld van meerkeuze-opdrachten

Voor een meerkeuze-opdracht luisteren kleuters naar de input en duiden ze aan welke van de vier meerkeuzeopties het best hiermee overeenkomt. In het concrete voorbeeld krijgen de kleuters een beschrijving van de leesvoorkeur van een persoon en gaan ze na met welk boek deze voorkeur het beste zou overeenstemmen.

Voorbeeld van een meerkeuze-opdracht: lievelingsboeken
Judith houdt van boeken over de **natuur**. Ze luister het **liefst** naar verhalen die over **dieren** gaan. Van welk verhaal zou Judith het meest houden? Trek een kring rond het boek voor Judith.


Figuur 5: Voorbeeld van een meerkeuze-opdracht: lievelingsboeken

3.3 Gerichte opdrachten met gestandaardiseerde observatie (doe-opdrachten)

3.3.1 Algemene beschrijving van gerichte opdrachten met gestandaardiseerde observatie

Tijdens de gerichte opdrachten met gestandaardiseerde observatie worden kleuters geobserveerd terwijl zij eenvoudige taken uitvoeren, bv. een instructie uitvoeren, reageren op een vraag naar voorkeuren… Dit soort doelstellingen is moeilijk direct te testen via meerkeuze-items. In SALTO werden deze doelstellingen indirect getest (bv. duid aan wie de instructie correct uitvoert), wat echter veel inlevingsvermogen en perspectiefname vraagt van 5-jarige kleuters. We besloten daarom om in KOALA bovenvermelde doelstellingen ook te testen door observatie-opdrachten toe te voegen aan de toetsbatterij. In deze observatie-opdrachten voeren kleuters een fysieke handeling uit of reageren ze op een vraag naar voorkeuren en worden hun reacties gescoord. Door gerichte opdrachten met gestandaardiseerde observaties toe te voegen aan de toetsbatterij kunnen we een aantal doelstellingen op een directe wijze toetsen.

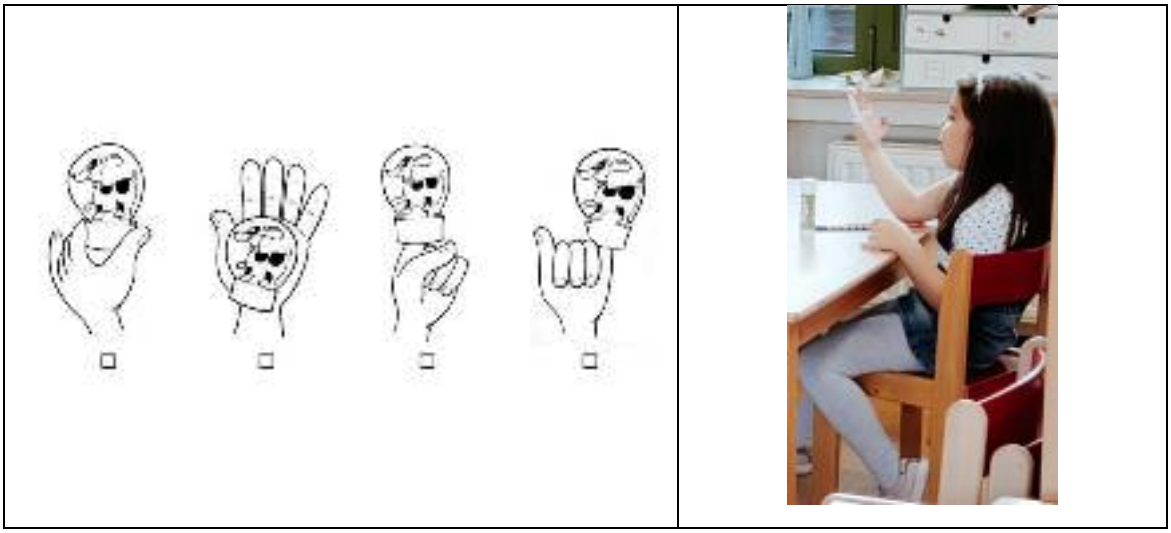
Andere voordelen van de gestandaardiseerde observatie-opdrachten is dat deze zorgen voor een actieve en eerder speelse betrokkenheid van de kleuter, afwisseling in toetsformat, aansluiting bij de leefwereld en context van de kleuter, en één-op-één contact met de bekende volwassene in de persoon van de leraar. Al deze elementen worden in de literatuur naar voren geschoven als belangrijke aandachtspunten voor het testen van jonge kinderen (Hasselgreen, 2000; Cameron, 2001).

In dit rapport benoemen we deze typetaak verder als ‘gestandaardiseerde observatie’. In documenten voor de leraren verwijzen we naar de gerichte opdrachten met gestandaardiseerde observatie als ‘doe-opdrachten’.

3.3.2 Voorbeeld van gerichte opdrachten met gestandaardiseerde observatie

Doorgaans gebeuren deze gerichte opdrachten met gestandaardiseerde observatie zonder papier of tablet, maar wordt concreet materiaal gebruikt dat voorhanden is in elke kleuterklas, zoals een popje, hoepel, knuffeldier ... De kleuters luisteren naar de input en gebruiken de materialen (lijmstift, papier) om de opdracht tot een goed einde te brengen.

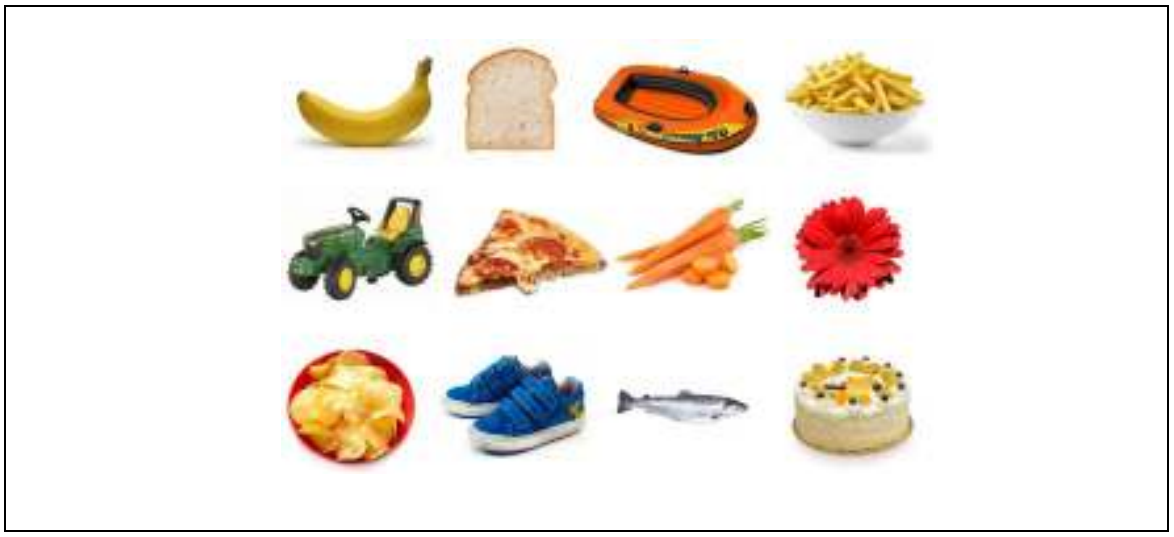
Voorbeeld van een gerichte opdracht met gestandaardiseerde observatie	
*Zet de vingerpop nu op je **pink**. Dat is je **kleinste vinger**.*	
Item in SALTO	Item in KOALA



Figuur 6: Voorbeeld van een gerichte opdracht met gestandaardiseerde observaties

In enkele gevallen, zoals in het voorbeeld in FIGUUR 7, worden illustraties gebruikt om kleuters te mogelijkheid te bieden om zonder verbale reactie hun begrip van de input duidelijk te maken. Op die manier kan een kleuter bijvoorbeeld het voorwerp van zijn/haar voorkeur aanduiden in plaats van het te moeten benoemen, zoals in het voorbeeld.

Voorbeeld van een gerichte opdracht met gestandaardiseerde observatie
Kijk goed naar de foto's. Welke dingen vind jij **héél lekker**? Wat lust jij héél graag? Duid maar aan of vertel maar wat jij lekker vindt.



Figuur 7: Voorbeeld van een gerichte opdracht met gestandaardiseerde observaties, met afbeeldingen ter ondersteuning

3.4 Bredere Observatie en (taal)Stimulering: inbedding van KOALA

Experten wijzen er veelvuldig op dat toetsen bij kleuters niet evident is en moet voldoen aan heel wat voorwaarden om tot betrouwbare resultaten te leiden. Ook onderzoek naar toetsen en evalueren van jonge kinderen bevestigt dit (zie Hoofdstuk 1: ToetsconstructHOOFDSTUK 2: LITERATUURSTUDIE EN EXPERTENBEVRAGING). Herhaaldelijk wordt door experten en onderzoekers gewezen op het belang van observeren en wordt geadviseerd breder te kijken dan enkel het meetinstrument en meetmoment.

Om een betrouwbaar beeld te geven van de luistervaardigheid van een kleuter, wordt KOALA gesitueerd in een breder kader van beeldvorming met meerdere bronnen van informatie en momenten van informatieverzameling. Het is op basis van de volledige vergaarde informatie dat schoolteams en leraren een goed beeld krijgen van waar een kleuter staat op vlak van luistervaardigheid (en de ruimere taalvaardigheid) en zij het best in staat zijn om gericht acties te ondernemen om een kleuter te ondersteunen, te stimuleren en de evolutie goed op te volgen.

De verschillende bronnen die in combinatie een breed en betrouwbaar beeld van de taalvaardigheid van de kleuters opleveren zijn:

- observaties van de leraar tijdens de dagelijkse interactie met kleuters;
- observaties van de leraar tijdens de afname van de taalscreening;
- analyse en interpretatie van de output gegenereerd door het taalscreeningsinstrument;
- aanvullende verbredende en verdiepende observaties van taalvaardigheid van de kleuter.

We stellen dit schematisch voor aan de hand van FIGUUR 8: Brede beeldvorming met KOALA:

Figuur 8: Brede beeldvorming met KOALA

Het screeningsinstrument luistervaardigheid (KOALA) - de toets dus - staat centraal in de brede beeldvorming, maar wordt bij voorkeur op meerdere momenten geflankeerd door andere (formele en informele) momenten van informatieverzameling over een kleuter. In de handleiding geven we aan hoe de gegevens uit de taalscreening en andere bronnen van informatie over de taalvaardigheid van kleuters op elkaar kunnen aansluiten.

Door de KOALA in te bedden in een breder geheel van handelingen van leraren draagt het geheel van KOALA en handleiding bij tot een betrouwbaar beeld van de taalvaardigheid van de kleuter (zie ook 3.4 BREDERE OBSERVATIE EN (TAAL)STIMULERING: INBEDDING VAN KOALA) én doet het recht aan deze professionaliteit van de kleuterleraar. De leraar heeft de regie in handen.

3.5 Beeldvorming door de leraar tijdens de klaswerking

Het screeningsinstrument KOALA staat niet op zichzelf, maar is een element dat de leraar helpt om een breed beeld te krijgen van de taalvaardigheid van een kleuter. De leraar heeft immers al heel wat indrukken opgedaan in verband met de taalvaardigheid van kleuters op basis van alledaagse interactie en de observatie tijdens het geven van instructies, het luisteren naar verhaaltjes, het vragen naar voorkeuren.

3.6 Gerichte beeldvorming door afname van KOALA

De afname van KOALA (het screeningsinstrument en de bijbehorende afnameprocedure) worden verder in dit rapport uitgebreid beschreven (zie 4 ▪ Afnameprocedure).

3.7 Verbreding en verdieping van de beeldvorming door lerarenteam

Voor en tijdens de afname heeft de leraar zich al een beeld gevormd van de kleuter en diens taalvaardigheid. Na de afname van het screeningsinstrument krijgt een leraar de resultaten van individuele kleuters en van zijn klas als geheel, alsook handvatten om deze resultaten te interpreteren.

In vele gevallen zullen de inschatting van leraar en de resultaten uit de screening sterk overeenkomen en vaak zullen de resultaten van de screening dit beeld verhelderen, elementen toevoegen, of toelaten om een leerling te situeren ten opzichte van een bredere populatie van kleuters in Vlaanderen. In sommige gevallen zullen de inschatting van de leraar en de resultaten elkaar tegenspreken (zie ook Hoofdstuk 6 1.11 INSCHATTING TAALVAARDIGHEID DOOR DE LERAAR). Dat kan bijvoorbeeld zijn omdat de inschatting van een leraar niet helemaal accuraat is: zo is het mogelijk dat een leraar zijn indruk van de luistervaardigheid sterk laat leiden door het spreekgedrag van een kleuter. Van extraverte, babbelgrage kleuters wordt met name al snel verondersteld dat zij ook veel begrijpen. Een andere mogelijke verklaring voor de discrepantie is dat de screening een beperkter beeld oplevert: een screening is immers een momentopname waarbij verschillende factoren een invloed kunnen hebben op de uitkomst (bv. de motivatie van de kleuter of het moment in de dag). Dit is een beperking die inherent is aan het gegeven van 'screening': een screening is breed inzetbaar en geeft signalen, maar heeft niet de bedoeling om een echte 'diagnose', een uitgebreid of diepgaand kleuterbeeld te schetsen.

Als uit de inschatting vooraf en/of uit de resultaten van KOALA 'een knipperlicht gaat branden' is het nodig dat het lerarenteam bredere of diepgaander informatie verzamelt. Deze verzameling van extra informatie hoeft niet strikt gescheiden te worden van de maatregelen die een leraar neemt om kleuters (verder) te doen groeien: tijdens het verzamelen van aanvullende informatie kan een leraar al gericht ondersteunen.

HOOFDSTUK 2: LITERATUURSTUDIE EN EXPERTENBEVRAGING

We voerden tijdens dit onderzoek een beperkte literatuurstudie en expertenbevraging uit om meer te weten te komen over:

- de rol en eventuele meerwaarde van gestandaardiseerde toetsing van taalvaardigheid bij kleuters;
- de mogelijkheden en beperkingen bij het toetsen van kleuters;
- de afnamecondities van een screening of toetsing bij kleuters.

De bevindingen van de literatuurstudie en de expertenbevraging hielpen om een degelijk instrument uit te werken en de mogelijkheden en de beperkingen van toetsing bij kleuters te verhelderen. Daarnaast hielp deze studie en bevraging ook om veel gestelde vragen uit het veld te beantwoorden. Om die reden zullen we de belangrijkste informatie uit dit hoofdstuk ook opnemen (o.a. onder de vorm van 'veel gestelde vragen') in de handleiding.

1 ▪ Methodologie

1.1 Methodologie van de expertenbevraging

De expertenbevraging gebeurde in drie fasen, met elk een eigen focus en bijbehorende experten:

(1) experten die uit eigen onderzoek ervaring hebben met toetsing, taalscreening en/of evalueren van kleuters,
(2) experten die ervaring hebben met de evaluatiepraktijken in het kleuteronderwijs;
(3) experten die specifieke informatie kunnen verschaffen over onderdelen van de handleiding.

Fase 1 vond plaats in de periode van eerste ontwikkeling van het instrument. De elementen die in de bevraging aan bod kwamen, konden dan ook meteen hierin worden opgenomen. In fase 1 werden vier experten bevraagd:

- *Bart Deygers (UGent)*, evaluatie-expert. Thema's als mensenrechten, rechtvaardigheid en billijke beoordeling in toetsing staan centraal in zijn onderzoek. Zijn meest recente onderzoek focust op toetsing van laaggeletterde taalleerders.
- *Marieke Vanbuel (KU Leuven)*, onderzoeker. Haar doctoraat focust op talenbeleid en de rol van taalscreening bij het lokale talenbeleid van scholen.
- *Carolien Frijns (Arteveldehogeschool)*, onderzoeker en lerarenopleider. Ze schreef een doctoraat over interacties in de kleuterklas. In het kader van dit onderzoek ontwikkelde ze een test voor kleuters.
- *Judith Vloedgraven (CITO)*, toetsdeskundige kleuter en projectleider *Kleuters in Beeld*.

Onderstaande vragen werden door de experten in fase 1 behandeld tijdens een open interview:

- Is het volgens jou mogelijk om de taalvaardigheid van 5-jarigen te meten op een valide en betrouwbare manier? Waarom wel/niet?
- Vind je het wenselijk/zinvol om de taalvaardigheid van 5-jarigen te meten?
- Wat zijn vanuit jouw expertise en onderzoek aandachtpunten en valkuilen bij de ontwikkeling van een dergelijke screening?
- Wat zijn vanuit jouw expertise en onderzoek aandachtpunten en valkuilen bij de implementatie van een dergelijke screening?

De inzichten uit deze gesprekken werden meegenomen bij de ontwikkeling van KOALA. Ze lagen met name mee aan de basis van:

- de criteria voor de toetsontwikkeling (bv. het belang van het criterium 'inzetbaarheid');

- de ontwikkeling van het screeningsinstrument (bv. het belang van brede beeldvorming en variatie in typetaken);
- de afnamecondities (bv. digitale afname en afname door bekend persoon).

Fase 2 vond plaats nadat het kader en de concretisering van het taalscreeningsinstrument was vastgelegd. Dat bood de mogelijkheid om in de gesprekken meteen een aantal heel concrete vragen te stellen en de reflecties daarop mee te nemen in de verdere ontwikkeling. In fase 2 werden volgende (groepen van) experten bevraagd:

- *Mieke Devlieger* en *Mathias Chlairie*, inspectieleden: werkgroep Nederlands en werkgroep kleuter, betrokken bij onderzoek kleuterparticipatie.
- *Kirsten Schraeyen (UAntwerpen)*, onderzoeker: actief op het vlak van taal- en leerstoornissen in een meertalige context.
- *Soraya Fret, Véronique De Cock* en *Ewan Claeyssens,* pedagogisch begeleiders: schoolbegeleiders kleuteronderwijs.
- *Ann De Roeck, Katrijn Boumon* en *Isabelle Deruyver,* CLB-medewerkers: psychologen en pedagogen, betrokken bij Prodia (https://prodiagnostiek.be/).
- *Ilse Aerden (UCLL)*, lerarenopleider: zorg en remediërend leren, optie kleuteronderwijs.
- *Onderzoeksteam Centrum voor Taal en Onderwijs*.

De gesprekken in fase 2 verliepen volgens een vast scenario (zie Bijlage 1: Scenario voor gesprekken met experten) waarbij de experten enerzijds antwoordden op vragen, anderzijds feedback gaven op de ontwikkeling van bv. de screening en de schoolfeedback. Een gesprek was opgebouwd uit verschillende componenten waarbij sommige experten meer of minder ingingen op een bepaald onderdeel afhankelijk van hun achtergrond en expertise:

1. percepties t.a.v. de taalscreening
2. het instrument
3. de afname
4. wat na de taalscreening
5. implementatie

In fase 3 werden nog een aantal experten gecontacteerd met specifieke vragen. De antwoorden hielpen om definitieve besluiten te nemen over nog enkele resterende vragen:

- *Bert De Smedt*, hoogleraar Pedagogische Neurowetenschappen aan de Faculteit Psychologie en Pedagogische Wetenschappen van de KU Leuven: uitwisseling rond cognitieve ontwikkeling/wiskundig denken.
- *Jan Van Hoof*, hoogleraar verbonden aan het Departement voor Onderwijs- en Opleidingswetenschappen (Faculteit Sociale Wetenschappen) van de Universiteit Antwerpen: uitwisseling rond schoolfeedback.

1.2 Methodologie van de literatuurstudie

Voor de literatuurstudie werd gezocht op volgende zoektermen (niet-exhaustief):

- assessment (language) Early Childhood Education
- assessment (language) kindergarten
- assessment (language) pre-primary education
- evaluation (language) kindergarten
- evaluation (language) Early Childhood Education
- evaluation (language) pre-primary education
- assessment of young learners
- screening of young language learners
- working memory pre-schoolers

De resulterende artikels werden gefilterd op publicatiejaar, waarbij enkel publicaties na 2000 werden meegenomen. Verder werden de manuscripten geselecteerd die peer-reviewed zijn en verschenen in het Engels of het Nederlands.

Daarnaast raadpleegden we ook enkele boeken over bovenvermelde thema's die de laatste 10 jaar verschenen. Naast deze wetenschappelijke publicaties bestudeerden we ook de handleidingen en achtergronden bij andere toetsen en screeningsinstrumenten voor kleuters in: KOBI-TV, TAL-K en CITO-toetsen (zoals *Kleuters in Beeld* en *Taal voor Kleuters*) en gingen we na welke praktijken er in het buitenland bestaan rond evalueren van jonge kleuters in schoolse contexten. Op die manier brachten we praktijken rond afnamecondities bij kleuters in kaart.

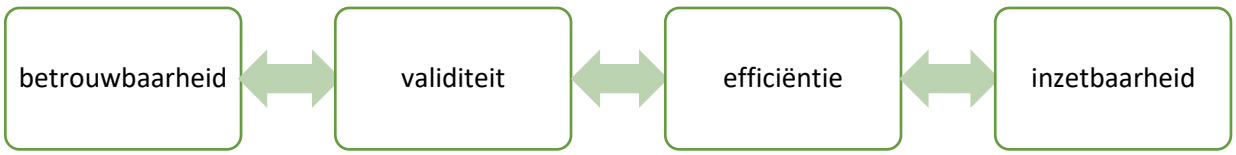
1.3 Rapportage

De verzamelde informatie werd thematisch opgedeeld en gerelateerd aan de vier vooropgestelde criteria. Vervolgens werden alle uitspraken en onderzoeksgegevens geclusterd en gerapporteerd. We rapporteren alleen over zaken waarover onderzoekers het eens zijn. Wanneer een mening of een advies van een expert maar één keer voorkwam, werd die niet opgenomen.

2 ▪ Resultaten

2.1 Vier criteria voor het taalscreeningsinstrument voor kleuters

De verschillende onderdelen van de expertenbevraging en de literatuurstudie geven samen meer informatie over de vier criteria die we in dit onderzoek vooropstellen bij het ontwikkelen van de taalscreening. De vier criteria zijn niet strikt van elkaar te scheiden en beïnvloeden elkaar. Doorheen dit rapport komen de vier criteria aan bod in de verschillende hoofdstukken.



'Betrouwbaarheid' en 'validiteit' zijn evidente begrippen in onderzoek en toetsonwikkeling. De vraag naar een betrouwbare en valide meting van taalvaardigheid bij 5-jarigen vormt dan ook de kern van ons onderzoek (zie onderzoeksvraag 1). De criteria 'efficiëntie' en 'inzetbaarheid' voegden we toe naar aanleiding van gesprekken met de experten en met de leden van de resonans- en stuurgroep en de brede validiteitskaders die gangbaar zijn in de literatuur (o.a. Hill and McNamara, 2011). De vier criteria komen in grote lijnen overeen met de VRIP-parameters zoals die ook in de 'toolkit breed evalueren' (2013) worden beschreven. VRIP staat voor 'Validitity' (validiteit), 'Reliability' (betrouwbaarheid), 'Impact' (impact) en 'Practicality' (praktische haalbaarheid).

Betrouwbaarheid en validiteit beschouwen we niet als eendimensionale constructen. De betrouwbaarheid heeft o.a. te maken met de afnameprocedures, de evaluatiecriteria en de kwaliteit van de toets. Betrouwbaarheid is bij 5-jarigen een extra aandachtspunt omdat onderzoek (Colpin et al., 2006) heeft uitgewezen dat een taaltoets slechts een betrouwbare indicator kan zijn voor taalvaardigheid vanaf een zekere onderwijservaring en leeftijd. Toetsafnames vroeger dan de derde kleuterklas en bij kleuters jonger dan zes jaar blijken niet altijd betrouwbaar. De validiteit van een test hangt dan weer af van een complex samenspel tussen factoren als de toetsresultaten, de interpretatie ervan en het gebruik van de toets (Messick, 1990 en Hill and McNamara, 2011). In de literatuur wordt een onderscheid gemaakt tussen verschillende soorten validiteit, waaronder

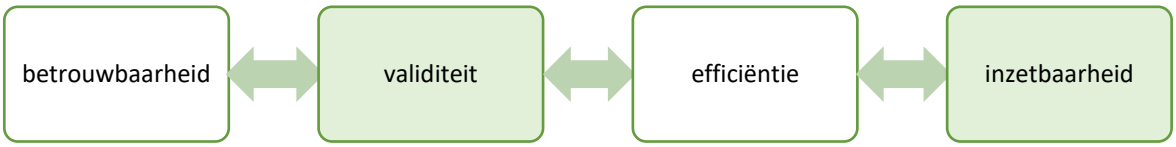
'construct validity[4]', 'face validity[5]' en 'content validity[6]'. Maar ook 'consequential validity' en dus de (sociale) gevolgen van de resultaten van een toets maken dat een toets al dan niet 'valide' wordt ingezet (Popham, 1997, Kane, 2010). Natuurlijk vormen de analyses van het kalibratie-onderzoek een noodzakelijke voorwaarde voor het controleren van de betrouwbaarheid en de validiteit van het instrument.

Efficiëntie en inzetbaarheid spelen een belangrijke rol bij de manier waarop de toets wordt gebruikt (getrouw, selectief en uitgebreid gebruik, Vermeir 2019), en al dan niet leidt tot een positief washback-effect op het onderwijs. Hoe meer rekening gehouden wordt met efficiëntie en inzetbaarheid, hoe meer leraren gemotiveerd zijn om de toets af te nemen en de resultaten in te zetten zoals bedoeld. Shohamy (2010) zegt daarover: *Most research of the past two decades, starting from Messick, demonstrated that the introduction of tests is not an isolated event; rather it is anchored in (political) motivations and intentions. Research also shows that these tests lead to impacts, in the form of intended and unintended consequences.*

Zowel in de literatuur als uit de expertenbevraging komen onderwerpen naar voren die van belang zijn voor één, maar vaker voor meerdere van de hierboven opgesomde criteria. Onderstaande is een samenvatting van de onderwerpen en bijhorende subcategorieën die we konden identificeren uit de expertenbevraging en literatuurstudie. Naast de resultaten geven we telkens een sprekend citaat van een expert weer en geven we aan met welke van de vier criteria de bevindingen vooral in verband kunnen worden gebracht.

2.2 Belang van aansluiting bij de visie op kleuteronderwijs en op evalueren binnen kleuteronderwijs

'Ik maak me vooral zorgen om het effect van de screening op het kleuteronderwijs en niet zozeer om de screening op zich.' (expert)



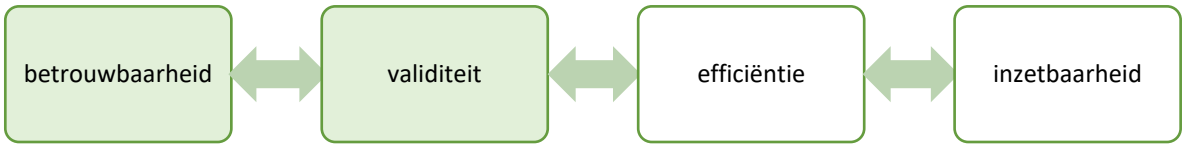
Experten geven de volgende aandachtspunten mee die gerelateerd zijn aan de essentie van onderwijs voor kleuters en evaluatie in deze context:

[4] *Betekenisvaliditeit* is de mate waarin een begrip meet wat er onder dat begrip moet worden verstaan / wat de betekenis van dat begrip is.
[5] *Indruksvaliditeit* betekent dat men (de onderzoeker in het bijzonder) de indruk heeft dat een meting valide is.
[6] *Inhoudsvaliditeit* is te omschrijven als '*Meet dit instrument precies wat er gemeten moet worden?'*.

- een taalscreening bij 5-jarigen kan zinvol zijn in het kader van kwaliteitsvol onderwijs (experten + Teddlie & Reynolds, 2000; De Fraine et al, 2003; De Maeyer & Rymenans, 2004), en het is ook mogelijk om 5-jarigen valide en betrouwbaar te meten (experten + Verhelst, 2002; Schouwstra & Vloedgraven, 2020); wel moeten we opletten met al te grote beslissingen (bv. zittenblijven) op basis van een momentopname (experten);
- beslissingen over taalscreening en over het omgaan met resultaten moeten worden gekaderd binnen een visie (experten):
    - ontwikkelingsgericht versus programmagericht: experten geven aan dat de wenselijke acties na de taalscreening afhangen van hoe we willen dat kleuteronderwijs eruit ziet (bv. vinden we een programma rond klankbewustzijn wenselijk in elke school?);
    - klasintern versus klasextern (remediërend) werken: enkel organisatorische ingrepen (zoals inrichten van zorguren) zijn niet wenselijk, we moeten vooral de klasleraar versterken (experten + Peleman et al, 2019);
    - gefragmenteerd inzetten op deelvaardigheden versus totaalontwikkeling: enkele experten wijzen op (1) de bekommernis dat er nog andere testen zouden komen die vooral typisch schoolse vaardigheden (bv. wiskunde) of deelvaardigheden (zoals letterkennis) in beeld brengen en (2) de bekommernis dat alleen taal aandacht krijgt in het kleuteronderwijs;
    - aanbodsgericht versus behoeftegericht werken: enkele experten wijzen op de gevaren van een te sterk accent op zaken die we kunnen trainen via een gericht aanbod (bv. woordenschat) waardoor de bredere ontwikkeling die aan de basis ervan ligt (en met andere woorden de werkelijke behoefte van de kleuter is) in het gedrang komt (bv. de wereld ontdekken en daar taal aan verbinden).

2.3 Brede beeldvorming en inbedding van de taalscreening

'Hopelijk heeft een school al een beeld van de kleuters vóór de afname van de kleuters, anders is het wel laat voor veel kleuters.' (expert)



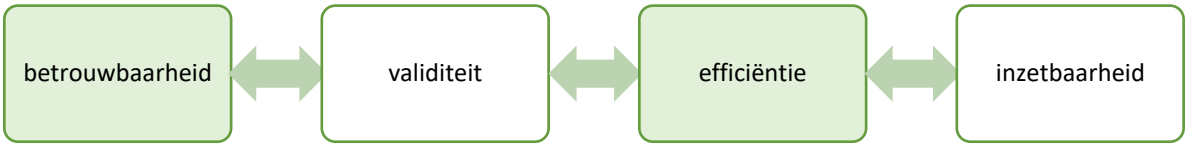
De literatuur en de expertenbevraging wijzen op het belang van het inbedden van een taalscreening binnen een breed beeld van kleuters:

- de screening aanvullen met observaties is noodzakelijk (experten + Gathercole et al, 2004; Colpin, 2006; Schouwstra & Vloedgraven, 2020). Dit kan gebeuren door (1) gebruik te maken van bestaande observaties en (2) bestaande observatiepraktijken kwaliteitsvoller te maken en het hele team (zoals de leraar kleuter) daarbij betrekken;

- kleuters kansen geven ten volle te laten zien wat ze kunnen: dat zijn vaak situaties ondersteund en in interactie met een volwassene (scaffolding), en bijvoorbeeld in spelsituaties (waar kleuters soms meer kunnen dan ze laten zien op een test) (experten + Bijlsma, 2004; Lonigan et al, 2011; Fulcher & Davidson, 2012; Blessing, 2019);
- velden voor verbreding beschrijven: wat kun je nog meer in kaart brengen in verband met taalvaardigheid (bv. gesprekken voeren, geletterdheid…)? (experten);
- leerling zelf betrekken bij de beeldvorming, bv. bij het maken van een taalprofiel (experten + Frijns, 2017).

2.4 Afnamecondities

'Veel factoren spelen een rol opdat een kleuter echt toont wat hij kan op het moment van afname van de taalscreening.' (expert)



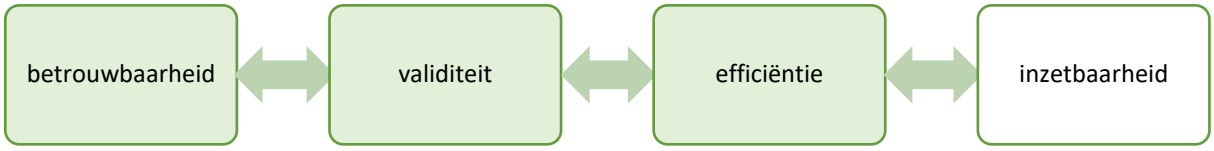
Afnamecondities blijken volgens de experten en de literatuurstudie ook een belangrijk element om mee te nemen bij een taalscreening voor kleuters:

- een implementatietraject voor leraren zou nuttig zijn (experten): (1) ondersteuning en feedback bij eerste afname, (2) hulp bij gestandaardiseerde afname, (3) ondersteunende rol van de zorgcoördinator verhelderen, (4) vormingstraject over acties 'na de taalscreening'… (experten);
- er zijn organisatorische uitdagingen (voldoende leraren, lokalen…): haalbaarheid voor scholen wordt door enkele experten in vraag gesteld (bv. grote scholen, scholen met weinig zorguren; er is voldoende ondersteuning nodig;
- afname door vertrouwd persoon wordt aangeraden door experten: daarvoor is geen evidentie in literatuur (bv. KOBI-TV geen verschil), maar kan wel van belang zijn in functie van welbevinden (zeker bij sommige zwakkere kleuters) (Gathercole et al, 2004);
- de ideale afnameduur: geen uitsluitsel in literatuur (behalve eensgezindheid over het feit dat kleuters een 'beperkte' aandachtspanne hebben (Gathercole et al, 2004) en dat ze beter geen tijdsdruk krijgen opgelegd (Schouwstra & Vloedgraven, 2020); een studie van Wheadon & Stockford (2010) laat zien dat kortere toetsen een lagere nauwkeurigheid hebben dan lange toetsen en toetsen met een grote spreiding in itemmoeilijkheid; veel experten raden 20 tot max. 30 minuten aan; afwisseling en bewegingstussendoortjes worden door alle experten aangeraden;
- het ideale afnamemoment: alle experten raden een 'laat' moment aan (dus eerder november) omwille van (1) effect van leeftijd, (2) ontwikkelingspsychologische argumenten (werkgeheugen, executieve functies, aanpakgedrag, concentratie (Fulcher & Davidson, 2012)), (3) tijd om schoolse vaardigheden te oefenen in de derde kleuterklas (bv. wennen

aan werken op het platte vlak) en (4) druk op leraren voor 'teaching-to-the-test' verminderen;
- zoveel mogelijk kleuters (behalve anderstalige nieuwkomers) laten meedoen om een representatief beeld van de school te krijgen: toch blijven er vragen rond kleuters met specifieke onderwijsbehoeften (bv. kleuter met ADHD of autisme spectrum stoornis) en zij-instromers (bv. Franstalige leerlingen die pas in de derde kleuterklas in het Nederlandstalig onderwijs instromen) (experten); het instrument *Kleuters in Beeld (CITO)* raadt aan om te wachten tot een leerling met een andere gezinstaal minimaal een half jaar op school zit;
- gedifferentieerde afname versus standaardisering: experten raden aan om een zekere vrijheid in te bouwen zonder de standaardisatie in gedrang te brengen (bv. minimale en maximale groepsgrootte, welke vorm van scaffolding kan wel/niet…).

2.5 Het screeningsinstrument

'Het instrument bevat heel veel zinvolle opdrachten die herkenbaar zijn voor kleuters en kleuterleraren.' (expert)



Een geschikt screeningsinstrument voor kleuters voldoet aan een aantal voorwaarden:

2.5.1 Instructies:

- aandachtspunten voor de Instructies: (1) beschrijven hoe leraren kleuters best voorbereiden op de afname van de screening en (2) hoe ze kunnen reageren wanneer kleuters fouten maken of om bevestiging vragen (experten);
- instructies in de vorm van filmpjes zijn een grote meerwaarde (experten).

2.5.2 Manieren van meten:

- verschillende manieren van meten best combineren (experten + Fulcher et al, 2016): doe-opdrachten zijn makkelijker en vertrouwder voor kleuters;

- verschillende vormen van ((semi-)gestandaardiseerde) observaties zijn mogelijk (Gysen et al, 2001; Schouwstra & Vloedgraven, 2020), bijvoorbeeld: de leraar scoort de gevolgen of resultaten van het gedrag (over een langere periode) en niet het gedrag zelf, de leraar observeert het gedrag op het moment dat het zich voordoet…;
- digitaal en papier kan allebei of een combinatie ervan (Choi et al (2003): digitaal kan zeer motiverend zijn voor kleuters; experten wijzen op het feit dat het vooral een 'implementatievraagstuk' is (m.a.w. zijn scholen en leraren er klaar voor?); het instrument *Kleuters in Beeld (CITO)* werkt digitaal met sleepvragen en hotspots en met meerkeuzevragen (drie of vier antwoordmogelijkheden) waar nodig;
- adaptieve toetsing of screening is een meerwaarde: (1) kleuters niet frustreren, (2) enkel tijd besteden aan toetsing of screening wanneer het iets oplevert en (3) zinvollere informatie voor leraren (experten); *KOBI-TV* werkt met een systeem waarbij de afname op een bepaald moment wordt afgebroken, *Kleuters in Beeld* werkt met drie verschillende versies.

2.5.3 Concrete aandachtspunten bij de items:

- aansluiten bij kleutergedrag en -ontwikkeling (experten + Hasselgreen, 2000): gebruik maken van spel en fantasie en inzetten op plezier;
- kenmerken van de afbeeldingen (experten + Ioannou-Georgiou, 2003): aantrekkelijk, herkenbaar en motiverend;
- opletten voor effect van 'immediate respons' (bv. eerste vraag bij verhaaltjes) (Hasselgreen, 2000).

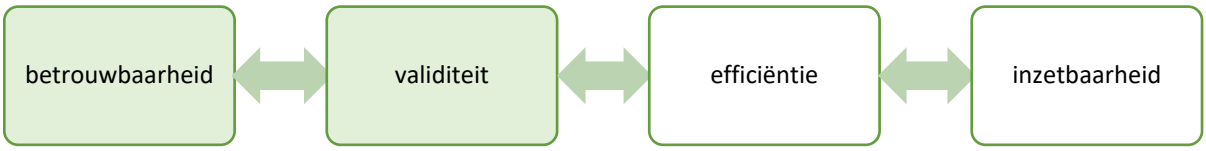
Bijna alle experten verwijzen naar andere bestaande testen voor jonge kleuters en de aandachtspunten die we daaruit kunnen meenemen:

- KOBI-TV: te weinig concrete handvatten om er nadien mee aan de slag te gaan.
- SALTO (zie ook SONO-rapport: *Helpen talenbeleid en taalscreening taalgrenzen verleggen?*, Vanbuel et al, 2017):
    - scholen doen weinig met de taalscreening eerste leerjaar omdat de informatie/output die ze krijgen te weinig handvatten geeft;
    - in sommige scholen scoren alle leerlingen groen: geen motivatie om iets te veranderen aan de onderwijsaanpak;
    - in sommige scholen scoren alle leerlingen rood: demotiverend voor leraren.
- TOETERS:
    - teveel gebruikt om kleuters te 'labelen' en advies te geven rond overgang naar eerste leerjaar;
    - heeft een sterk teaching-to-the-test effect.
- CITO TAAL:
    - te schools;
    - eenzijdige opdrachten: allemaal meerkeuzevragen;
    - niet motiverend voor kleuters;
    - sluit niet altijd aan bij leefwereld;
    - sterke focus op niet relevante woordenschat;

- wordt vaak afgenomen door zorgleraren, sommige klasleraren hebben geen zicht op de resultaten.

2.6 Interpretatie van de resultaten van de screening

'We moeten vermijden dat een schoolteam in een 'zorgkramp' schieten. Remediëren en aanbodsgericht werken in kleine groepjes of één-op-één zijn zinvol, maar waardeloos als basislaag van de piramide niet stevig genoeg is.' (expert)

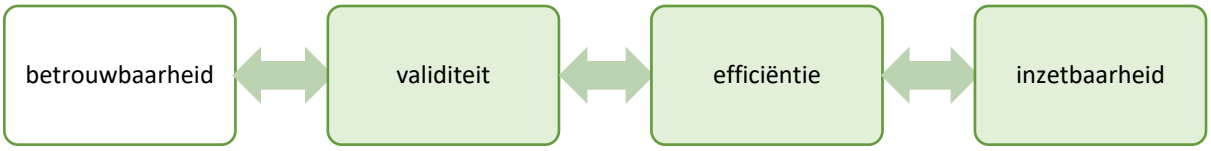


In de literatuur en bij de expertenbevraging wordt ook stilgestaan bij het interpreteren van resultaten:

- de taalscreening is één element in de beeldvorming van een kleuter en een moment in zijn ontwikkeling dat wordt vastgelegd (experten);
- aandacht is nodig voor, maar voorzichtigheid is aangewezen bij de interpretatie van de prestaties van meertalige leerlingen, bv. leerlingen die aan een inhaalbeweging in het Nederlands bezig zijn (Walter, 2008;, Heugh et al., 2009; Dale et al 2011);
- leraren kunnen leraren vrij goed inschatten, maar hebben de neiging om leerlingen met laag SES te onderschatten (Sneyers et al, 2020): de taalscreening kan een nuttig instrument zijn om de beeldvorming te 'objectiveren';
- een taalprofiel van (meertalige) leerlingen kan nuttig zijn om de resultaten beter te kunnen interpreteren: gezinstaal, aantal jaar in (Nederlandstalig) kleuteronderwijs, Nederlands taalaanbod, anderstalig taalaanbod, taalproductiekansen…(experten);
- het resultaat van een leerling die onder de verwachting scoort is vooral een signaal om beter te gaan kijken, maar nog geen reden tot directe ongerustheid (experten + Schouwstra & Vloedgraven, 2020).

2.7 Implementatie en gebruik van een screeningsinstrument

'Ik denk dat vooral het tijdsargument een rol zal spelen. Scholen steken soms tijd in toetsen waar ze dan niet altijd iets mee doen en waar ze dus het nut niet van inzien. (…) Leraren zijn nadien soms teleurgesteld in de 'return' die ze uit toetsen halen.' (expert)



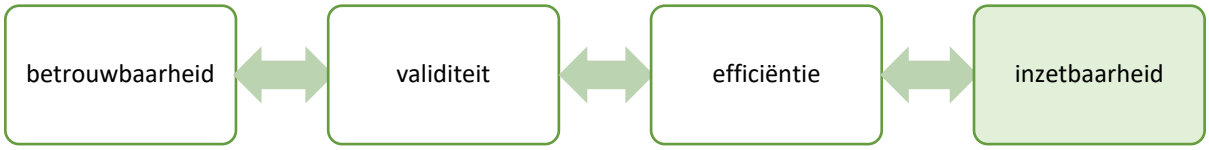
Met betrekking tot elementen die meespelen in een goede implementatie van de screening, geven het literatuuroverzicht en de experten de volgende aandachtspunten mee:

- het is belangrijk om een duidelijk doel te communiceren opdat de screening op die manier wordt ingezet (experten + Vanbuel et al, 2016; Penninckx et al, 2017; Vermeir, 2020); dit draagt ook bij tot het 'rechtvaardig' inzetten van het instrument en dus tot het verhogen van de gelijke onderwijskansen (Deygers, 2019);
- evenzeer is het essentieel om de meerwaarde van de taalscreening te expliciteren (experten + Vanbuel et al., 2017):
  - algemene meerwaarde (zoals een eenvoudige en snelle manier een beeld te krijgen van de taalvaardigheid van een kleuter, hulp bij goede acties om kleuters te ondersteunen, visie op taal en taalontwikkeling versterken in lerarenteam, meer onderbouwing voor gesprekken over kleuters met ouders, CLB, zorg…),
  - specifieke meerwaarde voor scholen met veel kleuters in de 'groene' zone én voor scholen met veel kleuters in de 'rode' zone;
- voor een goede implementatie dient de klasleraar van bij de start betrokken te zijn en de regie te krijgen (experten + Schouwstra & Vloedgraven, 2020);
- daarnaast moet een effectief screeningsinstrument oog hebben voor 'wat na de screening': (1) van screenen, over 'beeld' komen tot actie en (2) vanuit de taalscreening komen tot gezamenlijke taalpraktijk en teamwerking (experten + Vanbuel et al, 2017);
- screenings letten best op met 'labelen' en 'categoriseren', in het algemeen ('dit is een taalzwak kind') en zeker ook specifiek naar meertalige kleuters ('hij is taalzwak want praat alleen Frans') (experten + Byers-Heinlein & Lew-Williams, 2013; Conboy & Montanari, 2016; Miller et al, 2018);
- implementatie heeft meer kans op succes indien er linken gemaakt worden met bestaande praktijken en modellen van GOK en zorg (experten + Vanbuel, 2020): kindvolgsystemen, registratiesystemen als smartschool of questi, ondersteuningsmodel, handelingsgericht werken….
- het is van belang om duidelijkheid te creëren rond het gebruik van KOALA naast andere toetsen om 'overtesting' te vermijden (experten).

2.8 Informatie voor de gebruiker

'Het zou goed zijn als we aan leraren kunnen duidelijk maken dat de kleuterklas een taalbad is. (…) Taalintegratie is dan: hoe krijgen we dat taalbad tot bij die kleuters waarbij het niet lukt? Pre-teaching, een variatie aan homogene en heterogene groepjes, meespelen… kan allemaal. Dus, we moeten een gerichtere aanpak en meer aandacht en tijd aan bepaalde kleuters binnen de klas geven.

Integratie maakt ook dat de sterkere kleuters mee profiteren van de gedifferentieerde aanpak.' (expert)



Experten en de literatuur stellen dat het belangrijk is om voldoende informatie mee te geven aan de gebruikers, om op die manier de inzetbaarheid van het instrument te verhogen. Informatie die zeker aan bod moet komen:

- info over het verloop van taalontwikkeling, en meer specifiek over meertalige taalontwikkeling (successief en simultaan) (m.a.w. *wat kunnen we verwachten op welke leeftijd?*) (experten + Byers-Heinlein & Lew-Williams, 2013; Conboy & Montanari, 2016; Miller et al, 2018): leraren hebben immers vaak een verkeerde verwachting van bv. meertalige leerlingen en moeten rekening houden met grilligheid, sprongsgewijze ontwikkeling en grote verschillen tussen 5-jarige kleuters (experten + Goorhuis-Brouwer, 2005, Fulcher & Davidson, 2012);
- nood aan zinvolle maar 'leesbare' output na taalscreening: (1) belang van een school- en klasoverzicht, (2) informatie over verschillende doelstellingen of subdomeinen, (3) mogelijkheid tot verfijnen van de analyse door het geven van meer gedetailleerde informatie over kenmerken van de opdrachten (bv. schoolse woordenschat, rekentaal, vraag over een verhaal op macro- of op microniveau…) en (4) vergelijken met leerlingen of scholen met dezelfde kenmerken (bv. meertalige leerlingen, kansarme leerlingen) (experten + Van der Slik, 2000,; Sijtstra et al, 2002; De Fraine et al, 2003);
- nood aan aanpakken, adviezen, materialen om in te zetten na de afname van de screening vanuit drie vragen: *(1) wat doe je al?, (2) hoe kun je dat beter doen? en (3) wat moet je meer doen?* Daarnaast is een opsomming van kleine acties met groot effect op korte termijn motiverend voor leraren (experten);
- informatie over en concrete handvatten voor de communicatie met ouders en betrekken van ouders (experten);
- informatie en concrete handvatten over het opvolgen van evolutie (experten + Colpin, 2006): *(1) wat hebben je vastgesteld door KOALA?, (2) wat heb je intussen gedaan?* en *(3) welke evolutie stel je vast?;*
- informatie en concrete handvatten over het betrekken van het hele team bij bovenstaande punten (experten + Vanbuel, 2020).

In de volgende hoofdstukken tonen we aan hoe we met bovenstaande elementen rekening hielden bij de ontwikkeling van en het onderzoek naar KOALA. Veel bevindingen namen we ook mee bij het uitwerken van de handleiding (en bijhorende fiches) bij KOALA.

HOOFDSTUK 3: VOORONDERZOEK

In het vooronderzoek focusten we enerzijds op de geschiktheid van SALTO-taken voor de doelgroep van kleuters. Met name brachten we motivatie, aantrekkelijkheid, moeilijkheidsgraad van de typetaken, afbeeldingen en contexten voor deze nieuwe, jongere doelgroep in kaart en bekeken we de haalbaarheid voor een afnemer.

Zo keken we of 5-jarige kleuters voldoende concentratievermogen hebben en voldoende 'testrijp' zijn om de opdrachten uit te voeren. Ook gingen we van een aantal taken na of de nieuwe afbeeldingen werken, of ze duidelijk zijn en of ze motiverend werken voor de kleuters.

We brachten tijdens het vooronderzoek ook de reacties van kleuters op de nieuwe typetaken (zie 3 ▪ TYPETAKEN IN KOALA) in kaart: begrijpen de kleuters wat de bedoeling is, zijn de taken voor hen voldoende herkenbaar en zijn ze ook uitvoerbaar voor de toetsafnemer. Bij sommige opdrachten was het ook belangrijk na te gaan of kleuters de opdracht motorisch kunnen uitvoeren.

Voorbeeld van een opdracht waarbij motoriek een rol speelt: vingerpop
Plak de onderste strook dicht zodat je een rondje krijgt. Let op, zorg dat je een rondje hebt aan de achterkant van de vingerpop, niet aan de voorkant! Nu is je vingerpop klaar.
 *(beeld uit de testafname)*

Figuur 9: Voorbeeld van een opdracht waarbij motoriek een rol speelt: vingerpop

Tijdens deze testfase observeren de onderzoekers de kleuters en volgen ze hun denkwijze. Af en toe vragen ze aan kleuters om uit te leggen waarom ze een bepaalde keuze maken of handeling stellen. De onderzoekers proberen met andere woorden te achterhalen waarom kleuters bepaalde antwoordmogelijkheden wel/niet kiezen.

Ten slotte wilden we in het vooronderzoek nagaan of een digitale afname mogelijk is bij 5-jarigen.

In tegenstelling tot het piloot- en kalibratie-onderzoek vormden de prestaties van de kleuters niet de belangrijkste data, wel de observaties van de toetsafnemers en de reacties van de kleuters.

1 ▪ Werkwijze

1.1 Participanten

We testten 20 individuele kleuters. In totaal namen acht onderzoekers van het CTO de screening af in de thuissituatie. De selectie van die individuele kleuters gebeurde op basis van persoonlijke contacten en een oproep op Facebook.

Daarnaast namen we enkele taken af in een klas 5-jarigen in een multiculturele school in Gent in de periode van 5 tot en met 20 oktober. De school werd gekozen op basis van haar doelpubliek, namelijk een zeer diverse populatie met veel risicokleuters. De school zat niet in de selectie voor het piloot- of kalibratie-onderzoek.

De onderzoekers namen notities en maakten opnames. De afname gebeurde op gestandaardiseerde wijze, maar de kleuters mochten spontaan reageren tijdens de testafname, en de onderzoekers stelden gerichte vragen om de redenering achter een onverwachte of foutieve antwoordkeuze te achterhalen. De bevindingen werden in het onderzoeksteam besproken en dienden als basis voor aanpassingen en verdere ontwikkeling van de taken voor pilootafname.

1.2 Taken in het vooronderzoek

We testten in totaal 21 taken uit. Twee taken (van de 21) werden ook digitaal uitgetest: een doe-opdracht op papier/scherm en een meerkeuze-opdracht. Vaak werden er meerdere versies van een taak getest om na te gaan of een afbeelding of instructie beter werkte na aanpassing.

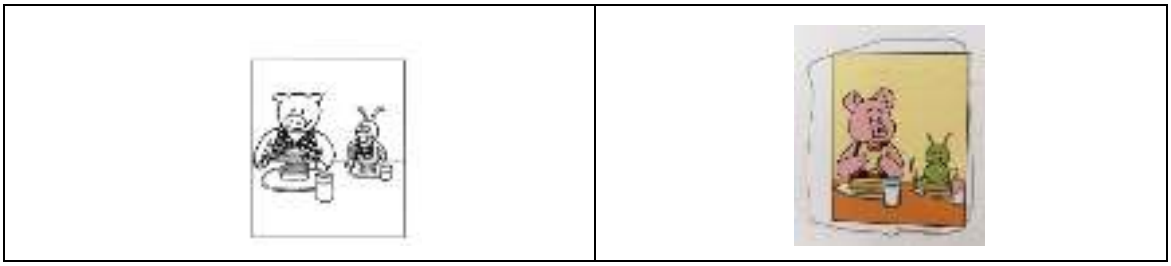
2 ▪ Bevindingen

De testafnames bevestigden een groot deel van de gemaakte keuzes en ingrepen. Op basis van de testafnames konden we vaststellen dat:

- de meeste kleuters zich makkelijk twintig minuten tot een half uur kunnen concentreren mits voldoende afwisseling in taken (maar die concentratiespanne is zeer wisselend en hangt onder andere samen met de rijpheid en de taalvaardigheid van de kleuter);
- kleuters baat hebben bij de integratie van doe-opdrachten in het toetsconstruct (omwille van concentratie, motivatie en testrijpheid);
- digitale afname op tablet mogelijk en motiverend is;
- kleuters de nieuwe tekeningen 'mooi', 'leuk' en duidelijk vinden (met andere woorden aantrekkelijk maar zonder dat de details teveel afleiden);
- alle kleuters (behalve enkele kleuters met weinig schoolse ervaring) in staat zijn de opdrachten motorisch uit te voeren;
- afname in groepjes van 5 het absolute maximum is;
- een grote hoeveelheid meerkeuze-opdrachten voor (taal)zwakkere kleuters moeilijk en demotiverend kan zijn.

Daarnaast gaven de testafnames ook informatie voor verdere aanpassingen en hielpen de observaties om knopen door te hakken. Op basis van de testafnames beslisten we bijvoorbeeld om een kring te laten trekken in plaats van een kruisje te laten zetten onder een tekening. Op die manier vermijden we een extra (denk)handeling (namelijk eerst de juiste tekening kiezen en dan het vakje onder de tekening 'zoeken' om aan te kruisen) en een fijn-motorische handeling. Tegelijkertijd beslisten we ook dat dit geen beoordelingscriterium vormt: de toetsafnemer kan de kring trekken in de plaats van de kleuter als hij/zij daar omwille van motorische redenen zelf niet in slaagt.

Voorbeeld aanpassing instructie: Varken en Rups	
Wat vonden Varken en Rups het allerleukst om samen te doen? Neem je potlood en zet een kruisje onder de juiste tekening.	*Wat vonden Varken en Rups het allerleukste om samen te doen? Trek een kring rond de juiste tekening.*



Figuur 10: Voorbeeld aanpassing instructie: Varken en Rups

We beslisten ook om geen meerkeuze-opdrachten op te nemen met meerdere correcte antwoordmogelijkheden. Zulke complexe cognitieve verwerking bleek hoog gegrepen voor de meeste kleuters en levert slechts beperkte extra informatie op over de luistervaardigheid van een kleuter. Bovendien is een combinatie van meerdere soorten antwoordmogelijkheden binnen een typetaak verwarrend voor kleuters. Dit was de reden van het wegvallen van een van de vier taken uit SALTO.

HOOFDSTUK 4: PILOOTONDERZOEK

Aan het pilootonderzoek namen twaalf scholen en meer dan 300 kleuters deel. De afnamecondities waren gelijkaardig aan de condities van het kalibratie-onderzoek. Naast een afname op papier werd ook een digitale afname op tablet grondig onderzocht.

Er namen 206 kleuters deel aan de papieren afname. De data van de afnames op papier werden kwantitatief en kwalitatief verwerkt. Aan de afname op tablet namen 107 kleuters deel. De digitale data werden kwalitatief verwerkt om op die manier informatie te verzamelen over de mogelijkheid van digitaal toetsen bij kleuters en de noodzakelijke afnamecondities. Omwille van het relatief klein aantal kleuters dat hieraan deelnam hebben we geen psychometrische analyses uitgevoerd op deze dataset. De onderzoekers verzamelden waardevolle kwalitatieve informatie via participatieve observatie, dit zowel bij de papieren als bij de digitale afnames en zowel over de screening als geheel als over de verschillende taken en items.

Het pilootonderzoek verschafte op deze manier informatie over de betrouwbaarheid en moeilijkheidsgraad van de taken, items en eventuele afleiders, de duidelijkheid van de instructies en de haalbaarheid van toetsing bij jonge kinderen. Op basis van de onderzoekersobservaties en de psychometrische analyses van dit pilootonderzoek werd de definitieve toetsbatterij voor het kalibratie-onderzoek samengesteld en werden, waar nodig, de taken aangepast (bv. omdat bepaalde instructies niet duidelijk waren of omdat bepaalde taken te lang uitvielen voor 5-jarige kleuters).

1 ▪ Selectie van de deelnemers

We vertrokken van een gelegenheidssteekproef, waarbij we variatie nastreefden op vlak van regio, graad van verstedelijking en leerlingpopulatie. Van de deelnemende scholen beschikten wij over een persoonlijk contact (m.a.w. contacten via het onderzoeksteam van het CTO of contacten via leden van de resonansgroep). De scholen maakten *geen* deel uit van de steekproeftrekking voor het kalibratieonderzoek. Uiteraard was de formele toestemming voor deelname van de school zelf een vereiste.

In totaal hebben er 12 scholen deelgenomen:

- Brusselse rand: 1
- Regio Aarschot: 1
- Regio Genk: 2
- Regio Gent: 1
- Regio Geraardsbergen: 1
- Regio Leuven: 6

Deelnamecriteria van de kleuters uit die scholen waren:

- geboren in 2015
- 3de kleuterklas
- geen kleuters die voor het eerst in Nederlandstalig onderwijs zitten

In totaal hebben er 313 kleuters deelgenomen aan het pilootonderzoek.

2 ▪ Afnamemodaliteiten

Alle afnames vonden plaats in de periode van 7 tot en met 20 oktober. Acht onderzoekers van het CTO namen toetsen af in de scholen.

De afnames vonden in de voormiddag plaats, op de school zelf, in een aparte ruimte buiten de klas. De taken werden afgenomen door de onderzoeker. (Minstens) een leraar (of zorgleerkracht, begeleidend personeel, assistent) werd gevraagd om aanwezig te zijn. Die leraar was actief betrokken bij de toetsafname: hij/zij zorgde er mee voor dat de afnames vlot verlopen, dat alle kleuters de instructies goed begrepen en hield de kleuters bezig tijdens individuele observatietaken. Kleuters werden zo ver mogelijk uit elkaar gezet om spieken en samenwerking te voorkomen. De digitale taken werden afgenomen door middel van iPads.

De bedoeling was om per voormiddag 3 groepjes van 5 kleuters na elkaar te testen. Voor ieder groepje werd een aparte takencluster gebruikt. In de praktijk kon er soms minder getest worden dan gepland (bijvoorbeeld: 2 groepjes in plaats van 3; of taken die binnen een cluster door tijdsgebrek niet meer getest konden worden; of minder kleuters aanwezig dan voorzien).

Uiteindelijk werden, op 2 taken na, van alle taken 30 prestaties of meer verzameld. Deze aantallen bleken voldoende voor psychometrische analyses die de betrouwbaarheid en moeilijkheidsgraad van items en taken nagaan.

3 ▪ Samenstelling van overlappende clusters

We testten in totaal 30 taken. Hiervan werden 18 taken niet enkel op papier, maar ook op tablet afgenomen. De 48 te piloten taken werden over 16 'clusters' (takenpakketten) verdeeld. De samenstelling van een cluster beantwoordde aan de volgende criteria :

- Elke cluster wordt getest door drie groepjes van 5 kleuters;
- Een cluster bestaat uit 6 taken;
- Iedere taak komt (minstens) in 2 verschillende clusters voor, om voldoende linken te kunnen leggen binnen de dataset;
- Digitale en papieren clusters zijn op dezelfde manier samengesteld, in functie van vergelijkbaarheid.

Volgens deze berekening zijn minstens 240 kleuters nodig om van alle taken 30 scores te verzamelen - een aantal dat zeker werd bereikt in dit pilootonderzoek.

Tabel 1 geeft de concrete samenstelling weer van de papieren en digitale clusters, gelinkt aan de typetaken en de doelstelling die in kaart wordt gebracht.

Nr.	Taak titel	Type taak	Doel: begrijpen van...	#	C 1	C 2	C 3	C 4	C 5	C 6	C 7	C 8	T	P
01	Bewegen	Doe-opdracht	Instructies (Fysieke reactie)	1 2						✓			✓	✓
02	Boekenhoek	Kies-opdracht	Vragen (Intenties en voorkeuren)	4	✓	✓							✓	✓
03	Dieren nadoen	Doe-opdracht	Vragen (Ervaringen en belevingen)	1								✓	✓	✓
04	Eten	Doe-opdracht	Vragen (Intenties en voorkeuren)	2					✓				✓	✓
05	Eenzaam	Kies-opdracht	Vragen (Gevoelens)	4	✓	✓							✓	✓
06	Hoepel	Doe-opdracht	Instructies (Fysieke reactie)	7							✓		✓	✓
07	Jelle	Zoek-opdracht	Informatieve mededelingen (Concrete gebeurtenissen)	5				✓	✓				✓	✓
08	Juf Is jarig	Kies-opdracht	Instructies (Mentale reactie of talige handeling)	5			✓	✓					✓	✓
09	Kabouters	Kies-opdracht	Verhalen (Verhalen)	5	✓	✓							✓	✓
10	Klasafspraken	Kies-opdracht	Informatieve mededelingen (Regels en voorschriften)	5				✓	✓				✓	✓
11	Klastaakjes	Kies-opdracht	Vragen (Gevoelens)	4			✓	✓					✓	✓
12	Konijntjes	Kies-opdracht	Verhalen (Verhalen)	6			✓	✓					✓	✓
13	Lievelingsboeken	Kies-opdracht	Vragen (Intenties en voorkeuren)	4							✓	✓	✓	✓
14	Lievelingsspeelgoed	Kies-opdracht	Vragen (Intenties en voorkeuren)	4			✓	✓					✓	✓
15	Mona's hoeken	Kies-opdracht	Informatieve mededelingen (Concrete gebeurtenissen)	4						✓	✓		✓	✓
16	Mug en olifant	Kies-opdracht	Verhalen (Verhalen)	4					✓	✓			✓	✓
17	Myriam	Kies-opdracht	Informatieve mededelingen (Regels en voorschriften)	6	✓							✓	✓	✓
18	Naar bad	Kies-opdracht	Verhalen (Verhalen)	4					✓	✓			✓	✓
19	Park	Kies-opdracht	Informatieve mededelingen (Regels en voorschriften)	5		✓	✓						✓	✓

20	Rommel in de eetzaal	Zoek-opdracht	Instructies (Mentale reactie of talige handeling)	5		✓						✓	✓
21	Rommel in de klas	Doe-opdracht	Instructies (Fysieke reactie)	4	✓							✓	✓
22	Speeltijd	Zoek-opdracht	Informatieve mededelingen (Concrete gebeurtenissen)	6		✓						✓	✓
23	Spelen	Doe-opdracht	Vragen (Intenties en voorkeuren)	2						✓		✓	✓
24	Turnles	Doe-opdracht	Informatieve mededelingen (Concrete gebeurtenissen)	6	✓							✓	✓
25	Vandaag	Doe-opdracht	Vragen (Ervaringen en belevingen)	1							✓	✓	✓
26	Varken en Rups	Kies-opdracht	Verhalen (Verhalen)	4							✓	✓	✓	✓
27	Verjaardagsfeest	Zoek-opdracht	Informatieve mededelingen (Concrete gebeurtenissen)	6			✓					✓	✓
28	Vingerpop	Doe-opdracht	Instructies (Fysieke reactie)	4								✓	✓	✓
29	Waar is…?	Zoek-opdracht	Vragen (Situaties, handelingen, voorwerpen)	6							✓	✓	✓	✓
30	Zandtafel	Zoek-opdracht	Vragen (Situaties, handelingen, voorwerpen)	4					✓	✓			✓	✓

Tabel 2: Overzicht samenstelling clusters voor pilootonderzoek

Om voor elke taak gegevens te verzamelen van een gevarieerde leerlingpopulatie, vermeden we om dezelfde cluster meerdere keren op dezelfde school te testen. In de planning hebben we steeds reserve-momenten voorzien te kunnen aanpassen en bijsturen.

Van de kleuters werden geen persoonlijke gegevens bijgehouden, met uitzondering van de school. Elke kleuter uit een groepje kreeg daartoe een ID-nummer toegewezen, samengesteld uit: schoolcode (2 letters), cluster nummer (C01, C02, C03…) en kleuter nummer (K1, K2, K3, K4, K5).

4 ▪ Taken in het pilootonderzoek

In het pilootonderzoek werden 30 taken getest. De kwantitatieve en kwalitatieve resultaten van de pilootafname bepaalden welke taken worden behouden voor het kalibratie-onderzoek, en welke taken moeten worden aangepast.

4.1 Context

Elke taak bestaat uit een introductie en een aantal toetsvragen. De bedoeling van de intro is om een herkenbare en vertrouwde context te creëren voor de kleuter. Op die manier maakt een vraag deel uit van een betekenisvol geheel. Als de kleuter de taak als zinvol ervaart, verhoogt dit de betrokkenheid.

Bijvoorbeeld bij taken waarbij het gaat over 'verhalen begrijpen', toont de toetsafnemer telkens een introductieprent van de hoofdpersonages in het verhaal.

Voorbeelden van introducties:

- TAAK 'SPELEN' (DOEL: VRAGEN BEGRIJPEN): *Ik wil graag met jou praten over de hoeken waar jij graag speelt. Kijk maar al eens wat er allemaal op de foto's staat.*
- TAAK 'KABOUTERS' (DOEL: VERHALEN BEGRIJPEN): *Ik ga een verhaal over twee kabouters vertellen. Eerst vertel ik een stukje van het verhaal. Bij het verhaal zijn veel tekeningen gemaakt. Jij moet straks kiezen welke tekeningen juist zijn. Dit is kabouter Toon. Dit is kabouter Mo.*

De toetsafnemer leest elke toetsvraag in principe twee keer voor. Wanneer een kleuter de opdracht niet begrepen heeft of vraagt om te herhalen, mag de toetsafnemer nog een derde keer voorlezen. De vetgedrukte woorden in de toetsvraag moeten heel duidelijk worden uitgesproken en mogen lichtjes worden benadrukt. De toetsafnemer maakt geen gebaren, tenzij dit in de instructie wordt meegegeven.

4.2 Soorten taken in het pilootonderzoek

Er zijn drie typetaken (zie ook HOOFDSTUK 1: TOETSCONSTRUCT). Elke cluster in het pilootonderzoek bevatte elke typetaak. Dit betekende dus ook dat elke kleuter elke typetaak heeft uitgeprobeerd.

4.3 Aantal taken en items

Het instrument dat voor het pilootonderzoek gebruikt werd bevatte 30 taken. De meeste taken (24) hebben 4, 5 of 6 items. Twee taken hebben meer items; het gaat hier om gerichte opdrachten met gestandaardiseerde observatie met 7 of 12 items. Vier taken hebben minder items. Dit zijn gerichte opdrachten met gestandaardiseerde observatie met hulp van afbeeldingen op papier, met 1 of 2 items.

5 ▪ Bevindingen uit kwalitatieve observaties

Tijdens het pilootonderzoek werden de ervaringen en observaties van toetsafnemers en indrukken uit de gesprekken met aanwezige leraren en kleuters bijgehouden. Een analyse van deze documenten bood inzichten op vlak van afnameconditie, instructies, afnamemodaliteiten en de verschillende soorten taken.

5.1 Afnamecondities

Alle toetsafnemers zijn het erover eens dat een testafname in groep moet doorgaan in een apart lokaal. De toets vraagt behoorlijk wat concentratie van de kleuters. De toets afnemen in aanwezigheid van andere kleuters zou voor te veel afleiding zorgen en zorgt er haast onvermijdelijk voor dat sommige instructies van de toetsafnemer verloren gaan omwille van (te veel) achtergrondlawaai.

Voorlopig doen we nog geen uitspraak over de wenselijkheid om de toets individueel of (gedeeltelijk) in groep te laten verlopen. De literatuur rond testafnames bij jonge kinderen geeft hierover geen uitsluitsel. Op basis van de pilootafname kunnen we stellen dat afname in groep mogelijk lijkt op voorwaarde dat er een apart lokaal voorzien wordt en er de nodige maatregelen genomen worden om afkijken en samenwerken te voorkomen (zie hieronder). Dat vereist wel dat er een andere (zorg)leerkracht in de klas bij de andere kleuters blijft.

Bij de gerichte opdrachten met gestandaardiseerde observatie kan in sommige gevallen de introductie in groep verlopen. De afname gebeurt steeds individueel. Bij een introductie in groep moeten we erover waken dat er niet teveel tijd zit tussen de introductie de eigenlijk opdrachten. Tijdens de pilootafname konden de kleuters die even moesten wachten vrij spelen in dezelfde ruimte, bijvoorbeeld met constructiemateriaal of tekenmateriaal. Dit verliep doorgaans vlot. Het helpt om afspraken te maken met de kleuters over waarmee ze mogen spelen (één of twee materialen volstaan) en om een duidelijke ‘speelruimte’ af te bakenen.

5.2 Afkijken

Tijdens de pilootafname stelden we vast dat zowel bij de papieren versie als bij de digitale versie kleuters probeerden samen te werken en af te kijken. Oorspronkelijk stelden we voor om een U-

vormige opstelling te maken met de aanwezige tafels. Hier zijn we vanaf gestapt omdat het in deze opstelling (te) gemakkelijk was om af te kijken en kleuters spontaan in interactie gingen. Het werkte beter om kleuters achter elkaar te laten zitten aan aparte tafeltjes.

Voor het kalibratieonderzoek stelden we daarom het volgende voor:

- Tafels achter elkaar zetten (zoals in Figuur 11) of in V-vorm.
- Tijdens of meteen na het oefenitem bij het begin van de afname afspraken maken met de kleuters, d.w.z. niet kijken op het papier/tablet van iemand anders, het antwoord niet voorzeggen, eigen papier of tablet niet tonen aan anderen.
- De toetsafnemer kan (subtiel) tussen twee kleuters gaan staan die blijven afkijken of samenwerken.
- Positief reageren, extra bevestiging en aanmoediging geven aan onzekere kinderen, bijvoorbeeld 'amai, jij hebt een mooie kring getrokken'.

Figuur 11: Voorbeeld van tafelopstelling die samenwerking tussen kleuters ontmoedigt

5.3 Afnamemodaliteit: op papier versus digitaal

Onderzoeksliteratuur geeft aan dat papieren en digitale taaltoetsen niet noodzakelijk verschillende vaardigheden meten (Choi, Kom and Boo, 2003). Ook de toetsafnemers van de pilootafnames meldden geen opvallende verschillen tussen de papieren afname en de afname op tablet. Zowel op papier als op tablet verliepen de afnames vlot. Voor enkele kleuters was het motorisch moeilijk om een kring te trekken rond een afbeelding, al vormt dit geen onoverkomelijk probleem. Alle kleuters waren voldoende vertrouwd met tablets om een afbeelding aan te tikken. Het lijkt dus zeker

interessant om de mogelijkheid tot betrouwbare metingen via digitale afname verder te verkennen in het kalibratie-onderzoek.

Enkele bemerkingen:

- Beide afnamemodaliteiten nemen ongeveer even veel tijd in beslag.
- De aangetikte afbeelding groen laten oplichten werkt prima op de tablets. Het is voor de kleuters meteen duidelijk dat ze een antwoord hebben gegeven.
- Kleine details op de afbeeldingen zijn iets moeilijker te zien op de tablets, dit omdat het scherm kleiner is dan een A4-blad.

5.4 Instructies voor toetsafnemers

De instructies mogen nog duidelijker worden geformuleerd voor toetsafnemers die niet betrokken waren bij de ontwikkeling (zoals de leraren in het kalibratie-onderzoek). Zo moet bij elke cluster duidelijk worden aangegeven wat op voorhand moet worden klaargelegd of uitgewerkt (bv. een afgewerkt vingerpopje).

De opdrachten die de toetsafnemer moet geven, werden als helder ervaren. Alleen bij de verhalen mocht duidelijker worden aangegeven wanneer de kleuters iets moeten doen, wanneer iets moet worden getoond, wanneer het volgende stukje van het verhaal mag worden verteld…

Lay-out en instructies mogen beter op elkaar worden afgestemd. De volgorde van de verschillende onderdelen is best ook steeds dezelfde. Een mogelijke toevoeging zijn pictogrammen bij de tekst (bv. voorlezen, instructie…).

5.5 Haalbaarheid van typetaken

5.5.1 Gerichte opdrachten met gestandaardiseerde observatie (9 taken)

○ Gerichte opdrachten met gestandaardiseerde observatie - met fysieke reactie van de kleuter

Literatuur rond toetsing en evaluatie van jonge kinderen wijst op hun behoefte aan speelse, fantasierijke en ‘leuke’ taken, een korte aandachtsspanne en het belang van bevestiging (eerder dan gevoelens van falen) (Hasselgreen, 2000). Gerichte opdrachten met gestandaardiseerde observatie vormen een effectief antwoord op deze behoeften en komen hieraan tegemoet, wat kon worden bevestigd vanuit de ervaringen tijdens de pilootafnames:

- De kleuters vonden dit type opdrachten doorgaans ‘leuk’, ‘tof’…
- De kleuters hadden bij dit type opdrachten voldoende succeservaringen (Mc Kay, 2005).

- Zeker voor beweeglijke kleuters waren taken waarbij ze iets met hun lichaam konden doen een welkome afwisseling, die het mogelijk maakten om nadien hun aandacht beter bij de uitvoering van de andere taken te houden.

De gerichte opdrachten met gestandaardiseerde observatie blijken eerder gemakkelijke taken, en kunnen daarom helpen om te discrimineren tussen kinderen met een matige en zwakke luistervaardigheid. Verdere analyses moeten deze indruk bevestigen (zie 6 ▪ BEVINDINGEN UIT DE KWANTITATIEVE ANALYSE).

Sommige opdrachten van deze typetaak nemen iets meer tijd in beslag omdat ze individueel worden afgenomen. Ze vragen ook meer organisatorische inspanningen omdat de toetsafnemer bepaalde materialen moet verzamelen en de kleuters zich even moeten bezighouden terwijl ze wachten. Om alles toch zo vlot mogelijk te laten verlopen, maken we gebruik van basismaterialen die in elke kleuterklas voorhanden zijn, zoals stiften, hoepel, knuffeldier, stoel, kast of ander meubilair, blokken, poppetjes.

Enkele aandachtspunten:

- In bepaalde situaties moet de toetsafnemer ervoor zorgen dat de kleuter of de materialen terug naar een neutrale positie gebracht worden zodat een handeling in het ene item geen invloed heeft op het volgende item.
  - *Voorbeeld uit 'Hoepel': Item 1 'Ga in de hoepel staan' en item 2 'Hou de hoepel vast voor je lichaam'. Belangrijk dat het kind niet meer de hoepel staat wanneer de toetsafnemer de instructie voor item 2 geeft.*
- De intentie van de kleuter is belangrijk. Sommige kleuters zijn motorisch niet in staat om een bepaalde handeling uit te voeren. Maar als de intentie er is, wordt dit als 'correct' aangeduid. Deze aanvulling moet in het beoordelingsmodel worden opgenomen.
  - *Voorbeeld uit 'Vingerpop': Dichtplakken van de strook van een vingerpop.*
- Opletten met woorden die minder frequent voorkomen in de kleuterklas.
  - *Voorbeeld uit 'Rommel in de klas': spreek over een 'doosje' of 'bakje' in plaats van een 'pennenzak'.*
- Vermijden dat bepaalde handelingen of reacties van de kleuters voor interpretatie vatbaar zijn. Beoordelingsmodel voor de toetsafnemer goed nakijken of uitbreiden met goed en af te keuren alternatieven.
  - *Voorbeeld uit 'Dansje/bewegen': 'Neem je knuffel goed vast. Spring nu samen met je knuffel in het rond zonder hem los te laten'. Op welke manier 'in het rond springen' interpreteren?*

Gerichte opdrachten met gestandaardiseerde observatie - met verbale reactie van de kleuter

Ook de ervaringen in verband met de gestandaardiseerde observaties met verbale reacties sluiten aan bij de literatuur, met name bij het belang van de actieve betrokkenheid. Ook deze taken geven heel wat kleuters succeservaringen (McKay, 2005).

Deze opdrachten vielen in de smaak bij alle kleuters. Ze maakten de testafname actiever en ze gaven aanleiding tot een persoonlijkere invulling door de kleuter. De kleuters kunnen bijvoorbeeld aangeven wat ze graag zelf eten, waar ze graag meespelen, wat ze vandaag gedaan hebben en hoe zij een bepaald dier nadoen. We merkten dat zowel sterk als minder sterk taalvaardige kleuters daar plezier aan beleefden. Met oog op variatie in de clusters van taken bleken deze opdrachten heel waardevol.

De diepgang van de antwoorden van de kleuters was uiteenlopend. Spreekvaardige en extraverte kleuters gaven soms een heel uitgebreide uitleg. Laagtaalvaardige en/of verlegen kleuters gingen op heel summiere wijze toelichting geven bij hun keuzes. Het gaat hier evenwel om de luistervaardigheid, dus zodra de toetsafnemer merkt dat de kleuter de opdracht begrepen heeft, wordt de score 'correct' gegeven.

We hadden de indruk dat deze opdrachten eerder gemakkelijk zijn. Maar enkel op basis van de feedback van de toetsafnemers is het moeilijk om uitspraken te doen over de discriminerende waarde van deze taken. Bijkomende analyses moeten ons hierover meer duidelijkheid geven. Eventueel moeten we het beoordelingsmodel herbekijken als blijkt dat deze taken te weinig discriminerend zijn.

Enkele aandachtspunten:

- De kleuters vinden de taken leuk om te doen, maar ze nemen behoorlijk wat tijd in beslag. Daarom stellen we voor om extra items te koppelen aan deze taken zodat er meer informatie uit kan worden gehaald, zonder een grote extra inspanning.
    - *Voorbeeld uit 'Vandaag'. De toetsafnemer tekent acht activiteiten die de kleuters in de klas gedaan hebben. Dit vraagt wel wat tijd. Nadien volgt er slechts één individuele vraag daarover.*
- Er zit soms wat tijd tussen de voorbereiding (in groep) en de daadwerkelijke vraagstelling. Waar mogelijk zorgen we er best voor dat de vraag sneller volgt op de input, en zorgen we voor gepaste (visuele) ondersteuning zodat kleuters die later aan de beurt komen niet benadeeld zijn.
    - *Voorbeeld uit 'Dieren nadoen'. De kleuters doen in groep(jes/en?) drie dieren na. Daarna komen ze één voor één bij de toetsafnemer en krijgen ze de vraag welk dier ze opnieuw willen nadoen en waarom. Zorgen dat ook de laatste kleuter nog weet dat het gaat over het opnieuw nadoen van de dieren, en niet over het dier dat je het leukst vindt.*
- Bij eventuele aanpassingen van het beoordelingsmodel moeten we erover waken dat steeds de luistervaardigheid getest wordt, en niet de spreekvaardigheid of extraversie.
    - *Voorbeeld uit 'Eten'. 'Kijk goed naar de foto's. Welke dingen vind jij héél lekker. Wat lust jij héél graag?'. Sommige kleuters geven spontaan een hele uitleg over waarom ze iets wel/niet lekker vinden. Andere kleuters duiden gewoon etenswaren aan met de uitleg 'omdat ik dit lekker vind'.*

Doe-en zoekopdrachten op papier/tablet (6 taken)

Net als de gerichte opdrachten met gestandaardiseerde observatie, bevestigen de reacties van de kleuters op de doe- en zoekopdrachten de literatuur rond het toetsen van jongen kinderen, met name als het gaat over actieve opdrachten (Hasselgreen, 2000) en de nood aan bevestiging en succeservaringen (Mc Kay, 2005).

In dit type taak maken we gebruik van één grote prent waarop de kleuters iets aanduiden, iets kleuren of iets tekenen. De prenten spreken de kleuters aan, en de meeste kleuters vinden dit een fijne afwisseling. De opdrachten lijken bovendien aan te sluiten bij typische vragen in een kleuterklas en waren dus herkenbaar. Deze taken lijken eerder gemakkelijk te zijn, en kunnen ons daarom helpen om laagtaalvaardige kinderen te onderscheiden van gemiddeld taalvaardige kinderen. In sterke groepen voelden we dat de kleuters (te) weinig uitdaging hadden aan sommige van deze taken. Verdere analyses moeten dit bevestigen (zie 6.5 MOEILIJKHEIDSGRAAD VAN DE VERSCHILLENDE ITEMS).

Bij de papieren versie is het belangrijk dat de toetsafnemer na elk item meteen registreert of het antwoord van de kleuter juist of fout is. Op het einde van de taak staan er immers meerdere dingen aangeduid of getekend op de prent, en wordt het moeilijker te achterhalen wat bij welk item aan de prent werd toegevoegd door de kleuter. Tijdens de pilootafname hebben we dit gedaan op het papier zelf, in een vakje onder de tekening. Dat werkte goed.

Bij de digitale versie wordt er na elk item doorgeklikt en verschijnt de prent opnieuw, waarop de kleuter dan het volgende antwoord kan aantikken. Omdat telkens dezelfde prent verschijnt op het scherm, kan de toetsafnemer niet in één oogopslag zien of alle kinderen bij het juiste item zijn. Een extra visuele indicatie kan dit oplossen.

Enkele aandachtspunten:

- Zorgen dat er geen verwarring of onduidelijkheid bestaat over de voorwerpen die op de prenten staan en aan bod komen in de instructies.
    - *Voorbeeld uit 'Rommel in de eetzaal': Als er gesproken wordt over een vuile lepel, moet het ook duidelijk zichtbaar zijn welke lepels op de tekening vuil zijn.*
- Voldoende afleiders voorzien. Elementen toevoegen aan sommige tekeningen.
    - *Voorbeeld uit 'Zandtafel'. Er zijn eerder weinig elementen te zien op de tekening, waardoor de items niet onafhankelijk zijn van elkaar. We lossen dit op door extra kinderen en materialen op de tekening te zetten.*

Meerkeuze-opdrachten op papier/tablet (15 taken)

De oefenitems van de meerkeuze-opdrachten werkten goed en moeten zeker behouden blijven. Het maakte op een eenvoudige manier en zonder veel uitleg duidelijk aan de kleuters wat van hen verwacht werd bij deze opdrachten.

Dit soort opdrachten vraagt duidelijk veel concentratie van de kleuters. Deze taken lijken het moeilijkst, en we verwachten dat deze typetaak zal helpen om de sterk taalvaardige kinderen te identificeren. Verdere analyses moeten dit bevestigen (zie ook Hoofdstuk 6, 1.7 RELATIE TUSSEN MOEILIJKHEIDSGRAAD VAN DE ITEMS EN EIGENSCHAPPEN VAN DE ITEMS). Sommige meerkeuze-opdrachten, zeker als het gaat om verhalen begrijpen, lijken te moeilijk voor laagtaalvaardige kleuters. Om te vermijden dat deze kleuters hun motivatie verliezen, moeten taken 'doenbaar' blijven; het kalibratie-onderzoek moet uitwijzen of het mogelijk is om betrouwbare informatie te krijgen over de luistervaardigheid van minder taalvaardige kleuters, zonder de moeilijkste taken voor te leggen aan álle kleuters.

De meeste kleuters hadden na een reeks van drie à vier meerkeuzetaken nood aan een (mentale) pauze. Op dat moment kan een andere typetaak afwisseling brengen.

Enkele aandachtspunten:

- Bij sommige vragen geven de kleuters een antwoord op basis van voorkennis of evidenties, niet zozeer op basis van wat er verteld wordt. Daarom is het goed een aantal minder evidente opties toe te voegen aan de instructies.
    - *Voorbeeld uit 'Myriam': 'Als je naar buiten gaat, moet je je jas dichtdoen'. Het is heel voor de hand liggend dat een kleuter zijn jas dicht doet bij het naar buiten gaan.*
- Alle details op de tekeningen moeten goed zichtbaar zijn, ook op de tablets.
    - *Voorbeeld uit 'Juf is Jarig': 'Leg enkele schijfjes appel op de taart. En leg daarna snoepjes rond de appelschijfjes'. De snoepjes moeten hier duidelijk zichtbaar zijn.*
- Als er emoties afgebeeld worden, is het belangrijk om twijfel uit te sluiten. Concreet betekent dit dat het verhaal coherent in elkaar moet zitten zodat er geen verwarring mogelijk is over hoe het personage zich voelt. Waar mogelijk is het beter te kiezen voor een close-up van het personage, zodat de emotie duidelijk(er) te zien is.
    - *Voorbeeld uit 'Klastaakjes'. Het onderscheid tussen 'droevig', 'boos' en 'bang' is niet heel duidelijk te zien op de tekening.*
- Vermijden dat er een bepaalde systematiek zit in de antwoordopties. Afleiders evalueren en zorgen dat de kleuters het juiste antwoord enkel kunnen weten door goed te luisteren.
    - *Voorbeeld uit 'Mona's hoeken'. Het juiste antwoord is in drie van de vier items de afbeelding met het meeste elementen. Heel wat kleuters lijken automatisch te kiezen voor de 'drukste' afbeelding.*
- Opletten met niet intuïtieve goede keuzes. Zorgen dat er geen twijfel kan zijn over het juiste antwoord.
    - *Voorbeeld uit 'Myriam'. De toetsafnemer vertelt over de gewoontes in de nieuwe klas van Myriam. In die klas is het de gewoonte om het licht aan te laten bij het verlaten van het toilet, terwijl veel kleuters net in het kader van energiezuinigheid leren om lichten te doven.*
- Zorgen dat de toetsafnemer heel goed weet wanneer de kleuters welke afbeeldingen te zien krijgen (tijdens of na het voorlezen).
    - *Voorbeeld uit 'Eenzaam'. De kleuters luisteren naar het verhaal en krijgen de afbeeldingen nog niet te zien. Een introprent kan helpen om meer in de sfeer van het verhaal te komen, en dit moet in de instructie/opleiding onder de aandacht gebracht worden.*

6 ▪ Bevindingen uit de kwantitatieve analyse

Via kwantitatieve analyses brengen we factoren in kaart die een impact hebben op de validiteit en betrouwbaarheid van de taalscreening. Zo willen we zicht krijgen op:

- de werking van de taalscreening als geheel: betrouwbaarheid van de toets als geheel op het niveau van items en kleuters;
- de werking van de items: moeilijkheid, betrouwbaarheid en discriminerend vermogen. Dit laatste gaat na in welke mate een item in staat is om een onderscheid te maken tussen kleuters met een hogere en een lagere taalvaardigheid;
- de werking van de afleiders: moeilijkheid en betrouwbaarheid. We willen met name nagaan of de correcte antwoordoptie gekozen wordt door de kleuters met de hoogste taalvaardigheid, en of verkeerde antwoordopties vooral aantrekkelijk zijn voor kleuters met een lagere taalvaardigheid.

We voeren bovenstaande analyses uit op de verzamelde data van het pilootonderzoek, maar beschouwen de resultaten als indicatief. Omdat de data verzameld werd bij een beperkte groep van kleuters, geven de resultaten uit de analyses vooral richting aan voor de herwerking: ze kunnen de kwalitatieve observaties uit het pilootonderzoek versterken, of indicaties geven dat bepaalde taken, items of afleiders nog eens grondig bestudeerd moeten worden.

6.1 Databestand

Er zijn 30 toetstaken en 139 items. Er hebben 206 kleuters meegedaan met de papieren testafname. We hebben 202 geldige cases (‘persons’) die worden meegenomen in de Winsteps-analyses. 4 kleuters hebben enkel missing values.

6.2 Betrouwbaarheid van het screeninginstrument als geheel

Tabel 3 geeft de betrouwbaarheid van het geheel van alle items weer. Er zijn 5 ‘extreme items’, d.w.z. items die door alle kleuters juist worden beantwoord. Deze werden uit de weergegeven resultaten in onderstaande tabel gehaald.

SUMMARY OF 134 MEASURED (NON-EXTREME) ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	20.4	28.9	.00	.57	.97	-.02	1.05	.07
SEM	.5	.4	.11	.02	.02	.09	.10	.09
P.SD	6.0	5.1	1.31	.18	.28	1.04	1.14	1.04
S.SD	6.0	5.1	1.31	.18	.28	1.04	1.15	1.05
MAX.	32.0	40.0	3.81	1.11	2.15	3.58	9.90	4.46
MIN.	4.0	11.0	-3.77	.40	.46	-2.47	.13	-1.74

REAL RMSE	.62	TRUE SD	1.15	SEPARATION	1.84	ITEM RELIABILITY .77
MODEL RMSE	.60	TRUE SD	1.16	SEPARATION	1.94	ITEM RELIABILITY .79
S.E. OF ITEM MEAN = .11						

MAXIMUM EXTREME SCORE: 5 ITEM 3.6%

Tabel 3: Betrouwbaarheid van de taalscreening in haar geheel

Gemiddeld werd een item door 29 kleuters beantwoord (M=28.9; SD=5.1), waarvan gemiddeld 20,4 keer correct (M=20,4; SD=6,0). Gemiddeld geeft 70,5% van de kleuters een correct antwoord op een item.

De 'item measure' geeft de 'moeilijkheidsscore' van de items weer. Hoe hoger de measure, hoe moeilijker het item is. Het gemakkelijkste item heeft een measure van -3,77 en het moeilijkste item heeft een measure van 3,81. Het gemiddelde van de 'item measures' is altijd 0. De 'person measure' is een maat voor de vaardigheid van een persoon, in ons geval de taalvaardigheid van een kleuter. Als de waarde van de 'person measure' gelijk is aan de waarde van de 'item measure' (moeilijkheidsscore), wil dat zeggen dat deze kleuter 50% kans heeft om dat item correct te beantwoorden. Is de measure van de kleuter groter dan die van het item, dan betekent dit dat de kleuter 'vaardiger' is en meer kans heeft om het item correct te beantwoorden. Is de measure van de kleuter lager dan die van het item, dan wil dat zeggen dat het item (te) moeilijk is voor deze kleuter en neemt de kans op een correct antwoord af.

De betrouwbaarheid van de measures voor de items ('item reliability') is met .77 redelijk goed. Idealiter zit de betrouwbaarheidscore voor de items boven .80 (Linacre, M., 2012).

6.3 Betrouwbaarheid van de metingen op het niveau van de kleuters

Tabel 4 vat de betrouwbaarheid op het niveau van de kleuters samen. Er zijn 14 'extreme cases', d.w.z. 13 kleuters die geen enkele fout hebben gemaakt en 1 kleuter die geen enkele vraag juist heeft. Deze worden automatisch uit de berekening van de betrouwbaarheidsscores gehaald omdat ze geen enkele variatie in hun antwoordenpatroon vertonen.

Gemiddeld hebben de leerlingen (die geen extreme score behaalden) 20 items beantwoord of uitgevoerd (M=19.9, SD=5,5)., waarvan gemiddeld 14 items correct (M=13,9; SD=6,6). Gemiddeld behalen de kleuters daarmee een score van 70% correct beantwoorde of uitgevoerde items.

De 'person measure' geeft de 'vaardigheidsscore' van de kleuters weer. Hoe hoger de measure, hoe meer luistervaardig een kleuter is. De minst luistervaardige kleuter heeft een measure van -3,07 en de meest luistervaardige kleuter heeft een measure van 4,70. Er zijn ook enkele kleuters die minimum of maximum scores haalden op de screening. Hun vaardigheid kunnen we niet voldoende registreren met de behulp van dit screeninginstrument. Als we hun vaardigheidsniveau toch erbij nemen, zit het bereik van de measures tussen -4,38 en 6,01.

De betrouwbaarheid van de measures voor de kleuters ('person reliability') is met .79 redelijk goed. Idealiter zit deze betrouwbaarheidscore boven .80 (Linacre, M., 2012).

SUMMARY OF 188 MEASURED (NON-EXTREME) PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	13.9	19.9	1.20	.68	.98	.03	1.03	.11
SEM	.5	.4	.12	.01	.02	.06	.07	.06
P.SD	6.6	5.5	1.60	.20	.26	.82	1.00	.88
S.SD	6.6	5.5	1.60	.20	.26	.82	1.01	.88
MAX.	31.0	32.0	4.70	1.26	1.99	2.74	9.45	3.57
MIN.	1.0	6.0	-3.07	.42	.37	-2.21	.10	-1.54

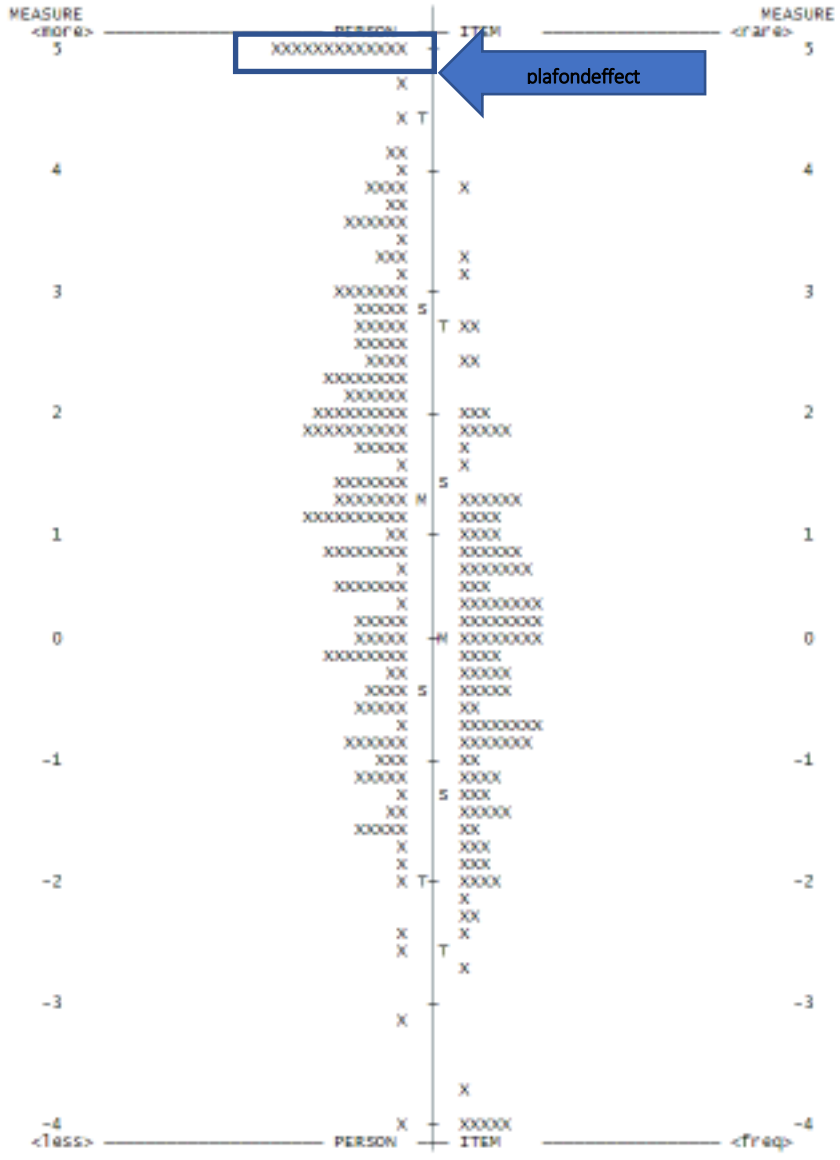
REAL RMSE	.74	TRUE SD	1.42	SEPARATION	1.91	PERSON RELIABILITY .79
MODEL RMSE	.71	TRUE SD	1.43	SEPARATION	2.01	PERSON RELIABILITY .80
S.E. OF PERSON MEAN = .12						

MAXIMUM EXTREME SCORE: 13 PERSON 6.4%
MINIMUM EXTREME SCORE: 1 PERSON .5%
LACKING RESPONSES: 4 PERSON

Tabel 4: Person Reliability

6.4 De verhouding tussen de moeilijkheidsgraad van de toets en de vaardigheid van de kleuters (Wright Map)

FIGUUR 12 biedt een visuele voorstelling van de verhouding tussen moeilijkheidsgraad van de items en de vaardigheidsscores van de kleuters.



Figuur 12: Visuele voorstelling van de verhouding tussen de moeilijkheidsgraad van de toets en de vaardigheid van de kleuters (Wright Map)

Links zien we de vaardigheidsscores van de kleuters, rechts de moeilijkheidsscores van de items. Wanneer vaardigheidsscore en moeilijkheidsscore op dezelfde hoogte staan, betekent dit dat het kind 50% kans heeft om dit item correct te beantwoorden. We stellen vast dat:

- zowel de measures voor de items als die voor de kinderen een normaal verdeelde curve hebben, weliswaar met een plafondeffect aan de kant van de 'person measures' (zie hierboven). De normaalverdeling bij de kleuters geeft aan dat we zelfs in deze beperkte pilootafname een goede vertegenwoordiging hebben gevonden in de geteste kleuters. De normaalverdeling bij de items geeft aan dat de gebruikte items in onze screening voldoende variëren.

- het gemiddelde (M) voor vaardigheid (person) een stuk hoger ligt dan voor moeilijkheid (item). Iemand met een gemiddelde taalvaardigheid zal met andere woorden meer dan de helft van de items correct oplossen. Het is dus een eerder 'gemakkelijk' instrument. Aan de bovenkant van de person measure zien we een groep kleuters die (bijna) het maximum scoren. Voor deze groep zijn er geen items meer die corresponderen met hun vaardigheidsmeasure. Dit wijst op een plafondeffect.
- er tussen measure -1 en +1 voldoende items lijken te zijn die corresponderen met de vaardigheidsscore van de kinderen. Dat wijst erop dat we met deze screening de laagtaalvaardige kleuters van de 'gemiddeld' taalvaardige kleuters kunnen onderscheiden.
- er voor de sterke tot zeer sterk taalvaardige kleuters minder items zijn. Vooral voor de kleuters met een vaardigheidsscore tussen 1 en 3 zijn er nu niet zoveel 'corresponderende items'.
- er voldoende gemakkelijke items lijken te zijn (tussen -2 en 0), we vinden dus zeker genoeg items om lagere taalvaardigheidsniveaus te onderscheiden.

6.5 Moeilijkheidsgraad van de verschillende items

6.5.1 Moeilijkste items (measure > 2)

Tabel 5 geeft meer informatie over de werking van de moeilijke items.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
16	7	22	3.81	.52	1.18	.77	1.07	.29	.37	.48	66.7	75.4	3.4cor
87	4	30	3.20	.60	1.17	.60	2.30	1.21	.22	.40	90.0	87.4	28.2car
134	5	31	3.13	.55	1.34	1.00	1.46	.81	.18	.41	80.6	85.9	51.1cor
17	11	21	2.65	.49	1.12	.71	1.27	1.07	.35	.45	65.0	66.6	3.5car
33	14	34	2.62	.45	.95	-.18	1.19	.52	.60	.60	77.4	76.2	6.4cor
62	15	34	2.42	.44	.74	-1.38	.59	-.85	.71	.61	83.9	75.7	14.4car
125	14	30	2.35	.41	1.27	1.69	2.03	2.38	.22	.45	53.3	68.8	43.2cor

Tabel 5: Gedetailleerde informatie over moeilijke items

We bekijken alle moeilijkste items (measure > 2) meer in detail. Daarnaast kijken we ook naar de misfits, de items die er niet goed in slagen de relevante informatie over de vaardigheid van een persoon mee te geven. Deze misfits zijn te onderscheiden door te kijken naar de 'infit' en 'outfit' statistieken.

Een hoge infitwaarde voor een item betekent dat er veel onverwachte antwoorden komen voor dit item voor personen met vaardigheidsscore die overeen komt met de moeilijkheidsgraad van het item. Hoge outfit-waarden vind je bij items die erg gemakkelijk of moeilijk zijn, waarbij een persoon met erg lage of hoge taalvaardigheid onverwacht antwoordt.

Items met een waarde groter dan 2 zijn outliers en bekijken we meer in detail. Alle 'infit'-waarden zijn lager dan 2. Er zijn 9 'outfit'-waarden hoger dan 2. Het gaat om item 28.2, item 43.2, item 28.4, item 7.1, item 32.5, item 17.4, item 32.2, item 43.1 en item 46.2.

We bestudeerden deze 9 items om de reden voor misfit te achterhalen en pasten het item, indien nodig, aan. In bijlage 2 (overzicht aanpassingen items) geven we een overzicht van alle items die we aanpasten of overwogen aan te passen voor het kalibratie-onderzoek op basis van misfit. Hieronder geven we één voorbeeld.

Item 6.4 (LIEVELINGSBOEKEN): *Let op: nu gaan we op zoek naar het boek dat Kobe **niet** kiest. Kobe houdt van boeken over **dieren**. Hij luister het liefst naar verhalen die gaan **over honden en over grote dieren**. Verhalen over **kleine diertjes** vindt hij **niet leuk**. Welk boek kiest Kobe **niet**? Trek een kring rond het boek dat Kobe niet kiest.*

- 14/34 kleuters geven hier een correct antwoord (boek met muis). De ontkenning in de vraagstelling maakt het hier moeilijk. Het kan zijn dat deze vraag goed discrimineert tussen goede en sterke kleuters, maar dat sommige kleuters de ontkenning niet opmerken om redenen die niet met taalvaardigheid te maken hebben. De afleiders zijn goed, want geen van de drie wordt opvallend vaker gekozen.
- De vraagformulering werd nog eens goed bekeken.
- Het item werd uiteindelijk niet aangepast maar opnieuw uitgeprobeerd. In het kalibratieonderzoek was het opnieuw een eerder moeilijk item (measure 1,36) maar niet problematisch (geen hoge outfit waarden) en met voldoende discriminerende waarde (0,90).

6.5.2 Gemakkelijkste items (measure < -2)

Ook de gemakkelijke items, met een vaardigheidsniveau lager dan -2, bekijken we apart. Tabel 6 geeft meer informatie over de psychometrische waarden van de gemakkelijk items.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
51	17	21	-2.02	.60	.63	-1.10	.44	-.38	.57	.34	90.5	82.7	12.3cor
14	22	22	-2.02	1.87	MINIMUM MEASURE				.00	.00	100.0	100.0	3.2cor
93	31	32	-2.14	1.08	.73	-.07	.13	-.58	.46	.26	96.6	96.6	29.4cor
124	29	30	-2.26	1.11	1.44	.75	3.26	1.50	-.03	.32	96.7	96.7	43.1cor
28	32	33	-2.28	1.06	.68	-.11	[illegible]	[illegible]	.42	.28	96.4	96.5	4.1cor
129	21	24	-2.45	.67	1.41	.98	6.95	2.76	-.09	.32	83.3	87.5	46.2cor
39	31	34	-2.65	.64	.79	-.35	.40	-.30	.39	.28	91.2	91.1	8.2cor
97	30	30	-3.12	1.97	MINIMUM MEASURE				.00	.00	100.0	100.0	32.3cor
98	30	30	-3.12	1.87	MINIMUM MEASURE				.00	.00	100.0	100.0	32.4cor
81	25	25	-3.16	1.94	MINIMUM MEASURE				.00	.00	100.0	100.0	24.1cor
118	28	28	-3.44	1.91	MINIMUM MEASURE				.00	.00	100.0	100.0	40.5cor
58	20	21	-3.77	1.05	.73	-.87	.20	-.41	.39	.28	95.2	95.2	12.2cor

Tabel 6: Gedetailleerde informatie over gemakkelijke items

Een aantal items worden door alle kleuters correct beantwoord of uitgevoerd. Voor deze items kunnen geen outfit-statistieken berekend worden. We kunnen aannemen dat deze items voor bijna niemand een uitdaging zijn. We bekijken daarom op welke manier we deze items kunnen aanpassen, met daarbij de bedenking dat enkele (zeer) gemakkelijke items in de screening ervoor kunnen zorgen dat alle kleuters een succeservaring hebben.

In Bijlage 2 (overzicht aanpassingen items) geven we een overzicht van alle items die we aanpasten of overwogen aan te passen. Hieronder geven we één voorbeeld.

Item 40.5 (KLASAFSPRAKEN): *Dit betekent dat als de kinderen de **planten water geven**, ze **een schort** moeten aandoen. Op welke tekening doen de kinderen het goed? Trek een kring rond de juiste tekening.*

- Misschien is dit item te vanzelfsprekend?
- Mogelijke oplossingen: afleiders herbekijken, vraag schrappen of vraag vervangen?
- De vraagstelling werd aangepast: *'Dit betekent dat de kinderen de planten moeten water geven met een gieter. Om hun kleren niet nat te maken, moeten ze ook een schort aandoen. Op welke tekening doet het kind het goed?'*. Deze aanpassing was succesvol: in het kalibratie-onderzoek is het item nog steeds eerder gemakkelijk (measure -0,69) maar niet meer té voor de hand liggend. De discriminerende waarde van het item in het kalibratie-onderzoek is goed (1,11).

De overige items met een hoge outfit-waarde (outliers) worden besproken bij Tabel 7.

6.5.3 Item polariteit

We verwachten dat items met een lage measure (gemakkelijke items) correct beantwoord worden door kleuters met een hoge measure (hoge luistervaardigheidsscore). We vragen ons af of er gemakkelijke items zijn waarop sterk taalvaardige kleuters, tegen de verwachtingen in, toch vaak een fout antwoord geven. We gaan daarvoor op zoek naar items met een lage PTMEASURE. Een negatieve waarde of lage waarde (< 0.20) vraagt om nader onderzoek. Tabel 7 identificeert de items met een slechte item polariteit.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
129	21	24	-2.45	.67	1.41	.98	6.95	2.76	-.09	.32	83.3	87.5	46.2cor
89	9	31	1.82	.48	2.15	3.58	5.42	4.46	-.09	.53	50.0	79.6	28.4car
124	29	30	-2.26	1.11	1.44	.75	3.26	1.50	-.03	.32	96.7	96.7	43.1car
72	16	19	-1.39	.70	1.28	.75	9.90	3.31	-.01	.39	84.2	85.2	17.4cor
96	29	30	-1.82	1.07	1.27	.57	3.18	1.47	.00	.23	96.8	96.8	32.2car
99	26	30	-.03	.61	1.70	1.63	2.80	1.91	.02	.40	76.0	86.1	32.5cor
101	20	33	.69	.49	1.42	2.45	1.98	3.19	.12	.46	61.3	69.2	33.1car
134	5	31	3.13	.55	1.34	1.00	1.46	.81	.18	.41	80.6	85.9	51.1cor

Tabel 7: Item Polariteit

Er zijn vier items met een negatieve PT-MEASURE. Sterk taalvaardige kleuters halen hier dus (tegen de verwachtingen in) vaker score 0. Daarnaast zijn er nog vier items met een (zeer) lage PT-MEASURE. Dit betekent dat de prestaties van de kleuters niet voldoende in de lijn der verwachtingen liggen.

Verder kijken we naar de kolom met PT-MEASURE CORRELATION, i.e. de correlatie tussen de antwoorden van de kleuters en hun measures (vaardigheidsscore). We verwachten een (sterk) positieve correlatie als het antwoord juist is en een (sterk) negatieve correlatie als het antwoord fout is. Afwijkingen wijzen op een probleem met het item.

In Bijlage 2 (overzicht aanpassingen items) geven we een overzicht van alle items die we aanpasten of overwogen aan te passen. Hieronder geven we één voorbeeld.

Item 32.2 (WAAR IS): *Welk kind steekt zijn **vinger in de lucht**? Zet een kruisje op het kind dat zijn vinger in de lucht steekt.*
Item 32.5 (WAAR IS): *Welke kinderen luisteren **niet** naar de juf? Zet een kruisje bij de kinderen die niet naar de juf luisteren.*

- Deze items hebben hoge outfit-waarden (>2), zie Tabel 7. Weinig kleuters (1/30 voor 32.2 en 4/30 voor 32.5) geven hier een fout antwoord. Het gaat dus om gemakkelijke items waar enkele kleuters van wie we het niet verwachten toch de mist in gaan. Misschien worden kleuters die geen uitdaging ervaren aan de taak (te) nonchalant?
- De PT-MEASURE correlaties benaderen 0 bij een correct antwoord (.00 voor item 32.2 en .02 voor item 32.5).
    - Mean Ability indien fout antwoord: 2.55 (item 32.2) en 2.45 (item 32.5)
    - Mean Ability indien juist antwoord: 2.54 (item 32.2) en 2.55 (item 32.5)
- Uit de feedback van de toetsafnemers en de 'item measures' blijkt ook dat taak 32 (WAAR IS) herwerkt moet worden.
- Het item 'welk kind steekt zijn vinger in de lucht' wordt geschrapt in het kalibratieonderzoek.
- Het andere item wordt herwerkt naar 'welk meisje praat tegen een vriendje'. 'Niet luisteren naar de juf' is immers voor interpretatie vatbaar. Het praten tegen het vriendje is duidelijker te zien. Dit heeft gewerkt. De PT-measure voor het item ligt in het kalibratie-onderzoek op 0,26 en de outfit MNSQ <2. Het is een gemakkelijk item (-1,38) met voldoende discriminerende waarde (0,94).

6.6 Vaardigheid van de kleuters

6.6.1 Kleuters met een (zeer) hoge taalvaardigheid (measure > 3,9)

Tabel 8 geeft meer informatie over mate waarin de screening informatie geeft voor kleuters met een zeer hoge taalvaardigheid.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	PERSON
135	26	26	6.01	1.86	MAXIMUM MEASURE				.00	.00	100.0	100.0	DA/C07/K2
159	19	19	5.81	1.86	MAXIMUM MEASURE				.00	.00	100.0	100.0	KB/CEXTRA1/K1
161	13	13	4.94	1.87	MAXIMUM MEASURE				.00	.00	100.0	100.0	KB/CEXTRA1/K3
183	32	32	4.78	1.84	MAXIMUM MEASURE				.00	.00	100.0	100.0	DA/CX2/K1
184	32	32	4.78	1.84	MAXIMUM MEASURE				.00	.00	100.0	100.0	DA/CX2/K2
134	25	26	4.70	1.07	1.03	.32	.39	-.11	.26	.23	95.7	95.6	DA/C07/K1
155	20	21	4.41	1.10	.57	-.37	.13	-.56	.46	.28	94.7	94.7	KB/CCR1/K2
189	25	25	4.32	1.84	MAXIMUM MEASURE				.00	.00	100.0	100.0	DA/CX3/K2
66	18	18	4.30	1.84	MAXIMUM MEASURE				.00	.00	100.0	100.0	DA/C04/K1
70	18	18	4.30	1.84	MAXIMUM MEASURE				.00	.00	100.0	100.0	DA/C04/K5
121	29	30	4.15	1.03	.97	.26	.40	-.08	.23	.17	96.7	96.6	DA/C06/K7
170	21	22	4.14	1.13	.51	-.50	.10	-.66	.53	.34	95.8	95.0	KB/CEXTRA2/K2
23	19	19	3.99	1.86	MAXIMUM MEASURE				.00	.00	100.0	100.0	SM/C02/K3
76	15	15	3.97	1.86	MAXIMUM MEASURE				.00	.00	100.0	100.0	MD/C05/K1
82	15	15	3.97	1.86	MAXIMUM MEASURE				.00	.00	100.0	100.0	DA/C05/K2
83	15	15	3.97	1.86	MAXIMUM MEASURE				.00	.00	100.0	100.0	DA/C05/K3
113	25	26	3.94	1.04	.82	.06	.22	-.36	.34	.20	96.2	96.1	MI/C06/K4

Tabel 8: Vaardigheid van de kleuters

We stellen vast dat 12 kleuters de maximumscore hebben behaald. In totaal zijn er 17 kleuters (8,5%) met een (zeer) hoge measure. Kleuters uit één school lijken oververtegenwoordigd, dus mogelijk houdt dit vrij grote aantal verband met sterk taalvaardige kleuters uit één - vrij grote - school.

6.6.2 Kleuters met een (zeer) lage taalvaardigheid (measure < -2)

Tabel 9 geeft meer informatie over mate waarin de screening informatie geeft over kleuters met een zeer lage taalvaardigheid.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	PERSON
26	3	16	-2.00	.67	1.16	.53	1.15	.48	.14	.28	81.3	81.0	WW/C02/K1
2	4	15	-2.35	.65	.70	-.90	.53	-1.03	.70	.43	86.7	77.4	SM/C01/K2
10	4	21	-2.55	.63	1.45	1.20	4.47	2.15	.08	.43	71.4	83.5	WW/C01/K5
29	1	11	-3.07	1.06	.97	.22	.64	.01	.30	.20	90.9	90.7	WW/C02/K4
34	0	12	-4.38	1.85	MINIMUM MEASURE				.00	.00	100.0	100.0	WW/C02/K9

Tabel 9: Kleuters met een lage taalvaardigheid

Slechts één kleuter beantwoordt geen enkele vraag juist. In totaal zijn er 5 kleuters (2,5%) met een measure lager of gelijk aan - 2. Zij behalen een zeer lage score en van hen kunnen we veronderstellen dat het zeer laagtaalvaardige kleuters zijn. Vier van deze vijf kleuters met een (zeer) lage vaardigheidsscore hebben een beperkt aantal items beantwoord (< 20 items). Het gaat dus om kleuters met een hoog aantal missings. Op basis van de kwalitatieve data (observaties en notities)

kan een mogelijke verklaring zijn dat deze kleuters de instructies van de leraar niet goed hebben begrepen, waardoor ze niets hebben omcirkeld, aangeduid of getekend.

De groep met een (zeer) lage taalvaardigheidsscore is klein (2,5%). Dat betekent dat de meeste kleuters succeservaringen kunnen hebben.

6.6.3 Kleuters met een niet-betrouwbaar resultaat

De 'outfit' statistiek helpt ons om outliers op te sporen onder de kleuters. We gaan op zoek naar leerlingen die tegen de verwachtingen een correct antwoord geven op (heel) moeilijke of een foutief antwoord geven op (heel) makkelijke items. Een beter zicht op de kleuters die niet-betrouwbare antwoorden geven, is belangrijk om de screening als geheel te verbeteren.

De 'outfit' waarden moeten rond '1' liggen. Bij een hogere waarde is er sprake van 'underfit': d.w.z. onvoorspelbare antwoorden o.b.v. het Rasch model. Bij een lagere waarde is er sprake van 'overfit', d.w.z. te voorspelbare antwoorden o.b.v. het Rasch model. MNSQ Outfit waarden groter dan 2 zijn problematisch. Deze leerlingen zijn outliers: hun resultaten verstoren het model.

In totaal worden 13 kleuters gedetecteerd als 'outlier' (6,4% van het sample).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	PERSON
146	18	19	3.79	1.07	1.29	.60	9.45	2.84	A -.20	.25	94.7	94.7	SM/C00/K3
17	21	23	2.56	.86	1.42	.81	6.79	2.42	B .05	.40	87.0	92.8	RD/C01/K2
89	21	24	2.31	.68	.95	.01	4.65	2.33	C .23	.36	90.9	86.3	DA/C05/K9
10	4	21	-2.55	.63	1.45	1.28	4.47	2.15	D .08	.43	71.4	83.5	WW/C01/K5
136	21	26	2.55	.57	1.37	1.19	4.15	2.94	E .19	.44	73.9	88.9	DA/C07/K3
120	9	28	-1.07	.48	1.75	2.74	3.29	3.57	F -.06	.51	60.7	76.5	DA/C06/K6
31	8	19	-.44	.52	.88	-.57	2.56	2.90	G .38	.43	84.2	78.2	WW/C02/K6
178	3	15	-1.60	.72	1.55	1.41	2.56	1.44	H -.02	.42	73.3	80.8	WW/CX1/K5
176	8	15	.47	.61	1.73	2.28	2.52	2.68	I .02	.53	46.7	73.4	WW/CX1/K3
91	11	14	1.91	.71	1.41	1.16	2.28	1.47	J -.05	.38	71.4	78.5	MD/C06/K1
81	14	15	2.64	1.08	1.28	.59	2.21	1.11	K -.02	.27	92.9	92.8	DA/C05/K1
11	8	23	-1.34	.51	1.46	1.75	2.16	1.78	L .18	.50	60.9	75.7	WW/C01/K6
9	12	19	.01	.56	1.32	1.29	2.10	1.73	M .29	.51	57.9	74.0	WW/C01/K4

Tabel 10: Kleuters met onverwachte resultaten

7 ▪ Conclusies Pilootonderzoek

7.1 Betrouwbaarheid van de taalscreening als geheel

De betrouwbaarheid op het niveau van de kleuter ligt in dit kleinschalig pilootonderzoek net op de grens voor 'voldoende' (.80). De itembetrouwbaarheid ligt net onder de grens voor 'voldoende' (.79). Beide nipt voldoende maten hoeven niet problematisch te zijn, aangezien dit pilootonderzoek kleinschalig was en geen representativiteit nastreefde. Bovendien kunnen aanpassingen aan de taken en items de itembetrouwbaarheid en de betrouwbaarheid op het niveau van de kleuter nog doen stijgen.

7.2 Bereik van de taalscreening

Zowel de measures voor de items als die voor de kleuters hebben een normaalverdeelde curve, waarbij het gemiddelde hoger ligt voor de vaardigheid van de kleuters dan voor de moeilijkheidsgraad van de items.

In het pilootonderzoek zijn er aanwijzingen voor een plafondeffect (8,5% van de kleuters heeft 95% of meer items correct). Dit hoeft geen probleem te zijn omdat we de screening in de eerste plaats willen inzetten om risicokleuters op te sporen. We moeten er wel over waken dat de taalscreening ook voor scholen met veel taalsterke kleuters interessant en motiverend blijft (piste van adaptief toetsen verder verkennen) (zie ook Hoofdstuk 3, 2.7 IMPLEMENTATIE EN GEBRUIK VAN EEN SCREENINGSINSTRUMENT).

In het pilootonderzoek zijn aanwijzingen dat de groep kleuters met zeer weinig correcte antwoorden (<20% correcte items) zeer beperkt is. Ook taalzwakke kinderen kunnen dus succeservaringen hebben, wat belangrijk is voor hun motivatie.

7.3 Betrouwbaarheid van de items

De items werden uitvoerig gescreend om eventuele problemen in het instrument op te sporen en bij te sturen. Zo werden moeilijke items, gemakkelijke items en items met een hoge outfit-waarde en lage itempolariteit geïdentificeerd.

7.4 Herwerking

De geïdentificeerde items werden verder onderzocht. We keken naar het aantal correcte/foutieve antwoorden, de afleiders en de feedback van toetsafnemers en resonansgroep voor deze items. Op basis van deze triangulatie werden de items herwerkt voor het kalibratieonderzoek (zie Hoofdstuk 4 7.4 HERWERKING).

HOOFDSTUK 5: KALIBRATIEONDERZOEK

Via een grootschalig kalibratie-onderzoek bij bijna 2000 kleuters werd vastgesteld of de afname van bepaalde toetsitems tot voldoende betrouwbare metingen kan leiden, en onder welke omstandigheden dat het geval is.

1 ▪ Steekproef Kalibratieonderzoek

1.1 Methode: gestratificeerde steekproef

De onderzoekspopulatie werd bepaald op basis van de beschikbare data aan het einde van schooljaar 2019-2020. Binnen deze data focusten we op de kleuters, geboren in 2015: de kleuters die in het schooljaar van het kalibratie-onderzoek volgens Onderwijsdecreet XXX leerplichtig zijn op de leeftijd van vijf jaar.

Deze volledige populatie van kleuters geboren in 2015 die schoollopen in Vlaamse basisscholen (N= 72 309), inclusief de Nederlandstalige scholen gelegen in het Brussels Hoofdstedelijk Gewest bevindt zich in 2306 scholen. Omdat KOALA als doel heeft kleuters te identificeren die moeilijkheden ervaren met onderwijstaal Nederlands, kozen we ervoor om de selectie te beperken tot de 644 scholen en 20 419 kleuters in Vlaanderen en Brussel waarbij minimum 25% van de schoolpopulatie aantikt op de indicator 'opleidingsniveau van de moeder'. Voorgaand onderzoek heeft immers uitgewezen dat SES (hier bepaald op basis van opleidingsniveau van de moeder), meer nog dan de indicator Nederlands niet-thuistaal, een goede voorspeller is voor het beheersen van de onderwijstaal (OECD, 2004; peiling Nederlands luisteren, 2019). Bovendien zien we in de selectie van deze 25% SES-kleuters dat de indicator 'SES' en indicator 'thuistalen' op schoolniveau heel vaak samen voorkomen.

Uit deze populatie van 644 scholen hebben we een gestratificeerde toevalssteekproef getrokken van 2000 kleuters. Dit aantal ligt hoger dan de vooropgestelde 1500 kleuters, wat de ruimte biedt om van bij aanvang rekening te houden met uitval omwille van ziekte van kleuters of leraren (die hoger zal liggen dan bij ander vergelijkbaar onderzoek, omwille van de COVID-19-pandemie). Strata waren 'provincie' (5 Vlaamse provincies plus Brussels Hoofdstedelijk gewest) en 'onderwijsverstrekker' (gemeenschapsonderwijs, vrij gesubsidieerd onderwijs, officieel gesubsidieerd onderwijs). Door verderzetting van de toevalssteekproef na selectie van 2000 kleuters werden voor elk van de strata vijf reservescholen geselecteerd.

Als return voor hun deelname kregen alle scholen een uitgebreid schoolfeedbackrapport (zie Bijlage 3: Anoniem schoolfeedbackrapport) met daarin de schoolresultaten, een stappenplan om met het schoolfeedbackrapport aan de slag te gaan en een scenario voor een personeelsvergadering met het team over de taalscreening. De inhoud van het schoolrapport werd toegelicht in een online sessie en schoolteams kunnen in de loop van april-mei aan een CTO-medewerker vragen stellen bij

onduidelijkheden. Dit schoolfeedbackrapport kan uiteindelijk ook als voorbeeld dienen voor de output van het definitieve instrument.

1.2 Vooropgestelde steekproef

Tabel 11 geeft een overzicht van de verdeling van de scholen over de verschillende strata van de steekproef.

Provincie	Gemeenschaps-onderwijs	Vrij gesubsidieerd onderwijs	Officieel gesub-sidieerd onderwijs	TOTAAL
Brussels HG	1	2	1	4
Vlaams-Brabant	3	6	2	11
Antwerpen	5	12	6	23
Limburg	2	7	1	10
West-Vlaanderen	3	10	2	15
Oost-Vlaanderen	5	9	4	18
TOTAAL	19	46	16	81

Tabel 11: Overzicht verdeling geselecteerde scholen steekproef kalibratie-onderzoek

In Tabel 11 kunnen we lezen dat, om via de gestratificeerde steekproef te komen tot de 2000 kleuters die vooropgesteld zijn, we 81 scholen geselecteerd hebben. De selectie vormt een mooie afspiegeling van de verdeling over de onderwijsverstrekkers en provincies plus het Brussels Hoofdstedelijk Gewest.

Tabel 12 vergelijkt de gemiddeldes van twee OKI-indicatoren op schoolniveau uit de steekproef met de gemiddeldes van deze OKI-indicatoren in de populatie.

	Aantal Scholen	Gemiddelde OKI	Gemiddelde Op-leiding moeder	Gemiddelde Gezinstaal
Populatie	2302	0,90	20,24%	23,56%
Vooropgestelde steekproef	81	1,85	42,81%	48,54%

Tabel 12: Schoolkenmerken: populatie- vooropgestelde steekproef

Uit Tabel 12 kunnen we afleiden dat de gemiddelde OKI-waarde van de scholen in deze steekproef 1,85 bedraagt en dus hoger ligt dan het populatie-gemiddelde (0,90). Dit is ook het geval voor het gemiddeld aantal meertalige leerlingen op schoolniveau (48,54% tegenover 23,56% in de volledige

populatie), en het gemiddeld aantal leerlingen met een moeder met een laag opleidingsniveau (42,82% tegenover 20,24% in de volledige populatie). Deze afwijkingen zijn het gevolg van de bewuste keuze om te selecteren uit scholen waarvan 25% of meer van de leerlingen aantikken op de indicator 'opleidingsniveau moeder'.

In oktober werden mails uitgestuurd naar de 81 scholen uit de steekproeftrekking met een verzoek tot deelname aan het kalibratie-onderzoek. Eind oktober zegde ongeveer de helft van de geselecteerde scholen toe om deel te nemen aan het kalibratieonderzoek. Bij weigering werd onmiddellijk een reserveschool van de bijbehorende straat aangeschreven of een school die zich vrijwillig had aangemeld (maar met gelijkaardige populatie) opgenomen. Redenen om te weigeren hielden meestal verband met corona, zowel de organisatorische moeilijkheden als de veiligheidsvoorschriften die gepaard gaan met de corona-maatregelen. Eén school had deontologische bezwaren tegen het afnemen van toetsen bij kleuters.

1.3 Deelnemende scholen

De lijst van deelnemende scholen (zie Bijlage 4: Vergelijking steekproeftrekking taalscreening versus reële set voor dataverzameling) kende nog enkele grondige wijzigingen omwille van corona (zie Hoofdstuk 5: 2 ▪ DATAVERZAMELING IN CORONATIJDEN) ten opzichte van de vooropgestelde en aanvankelijk aangeschreven scholen. Desalniettemin konden we bereiken dat voldoende kleuters deelnamen uit de verschillende strata.

Provincie	Gemeenschaps-onderwijs	Vrij gesubsidieerd onderwijs	Officieel gesub-sidieerd onderwijs	TOTAAL
Brussels HG	2	2	1	5
Vlaams-Brabant	3	7	2	12
Antwerpen	4	9	3	16
Limburg	5	9	2	16
West-Vlaanderen	2	9	2	13
Oost-Vlaanderen	5	14	4	23
TOTAAL	21	50	14	85

Tabel 13: Overzicht verdeling deelnemende scholen

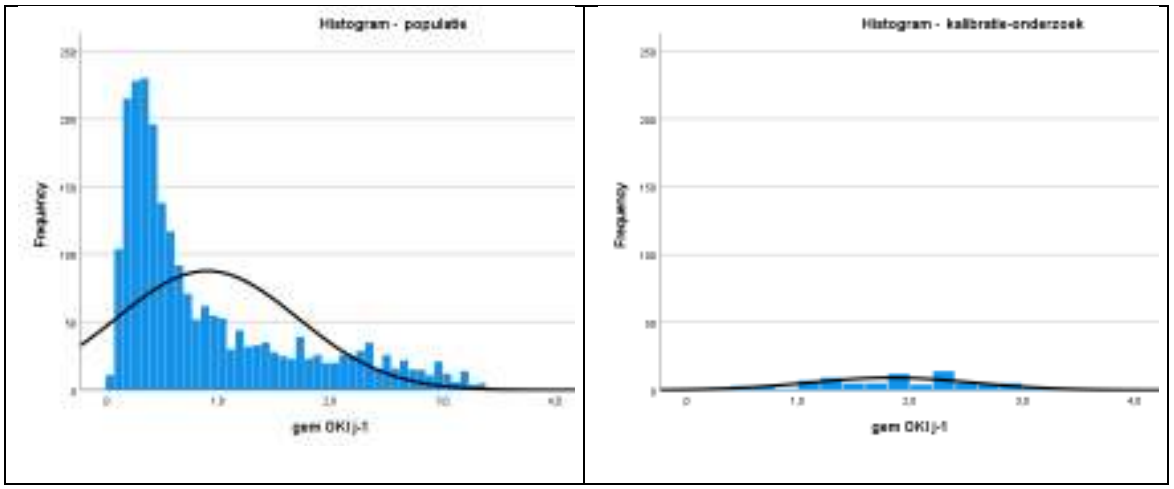
Net als in de oorspronkelijk bedoelde steekproef bereikten we meer dan 80 scholen: waar we streefden naar 81 scholen, hebben we uiteindelijk zelfs 85 scholen kunnen opnemen in het kalibratie-onderzoek verspreid over de verschillende provincies en netten.

Net als in de vooropgestelde steekproef ligt de gemiddelde OKI-waarde van de scholen die deelnamen aan het kalibratie-onderzoek met 1,85 hoger dan het populatie-gemiddelde (0,90). Ook het gemiddeld aantal meertalige leerlingen (steekproef 48,54%, deelnemende scholen 46,57%) en het gemiddeld aantal leerlingen met een moeder met een laag opleidingsniveau (steekproef 42,81%, deelnemende scholen 40,99%) is nauwelijks verschillend van de vooropgestelde steekproef. Tabel 14 zet de belangrijke elementen uit de vergelijking tussen populatiekenmerken, vooropgestelde steekproef en uiteindelijke scholen deelgenomen aan het kalibratie-onderzoek nog eens op een rij.

	Aantal scholen	Gemiddelde OKI	Gemiddelde Opleiding moeder	Gemiddelde Gezinstaal
Populatie	2302	0,90	20,24%	23,56%
Vooropgestelde steekproef	81	1,85	42,81%	48,54%
Deelnemers kalibratie-onderzoek	85	1,83	40,99%	46,57%

Tabel 14: Vergelijking schoolkenmerken: populatie- vooropgestelde steekproef - uiteindelijk deelgenomen scholen

De gemiddelde OKI-waarden van de populatie en de deelnemers aan het kalibratie-onderzoek worden in Figuur 13 tegenover elkaar afgezet in een histogram.



Figuur 13: Vergelijking gemiddelde OKI: populatie - scholen kalibratie-onderzoek

Uit deze figuur kunnen we afleiden dat de steekproef zoals vooropgesteld een oververtegenwoordiging heeft van scholen met een hogere OKI-waarde, waarbij de top van de normaalverdeling in de populatie rond een gemiddelde OKI-waarde van 1 ligt. In het kalibratie-onderzoek ligt deze top bijna op de gemiddelde OKI-waarde van 2.

1.4 Leerlingen in de steekproef

Via de databanken van het Departement Onderwijs en Vorming werden de achtergrondgegevens op leerlingenniveau en schoolniveau opgevraagd van de geselecteerde scholen. Op die manier werden scholen niet met extra administratie belast bij deelname aan het kalibratie-onderzoek.

1983 kleuters namen deel aan de dataverzameling voor het kalibratie-onderzoek. Dat is meer dan het vooropgesteld minimum van 1500 leerlingen.

2 ▪ Dataverzameling in coronatijden

In deze paragraaf beschrijven we telkens de specifieke problemen die scholen ondervonden door de Covid-19-problemen en de ingrepen die we ondernamen om hiermee om te gaan.

2.1 Respons

Scholen ondervonden een hoge werkdruk door de coronamaatregelen en door de vele afwezigheden (door ziekte, wachten op resultaten coronatest, besmetting met corona) waardoor het vaak niet hun belangrijkste prioriteit was om onze vraag naar deelname te bevestigen. Bovendien beperkte de verlengde herfstvakantie de tijd voor communicatie.

OPLOSSING

Een vrijwillig medewerker werd ingeschakeld om scholen op te bellen, naast de reeds aangeworven junior. Op die manier moesten we niet wachten op antwoord op mails en konden we onmiddellijk nadat ze weer openden veel scholen op korte termijn bereiken.

2.2 Toetsassistenten

Toetsassistenten vinden was geen sinecure: we konden geen beroep doen op gepensioneerde leraren (risicogroep voor corona), we moesten de mobiliteit van toetsassistenten beperken om het besmettingsgevaar tussen scholen te verkleinen en een toetsassistent voor de digitale afnames voor de provincie Oost-Vlaanderen moest in quarantaine.

OPLOSSING

We schakelden meer toetsassistenten in die op een beperkt aantal scholen kwamen en zochten vervanging voor oudere toetsassistenten. Daartoe breidden we de toetsassistentenpool uit met studenten lerarenopleiding (UCCL, Arteveldehogeschool, KUL), met CTO-medewerkers, met pedagogisch begeleiders en met vrijwillig medewerkers.

De scholen waar geen toetsassistent kon langsgaan, kregen een 'digitale' toetsassistent toegewezen: de scholen streamden de afname en de monitoring gebeurde door een CTO-medewerker.

Toetsafnemers en toetsassistenen konden elke dag terecht op een helpdesk bemand door een CTO-mederwerker: telefoonpermanentie tussen 8-12u, voor dringende vragen en assistentie.

2.3 Opleiding van toetsafnemers en toetsassistenten

De face-to-face opleiding van toetsassistenten en toetsafnemers tijdens een opleidingsdag, was door coronarestricties op samenkomsten onmogelijk.

OPLOSSING

We organiseerden vijf digitale sessies 's avonds.

2.4 Verdeling toetspakketten

De verdeling van het materiaal vergde meer organisatie en een hogere kostprijs omdat het materiaal niet kon worden opgehaald en teruggebracht door toetsassistenten en meer materiaal moest verzonden worden (omdat de toetsassistent geen ongebruikt materiaal opnieuw kon meebrengen zodat het opnieuw kon worden ingezet).

OPLOSSING

We hebben het materiaal naar scholen verzonden met Bpost, met daarbij een (voor de scholen gratis) retourticket.

We signaleerden bijkomende kosten naar stuurgroep: we kunnen de bijkomende kosten opvangen door de verminderde vergoeding van toetsassistenten.

2.5 Uitval van scholen, toetsafnemers en kleuters

Er was een hoge uitval tijdens de data-verzameling:

- één school ging in quarantaine en één school ging gedeeltelijk in quarantaine en haakten dus af;
- drie toetsassistenten werden ziek (mogelijk door corona) en twee toetsassistenten haakten af (omdat ze moesten wachten op resultaten coronatest);
- vier scholen haakten af tijdens de dataverzameling: organisatorische druk, te weinig personeel…;
- tien scholen testten minder dan de helft van het aantal voorziene kleuters: organisatorische druk, te weinig personeel…;
- zes scholen moesten nog een deeltje van de testen afnemen de eerste week van januari: organisatorische druk, te weinig personeel…;
- meer ouders dan voorzien gaven geen toestemming: scholen geven aan dat ze ouders moeilijker konden bereiken (bv. geen schoolpoortcontacten of contact in de klas 's ochtends door corona).

OPLOSSING (steeds met afstemming op oorspronkelijke steekproef)
Reservescholen werden preventief ingeschakeld om uitval van kleuters op te vangen. Enkele scholen die zich spontaan hadden aangemeld, werden toegevoegd. Enkele studenten namen als toetsassistent de screening ook af in hun stageschool.

3 ▪ Instrument voor het Kalibratie-onderzoek

Het instrument dat we gebruikten voor het kalibratie-onderzoek bestond uit 30 taken. Dit zijn de taken we ook testten in het pilootonderzoek. Dit instrument werd in het vorig tussentijds rapport uitvoerig geschreven. In Hoofdstuk 1: Toetsconstruct3 ▪ Typetaken in KOALA' geven we een algemene beschrijving met voorbeelden van verschillende typetaken.

4 ▪ Afnameprocedure

4.1 Overlappende clusters

Net als in het pilootonderzoek testten we in het kalibratie-onderzoek in totaal 30 taken. Hiervan werden 18 taken niet enkel op papier, maar ook op tablet afgenomen. De 48 te piloten taken werden over verschillende 'clusters' (takenpakketten) verdeeld. De samenstelling van een cluster beantwoordde aan de volgende criteria: :

- Elke cluster bestaat uit 6 taken.
- Elke taak komt (minstens) in 2 verschillende clusters voor, om voldoende linken te kunnen leggen binnen de dataset en om te vermijden dat de resultaten van een bepaalde taak afhankelijk zijn van de plaats van de taak of van het takenpakket in de cluster).
- De verschillende clusters bevatten een gelijk aantal van de drie typentaken. Elke cluster bestond uit 3 meerkeuze-opdrachten (met maximum 1 verhaal), minstens 1 doe-en zoekopdracht en minstens 1 doe-opdracht met gestandaardiseerde observatie.
- Elke cluster heeft een vergelijkbare moeilijkheidsgraad (op basis van de analyses in het pilootonderzoek).
- Elke cluster toetst een variatie van doelen (zie 2.2 TOETSMATRIJS en, zie ook 2.3 VARIATIE IN COMPLEXITEIT).
- Digitale en papieren clusters zijn op dezelfde manier samengesteld, in functie van vergelijkbaarheid.

Op basis van bovenstaande criteria hebben we 10 verschillende clusters samengesteld (30 taken x 2 (iedere taak in 2 clusters) / 6 taken per cluster = 10). Elke cluster heeft een nummer. Omwille van

de herkenbaarheid hebben we aan elke cluster ook een dierennaam gegeven (bv. panda, kangoeroe, otter…).

NR.	Titel	Typetaak	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
24	Turnles	Doe-opdracht	✓				✓					
6	Hoepel	Doe-opdracht		✓				✓				
21	Rommel in de klas	Doe-opdracht			✓				✓			
28	Vingerpop	Doe-opdracht				✓				✓		
3	Dieren nadoen	Doe-opdracht					✓				✓	
25	Vandaag	Doe-opdracht	✓					✓				
1	Bewegen	Doe-opdracht							✓			✓
5	Eten	Doe-opdracht		✓						✓		
23	Spelen	Doe-opdracht			✓						✓	
7	Jelle	Zoek-opdracht				✓						✓
22	Speeltijd	Zoek-opdracht	✓					✓				
27	Verjaardagsfeest	Zoek-opdracht		✓					✓			
20	Rommel in de eetzaal	Zoek-opdracht			✓					✓		
29	Waar is	Zoek-opdracht				✓					✓	
30	Zandtafel	Zoek-opdracht					✓					✓
15	Mona's hoeken	Kies-opdracht	✓			✓						
10	Klasafspraken	Kies-opdracht		✓			✓					
17	Myriam	Kies-opdracht			✓			✓				
19	Park	Kies-opdracht				✓			✓			
2	Boekenhoek	Kies-opdracht					✓			✓		
8	Juf is jarig	Kies-opdracht						✓			✓	
9	Kabouters	kies-opdracht (verhaal)	✓	✓								
12	Konijntjes	kies-opdracht (verhaal)			✓	✓						
16	Mug en olifant	kies-opdracht (verhaal)					✓	✓				
18	Naar bad	kies-opdracht (verhaal)							✓	✓		
26	Varken en rups	kies-opdracht (verhaal)									✓	✓
4	Eenzaam	Kies-opdracht							✓			✓
11	Klastaakjes	Kies-opdracht	✓							✓		
13	Lievelingsboeken	Kies-opdracht		✓							✓	
14	Lievelingsspeelgoed	Kies-opdracht			✓							✓

Tabel 15: Clustersamenstelling voor het kalibratieonderzoek

Tot slot werd de volgorde van de taken op voorhand vastgelegd. Binnen elke cluster hanteerden we dezelfde volgorde:

1. doe-en zoekopdracht(en)

2. oefenitem voor de meerkeuze-opdrachten
3. meerkeuze-opdrachten
4. gestandaardiseerde observatie(s).

Deze volgorde is de meeste logische omdat kleuters zeer gemotiveerd zijn voor doe-en zoekopdrachten (zie HOOFDSTUK 4: PILOOTONDERZOEK) en deze typetaak dus een goede start vormt. De gestandaardiseerde observaties komen dan weer aan het einde omdat die individueel worden afgenomen. De kleuters die al klaar zijn met de afname kunnen dan gaan spelen of eventueel al terug naar de klas.

4.2 Beoogd aantal prestaties

Voor de afnames op papier streefden we naar minstens 200 observaties van elke taak (we wilden met andere woorden elke taak bij minstens 200 kleuters afnemen). Om hieraan te komen, maakten we de volgende berekening:

- Papieren clusters werden steeds afgenomen bij groepjes van 5 kleuters. Dit wil zeggen dat de afname van 1 papieren cluster neerkomt op 5 observaties voor alle taken in die cluster.
- Om aan 200 observaties voor alle taken te geraken moeten alle 10 clusters dus (minstens) 20 keer afgenomen worden.
- 1000 kleuters zijn nodig voor 20 afnames van 10 clusters.
- Binnen een school worden verschillende clusters getest, om zo variatie over scholen heen te bewaken. Enkel wanneer er voor een school meer dan 10 clusters moesten worden ingepland, weken we hiervan af.

Vooruitlopend op de concrete praktijk van datacollectie (waarbij soms kleuters niet aanwezig zijn, scholen tijdelijk moeten sluiten door corona…) werden in totaal 410 papieren clusters ingepland en naar de scholen verstuurd.

Voor de digitale afnames via tablets streefden we naar een minimum van 100 observaties van elke taak. Om deze aantallen te bereiken, maakten we de volgende berekening:

- Clusters op tablet werden steeds afgenomen bij groepjes van 4 kleuters. Dit wil zeggen dat de afname van 1 cluster op tablet neerkomt op 4 observaties voor alle taken in die cluster.
- Om aan 100 observaties voor alle taken op tablet te geraken moeten alle 10 clusters dus (minstens) 13 keer afgenomen worden.
- 520 kleuters zijn nodig voor 13 afnames van 10 clusters.

Ook voor de afnames op tablet probeerden we meer scholen en kleuters in te plannen dan strikt noodzakelijk. In totaal werden 217 digitale clusters ingepland en naar de scholen verstuurd, en bestond de geplande steekproef uit 766 ingeplande kleuters die de taken op tablet aflegden.

Op deze manier bereikten we in het kalibratie-onderzoek 1876 kleuters in totaal (zie 1 ▪ STEEKPROEF KALIBRATIEONDERZOEK).

4.3 Codering van de kleuters

Elke deelnemende kleuter uit een groepje kreeg een uniek ID-nummer, samengesteld uit schoolcode (3 letters), clusternummer (C01, C02, C03…) en kleuternummer (K01, K02, K03, K04, K05).

Daarnaast bezorgden alle deelnemende scholen klaslijsten met de namen en stamboeknummers van de deelnemende kleuters. Deze gegevens konden we nadien koppelen aan een databestand verkregen via het Departement Onderwijs en Vorming met de achtergrondgegevens van de kleuters. De uiteindelijke dataset werd geanonimiseerd zodat we met de (anonieme) ID-nummers van de kleuters de toetsscores en de achtergrondgegevens konden linken.

Op de klaslijsten duidden de leraren ook hun inschatting van de kleuters aan: enerzijds maakten ze een rangschikking (van de meest taalvaardige tot de minst taalvaardige kleuter) en anderzijds gaven de kleuters een kleur (waarbij ze aangaven hoe ze de prestatie op de taalscreening van die kleuter inschatten).

4.4 Afname

De afname van de screening op papier gebeurde door de leraren zelf (voor details in verband met de afnamecondities in coronatijden, zie 2 ▪ DATAVERZAMELING IN CORONATIJDEN). Hierbij kregen de leraren steeds assistentie, hetzij ter plaatse door opgeleide toetsassistenten (projectmedewerkers, (job)studenten of vrijwilligers), hetzij in de vorm van digitale assistentie door een projectmedewerker (via *Skype For Business*). Bovendien was er steeds een digitale helpdesk voorzien waar de toetsafnemers terecht konden in geval van twijfels, problemen of vragen. De toetsafname met tablets gebeurde eveneens door de leraar zelf, onder toezicht van een (fysiek aanwezige) toetsassistent.

4.5 Standaardisatie via toetsassistenten

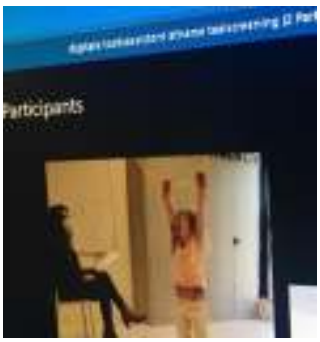
Toetsassistenten waren digitaal of fysiek aanwezig om de afnames op te volgen en standaardisatie te verzekeren. De rol van de toetsassistenten bestond enerzijds uit het bewaken van standaardisatie en anderzijds ook uit het praktisch ondersteunen. Dat laatste kon bijvoorbeeld betekenen dat de toetsassistent een kleuter hielp bij het omdraaien van een blad, naast een faalangstige kleuter plaatsnam om hem aan te moedigen, meevolgde in de instructiebundel om de toetsafnemer te herinneren aan het aantal toegestane herhalingen van het verhaaltje… Bij de digitale afname hielp de toetsassistent ook om de IPads klaar te maken voor de afname.

De digitale toetsassistent volgde via *Skype For Business* meerdere afnames tegelijkertijd. De toetsafnemers konden ten allen tijde vragen stellen via de chat of via een call. De digitale toetsassistent gaf vóór de afname instructies of antwoordde op vragen in verband met de afname, gaf tussentijds feedback en voerde na de afname een kort feedbackgesprek. De digitale toetsassistent greep in wanneer de standaardisatie in het gedrang kwam.

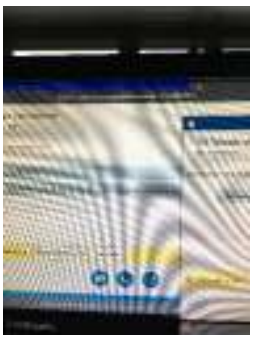
Voorbeelden digitale toetsassistentie met *Skype for business*

systeem klaarzetten om meerdere afnames te volgen

volgen van de afname



vragen beantwoorden via de chatfunctie



Alle toetsassistenten werden op voorhand opgeleid. Ook de toetsafnemers kregen de kans om hierbij aan te sluiten. Er waren meerdere tools voorzien:

- één online infosessie voorzien voor papieren afnames en één sessie voor digitale afnames. Tijdens deze sessies gaven projectmedewerkers online met ondersteuning van een PowerPointpresentatie uitleg over het verloop van de afname en kregen de deelnemers bovendien de kans om vragen te stellen. De sessie werd opgenomen en ter beschikking gesteld van de deelnemers.
- Informatieve PowerPoint: Iedereen die zich voor een online infosessie had ingeschreven, kreeg een informatieve PowerPointpresentatie toegestuurd. Ook degenen die niet aan de infosessies konden deelnemen of via mail aangaven meer info te willen, kregen de slides opgestuurd.
- Illustratieve videofragmenten: In de slides die getoond werden tijdens infosessie stonden links naar drie filmpjes, één voor elke typetaak. Het gaat hierbij niet om bewerkte filmpjes, maar om een opname van een afname van de screening bij één kleuter. Deze filmpjes werden niet getoond tijdens de sessie, maar ze konden op een later moment bekeken worden.

4.6 Documenten per cluster

Voor elke cluster werd een bundel voorzien met instructies, begeleidende prenten en registratieformulieren voor de afnemer en toetsdocumenten voor de vijf kleuters die dezelfde cluster afleggen.

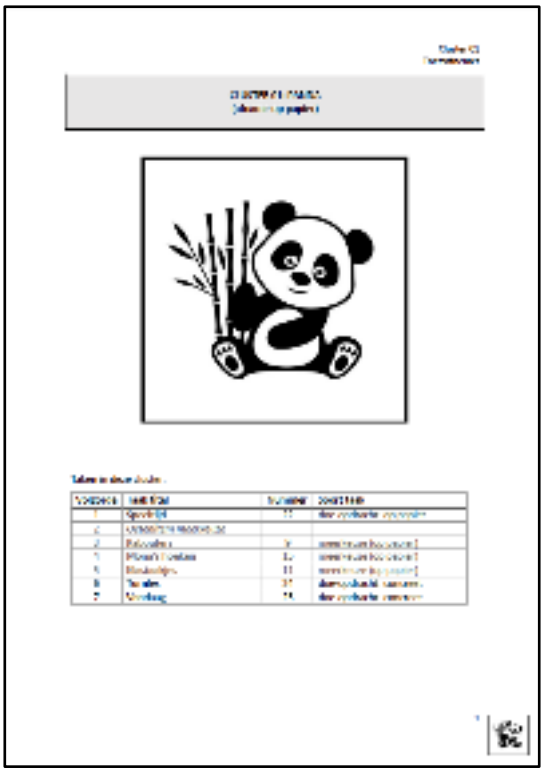
4.6.1 Instructiebundels voor de toetsafnemers

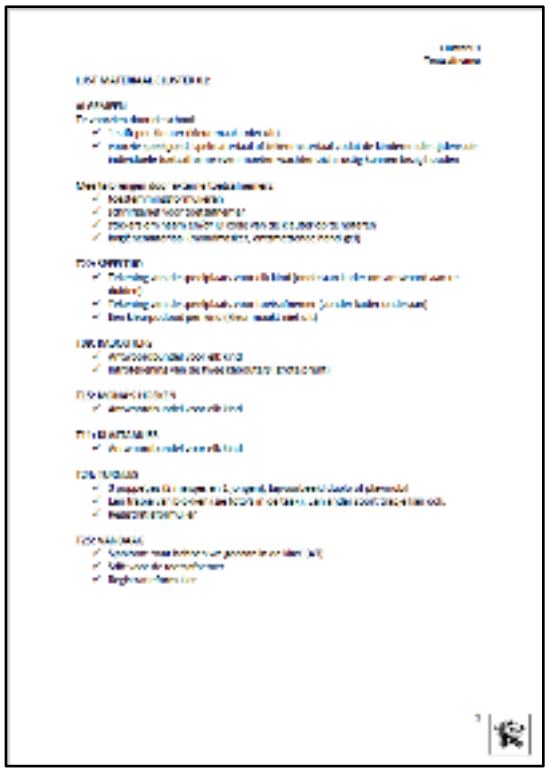
Voor elke takencluster was er een instructiebundel voor de toetsafnemer. Deze bevat gedetailleerde instructies over hoe de toetsafnemer de taken hoorde af te nemen (zie voorbeelden hieronder). Ook bij digitale afnames was een instructiebundel op papier voorzien.

De instructiebundel bestond uit:

- Een voorblad met algemene informatie over de cluster en een overzicht van alle taken uit de cluster.
- Een lijst met alle benodigde materialen per taak.
- De instructies per taak, in de volgorde van de afname.
- Per taak een duidelijke inleiding en contextschepping voor de toetsafnemer. Bv. taak Vingerpop: *Dit is een concrete doe-opdracht. Bij deze taak knutselen de kleuters stapsgewijs een papieren vingerpop. De kleuters werken individueel. Voorzie speelgoed of spelmateriaal zodat de kinderen die tijdens de individuele toetsafname even moeten wachten zich rustig kunnen bezighouden.*
- Miniaturen van de illustraties, i.e. afbeeldingen om te tonen aan de kleuters tijdens een bepaald onderdeel van de taak (zie hieronder: kleuterdocumenten). Op deze manier werd het duidelijk voor toetsafnemers welke afbeeldingen de kleuters op welke momenten te zien krijgen.
- De lay-out van instructiebundels werd na het pilootonderzoek over de taken heen gelijkvormig gemaakt. We voegden pictogrammen toe om de instructies op een heldere manier te presenteren, bijvoorbeeld pictogrammen voor instructies, vragen of voorlezen van een verhaal.

Voorbeeld instructiebundel: voorblad en materiaallijst





Voorbeeld instructiebundel: instructies bij meerkeuzetaak (kabouters)

4.6.2 Begeleidende prenten voor de toetsafnemer

Naast de instructiebundels zaten er bij iedere takencluster ook begeleidende prenten voor de toetsafnemer. Het gaat hierbij bijvoorbeeld om grote prenten van hoofdpersonages die de toetsafnemer aan de kleuters toont tijdens het voorlezen van een verhaal, of om de prenten van de taken Eten, Dieren nadoen of Speelhoeken die de toetsafnemer aan de kleuters toont tijdens gestandaardiseerde observaties.

4.6.3 Registratie van de antwoorden

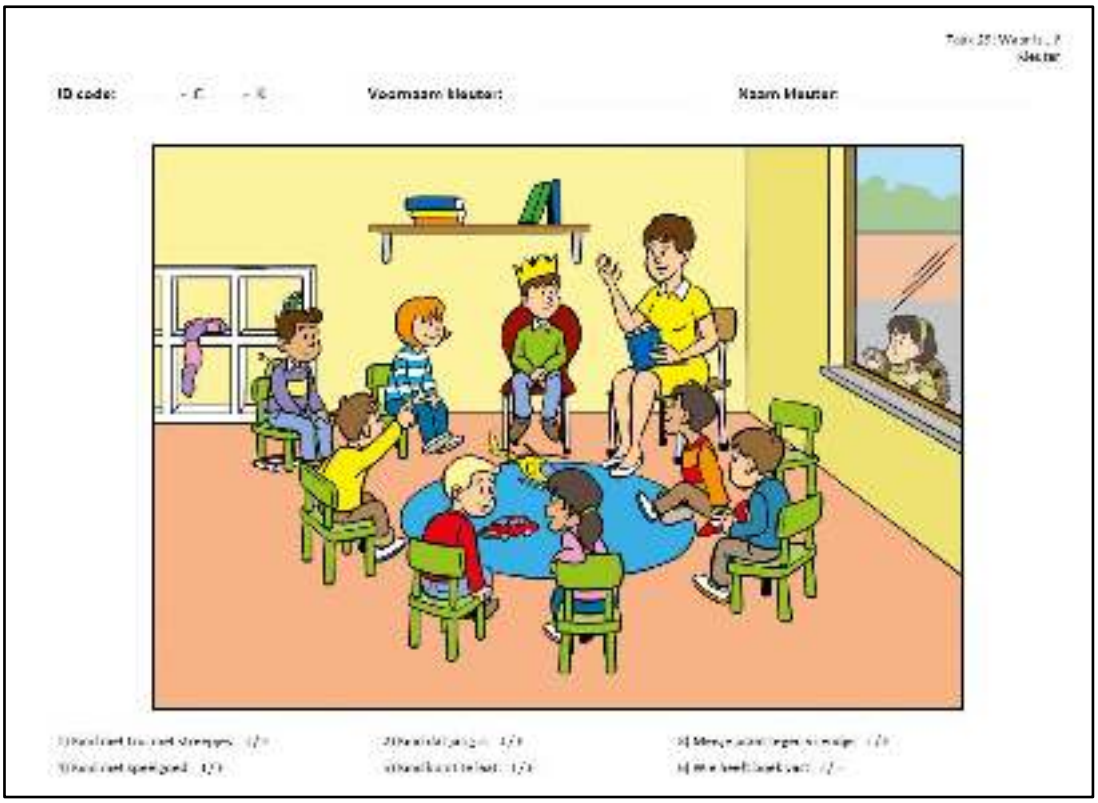
Voor papieren afnames waren er (papieren) registratieformulieren bijgevoegd aan de clusters. Op deze formulieren noteerden toetsafnemers de beoordelingen van kleuters bij gestandaardiseerde observaties. Ook was op deze formulieren ruimte voorzien voor opmerkingen. De registratieformulieren van de individuele taken werden gebundeld per cluster. Bij digitale afnames gebeurde de registratie van de gestandaardiseerde observaties digitaal via de tablet.

De registratie van de antwoorden was afhankelijk van de typetaken.

4.6.4 Registratie van doe-en zoekopdrachten

- Papieren afname: Doe-en zoekopdrachten op papier werden onmiddellijk beoordeeld (juist/fout) per item in een kadertje onderaan de prent.
- Afname op tablet: Bij doe-en zoekopdrachten die op de tablet werden afgenomen, gebeurde de registratie automatisch via Qualtrics. Enkele doe-en zoekopdrachten werden ook bij digitale afnames op papier afgenomen. In dat geval voerde de toetsafnemer de registratie onmiddellijk digitaal in via de tablet.

Voorbeeld: registratie van doe-en zoekopdrachten op papier (bij papieren afname); Taak Zandtafel



Voorbeeld: registratie van doe-en zoekopdrachten op papier (bij digitale afname, ingevuld);

Taak Speeltijd

Voorbeeld: automatische registratie van gedigitaliseerde doe-en zoekopdrachten op de tablet;

Taak Zandtafel (ingevuld: groene zone geeft aan waar kleuter op heeft getikt)

4.6.5 Registratie van meerkeuze-opdrachten

De registratie van meerkeuze-opdrachten was verschillend voor papieren afname en afname op tablet.

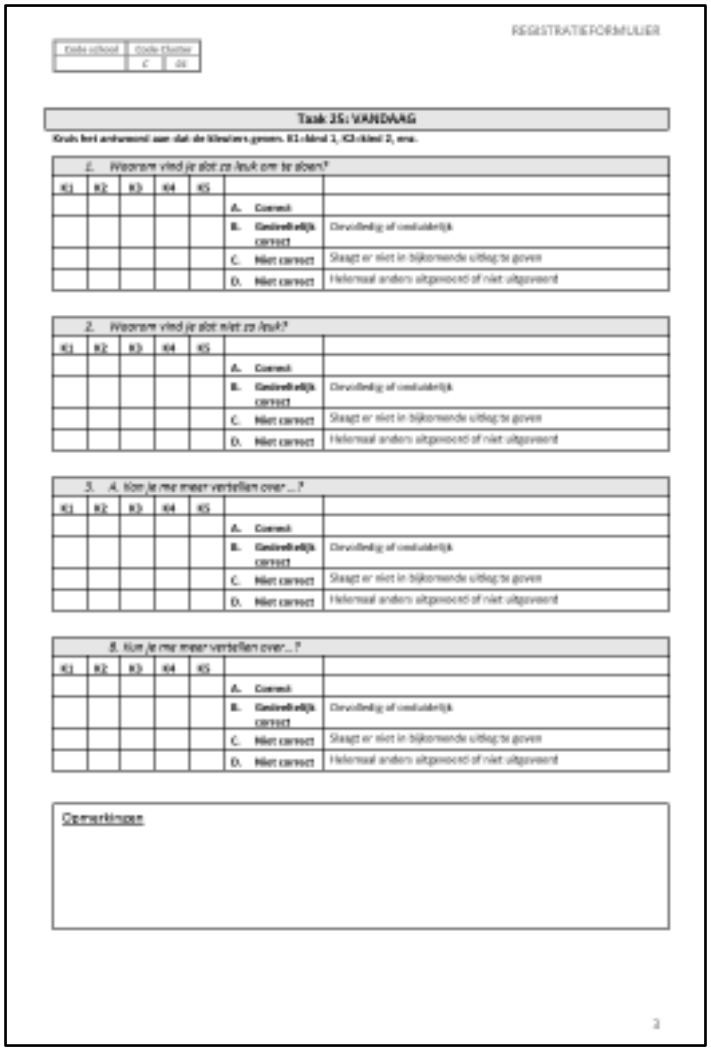
- Papieren afname: De antwoorden op de meerkeuze-opdrachten werden niet tijdens de afname geregistreerd maar op een later moment digitaal ingevoerd door jobstudenten.
- Afname op tablet: De antwoorden op de meerkeuze-opdrachten van de kleuters werden automatisch digitaal opgeslagen via Qualtrics.

4.6.6 Registratie van de gestandaardiseerde observaties

De gestandaardiseerde observaties werden voor het kalibratieonderzoek uitgebreid met uitvoeriger omschrijvingen van de beoordelingsopties (A, B, C, D) en met extra voorbeelden bij opties B (gedeeltelijk correct) en C (kind doet iets helemaal anders).

Voorbeeld: registratieformulier (voor gestandaardiseerde observaties bij papieren afname);

Taak Vandaag



REGISTRATIEFORMULIER

Code school	Code Cluster	
	C	01

Taak 15: VANDAAG

Kruis het antwoord aan dat de kleuters geven. K1=kind 1, K2=kind 2, enz.

1. Waarom vind je dat leuk om te doen?

K1	K2	K3	K4	K5		
					A. Correct	
					B. Gedeeltelijk correct	Onvolledig of onduidelijk
					C. Niet correct	Slaagt er niet in bijkomende uitleg te geven
					D. Niet correct	Helemaal anders uitgevoerd of niet uitgevoerd

2. Waarom vind je dat niet zo leuk?

K1	K2	K3	K4	K5		
					A. Correct	
					B. Gedeeltelijk correct	Onvolledig of onduidelijk
					C. Niet correct	Slaagt er niet in bijkomende uitleg te geven
					D. Niet correct	Helemaal anders uitgevoerd of niet uitgevoerd

3. 4. Kan je me meer vertellen over ...?

K1	K2	K3	K4	K5		
					A. Correct	
					B. Gedeeltelijk correct	Onvolledig of onduidelijk
					C. Niet correct	Slaagt er niet in bijkomende uitleg te geven
					D. Niet correct	Helemaal anders uitgevoerd of niet uitgevoerd

5. Kan je me meer vertellen over ...?

K1	K2	K3	K4	K5		
					A. Correct	
					B. Gedeeltelijk correct	Onvolledig of onduidelijk
					C. Niet correct	Slaagt er niet in bijkomende uitleg te geven
					D. Niet correct	Helemaal anders uitgevoerd of niet uitgevoerd

Opmerkingen

Voorbeeld: registratieformulier (voor gestandaardiseerde observaties bij digitale afname, ingevuld); Taak Bewegen

4.6.7 Kleuterdocumenten

In elke cluster waren ook documenten voor de kleuter opgenomen:

- Bij papieren afnames waren er voor alle taken 5 exemplaren voor de kleuters. De doe- en zoekopdrachten staan op 1 enkel blad. Meerkeuze-opdrachten zitten in een bundel met steeds één pagina per item, samengevoegd met een nietje. Na ieder item bladeren de kleuters met hulp van de toetsafnemers of toetsassistenten naar de volgende pagina. Op alle documenten was ruimte voorzien om de naam en het ID-nummer van de kleuter te noteren.
- Bij de digitale afnames waren er voor de kleuters enkel papieren invulbladen voor de doe- en zoekopdrachten waarvan er geen digitale versie bestond (taak Rommel in de eetzaal, taak Speeltijd, taak Verjaardagsfeest).

4.7 Papieren afname versus digitale afname

Papieren en digitale afname verliepen grotendeels op dezelfde manier. Toch waren er enkele verschillen voor toetsafnemers en kleuters.

- verschillen voor toetsafnemers:
    - andere manier van registreren (zie hierboven);
    - bij digitale afnames was er altijd een toetsassistent fysiek aanwezig.

- verschillen voor kleuters:
    - andere respons: een kring of kruisje tekenen bij papieren afnames vs. op de afbeelding tikken bij digitale afnames;
    - grootte van de afbeeldingen bij meerkeuze-opdrachten: op de tablet zijn de afbeeldingen kleiner.

Voorbeeld: digitale meerkeuzevraag: taak Lievelingsspeelgoed

HOOFDSTUK 6: RESULTATEN KALIBRATIEONDERZOEK

1 ▪ Kwantitatief: Statistische Analyses

De resultaten van de dataverzameling werden geanalyseerd via een multidimensionele Raschanalyse. Raschanalyse laat toe een objectieve moeilijkheidsgraad van de toetsitems te bepalen, onafhankelijk van de vaardigheid van de groep getoetste leerlingen. Hiervoor worden de toetsitems volgens een latente schaal gekalibreerd, die zowel de moeilijkheidsgraad van het toetsitem weergeeft als de vaardigheid van de leerling. Deze schaal is implicationeel.

Naast de analyse van de betrouwbaarheid van de screening gingen we ook een via biasonderzoek na of bepaalde toetsitems bepaalde kleuters (bv. van een bepaald geslacht of van een bepaalde achtergrond) bevoordelen of benadelen. Op basis van de analyses kan een selectie van toetstaken voor de cesuurbepaling gebeuren en voor het definitief instrument (zie Hoofdstuk 7: Cesuren bij koala en Hoofdstuk 9: Samenstelling Definitieve instrument)

De analyses werden uitgevoerd met behulp van de softwareprogramma's WINSTEPS en SPSS.

1.1 Samenstelling van het databestand

We maakten een databestand op basis van drie bestanden:

- klaslijsten
- bestand met achtergrondinfo over kleuters van het Departement Onderwijs en Vorming
- binnengekomen resultaten (papier + digitaal)

Door toekenning van unieke codes bleven de gegevens in het samengevoegd bestand anoniem.

De grootte van de 'sample' verschilt naargelang de criteria:

- aantal kleuters op klaslijsten: N=1983
- aantal kleuters in databestand met achtergrondgegevens van Departement Onderwijs en Vorming én op klaslijsten: N=1920
- aantal kleuters met zekerheid geboren in 2015: N=1949
- aantal kleuters met binnengekomen toetsobservaties: N=1955
    - afname op papier: N=1372 (70,2%)
    - afname met tablet: N=583 (29,8%)
- aantal kleuters met zekerheid geboren in 2015 én met binnengekomen toetsobservaties: N=1924
- aantal kleuters die wel op klaslijsten staan maar geen toetsobservaties hebben: N=28

We nemen alle kleuters met binnengekomen toetsobservaties (N=1955) mee in onderstaande analyses, tenzij anders vermeld.

1.2 Dataset van betrouwbare toetsitems

Er zijn 30 taken, bestsaande uit 148 items. In de analyses werden 5 toetsitems geïdentificeerd als onbetrouwbare items, met hoge outfit-waarden (outfit MNSQ > 2). Deze 5 items werden niet meegenomen in de analyses. Het gaat om:

- Item 1.4 (bewegen): 'zet de knuffel op je hoofd'
- Item 1.7 (bewegen): 'tik met je vinger op je neus'
- Item 4.2 (eenzaam): 'waarom moet Ana-Lucia huilen/wenen? Waarom is Ana-Lucia verdrietig? Trek een kring rond de juiste tekening'.
- Item 12.2 (konijntjes): 'Op welk tekening speelt Paultje met al zijn vriendjes? Waar zie je Paultje met al zijn vriendjes? Trek een kring rond de juiste tekening'.
- Item 18.4 (naar bad): 'Wat doet de mama van Rosanne helemaal aan het einde van het verhaaltje? Trek een kring rond de juiste tekening.

Onderstaande analyses vertrekken van de 143 resterende items.

1.3 Betrouwbaarheid van het screeningsinstrument als geheel

Tabel 16 geeft een overzicht van de betrouwbaarheid van de volledige screening met uitzondering van vijf items (zie 1.2 Dataset van betrouwbare toetsitems). Er zijn 143 betrouwbare items. Geen enkel item is gedefinieerd als 'extreem', dit wil zeggen dat geen enkel item door alle kleuters juist of fout wordt beantwoord.

SUMMARY OF 143 MEASURED (NON-EXTREME) ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	277.7	381.0	.00	.15	.99	-.03	1.00	.02
SEM	5.3	3.9	.09	.00	.01	.15	.02	.15
P.SD	63.3	46.5	1.04	.04	.12	1.85	.25	1.80
S.SD	63.6	46.7	1.04	.04	.12	1.85	.26	1.81
MAX.	398.0	437.0	2.42	.45	1.29	5.42	1.87	6.09
MIN.	85.0	91.0	-3.19	.11	.73	-5.34	.50	-4.81

REAL RMSE	.16	TRUE SD	1.02	SEPARATION	6.52	ITEM RELIABILITY .98
MODEL RMSE	.15	TRUE SD	1.02	SEPARATION	6.63	ITEM RELIABILITY .98
S.E. OF ITEM MEAN = .09						

Tabel 16: Analyse van de betrouwbaarheid van alle items

Gemiddeld wordt een item door 381 kleuters beantwoord, waarvan gemiddeld 278 keer correct. Dat betekent dat gemiddeld 72,8% van de kleuters een correct antwoord op een item geeft .

De 'item measure' geeft de 'moeilijkheidsscore' van de items weer. Hoe hoger de measure, hoe moeilijker het item. Het gemakkelijkste item heeft een measure van -3,19; het moeilijkste item heeft een measure van 2,42.

De betrouwbaarheid van de measures voor de items ('item reliability') is met .98 zeer goed. We zitten hiermee ruim boven de gewenste waarde van .80 (Linacre, M., 2012).

1.4 Betrouwbaarheid van de metingen op het niveau van de kleuters

In onderstaande tabel worden de kleuters met de maximumscore (N=69) en de minimumscore (N=2) niet meegenomen. Omdat in de antwoorden van deze kleuters geen variatie zichtbaar is, kunnen we voor deze kleuters geen betrouwbaarheidsgegevens (infit en outfit) berekenen.

SUMMARY OF 1884 MEASURED (NON-EXTREME) PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	20.1	27.9	1.38	.55	1.00	.07	1.00	.09
SEM	.1	.1	.03	.00	.00	.02	.01	.02
P.SD	5.8	3.0	1.26	.17	.17	.78	.60	.87
S.SD	5.8	3.0	1.26	.17	.17	.78	.60	.87
MAX.	46.0	55.0	3.98	1.07	2.59	4.96	9.90	5.16
MIN.	1.0	15.0	-3.84	.29	.43	-3.16	.17	-2.74

REAL RMSE	.59	TRUE SD	1.11	SEPARATION	1.88	PERSON RELIABILITY .78
MODEL RMSE	.57	TRUE SD	1.12	SEPARATION	1.94	PERSON RELIABILITY .79
S.E. OF PERSON MEAN = .03						

MAXIMUM EXTREME SCORE: 69 PERSON 3.5%
MINIMUM EXTREME SCORE: 2 PERSON .1%
LACKING RESPONSES: 28 PERSON

Tabel 17: Analyses van de betrouwbaarheid van de metingen op het niveau van de kleuter

Gemiddeld hebben de resterende kleuters (die geen extreme score behaalden) 28 items beantwoord of uitgevoerd (M=27.9, SD=3), waarvan gemiddeld 20 items correct (M=20.1, SD=5.8). Gemiddeld behalen de kleuters daarmee een score van 72% correct beantwoorde of uitgevoerde items.

De 'person measure' geeft de 'vaardigheidsscore' van de kleuters weer. Hoe hoger de measure, hoe beter de luistervaardigheid van een kleuter is. De minst luistervaardige kleuter heeft een measure van -3,84 en de meest luistervaardige kleuter heeft een measure van 3,98. Als we de kleuters die het minimum of maximum behaalden erbij nemen, zit de range van de measures tussen -4,93 en 5,12.

De betrouwbaarheid van de measures voor de kleuters ('person reliability') is met .78 redelijk goed. Idealiter zit deze betrouwbaarheidscore boven .80 (Linacre, M., 2012).

1.5 De verhouding tussen de moeilijkheidsgraad van de screening en de vaardigheid van de kleuters (Wright Map)

Figuur 14 biedt een visuele voorstelling van de verhouding tussen moeilijkheidsgraad van de items en de vaardigheidsscores van de kinderen (Wright Map). In de item-person map zien we links de measures van de kleuters (vaardigheidsscores), en rechts de measures van de items (moeilijkheidgraad).

Het gemiddelde van de ‘item measures’ is altijd 0. Als de waarde van de ‘person measure’ (vaardigheidsscore) gelijk is aan de waarde van de ‘item measure’ (moeilijkheidsscore), wil dat zeggen dat deze kleuter 50% kans heeft om dat item correct te beantwoorden. Is de measure van de kleuter groter dan die van het item, dan betekent dit dat de kleuter ‘vaardiger’ is en meer kans heeft om het item correct te beantwoorden. Is de measure van de kleuter lager dan die van het item, dan wil dat zeggen dat het item (te) moeilijk is voor deze kleuter en neemt de kans op een correct antwoord af.

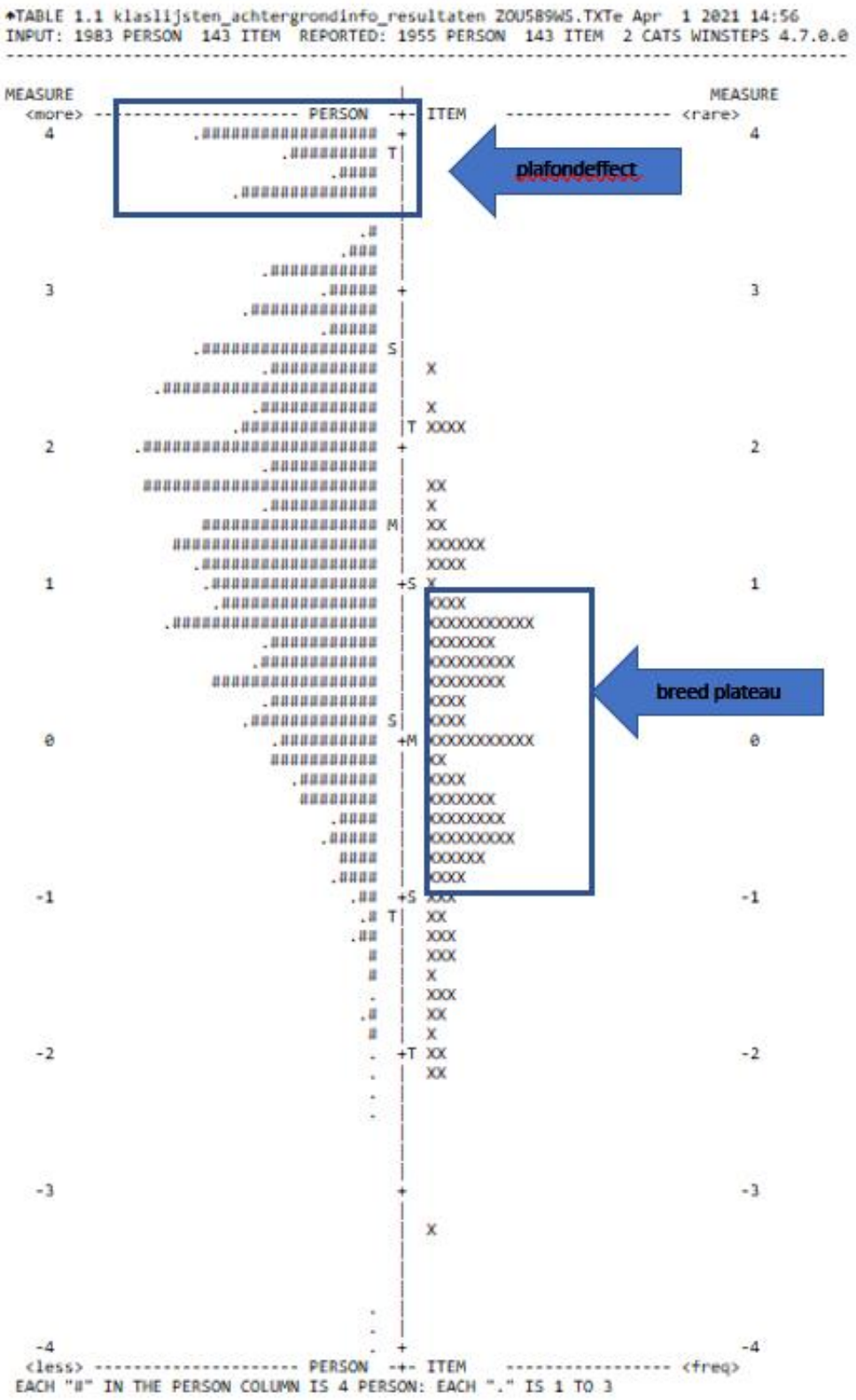


TABLE 1.1 klaslijsten_achtergrondinfo_resultaten ZOU589WS.TXTe Apr 1 2021 14:56
INPUT: 1983 PERSON 143 ITEM REPORTED: 1955 PERSON 143 ITEM 2 CATS WINSTEPS 4.7.0.0
MEASURE
<more>
PERSON
ITEM
<rare>
plafondeffect
breed plateau
<less>
<freq>
EACH "#" IN THE PERSON COLUMN IS 4 PERSON: EACH "." IS 1 TO 3

Figuur 14: Visuele voorstelling van de verhouding tussen de moeilijkheidsgraad van de toets en de vaardigheid van de kleuters (Wright Map)

We doen onderstaande vaststellingen op basis van de Wright Map:

- De measures van de items volgen een normaalverdeelde curve met een behoorlijk breed 'plateau'. Deze ruime spreiding van de items in de 'middenzone' (tussen measure -1 en +1) laat toe om twee cesuren te plaatsen. Het is immers van belang om voldoende items ter beschikking te hebben in de zone(s) waarin de cesuur vermoedelijk geplaatst zal worden.
- Bij de measures van de kleuters zien we tussen measure -2 en measure 3,5 een normaalverdeelde curve. Aan de bovenkant (measure > 3,5) zien we een groep kleuters die (bijna) het maximum scoren. Voor deze groep zijn er geen items meer die corresponderen met hun vaardigheidsmeasure. We zien een duidelijk plafondeffect. Dit is sluit aan bij de doelstellingen van de toets: de screening focust zich immers op de detectie van risicokleuters, niet op het identificeren van zeer sterk taalvaardige kleuters.
- Het gemiddelde (M) ligt voor de person measure (luistervaardigheid) een stuk hoger dan voor de item measure (moeilijkheidgraad). Dit wil zeggen dat het een eerder 'gemakkelijke' screening is. Zelfs kleuters met een (zeer) lage measure zullen toch een aantal items correct kunnen beantwoorden. Dit is belangrijk om de motivatie van de zwakkere presteerders op peil te houden doorheen de afname.
- Tussen measures 0 en 1 lijken er (ruim) voldoende items te zijn die corresponderen met de vaardigheid van de kleuters. Dit wijst erop dat we met dit screeningsinstrument de laagtaalvaardige kleuters kunnen onderscheiden van de kleuters die (net) voldoende taalvaardig zijn.

1.6 Moeilijkheidsgraad van de verschillende items

Uit de Wright Map konden we afleiden dat de gemiddelde vaardigheid van de kleuters doorgaans hoger ligt dan de gemiddelde moeilijkheidsgraad van de items. Om die reden bestuderen we de moeilijkste en gemakkelijkste items uit de toetsenbatterij in detail.

1.6.1 Moeilijkste items (measure > 2)

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ESTIM DISCR	ITEM
56	118	368	2.42	.13	1.24	4.08	1.48	3.93	.25	.45	65.5	73.7	.47	g12.1
73	135	352	2.19	.13	1.23	3.71	1.42	3.78	.37	.52	67.0	73.4	.51	g16.1
88	147	384	2.10	.12	1.18	3.28	1.49	4.05	.34	.47	66.9	72.5	.56	g19.3
21	139	345	2.08	.13	1.24	4.15	1.59	4.58	.32	.50	66.5	72.5	.39	g4.4
4	141	347	2.03	.13	.93	-1.22	.95	-.36	.52	.49	76.3	72.5	1.12	g1.5
128	168	411	2.02	.12	1.00	.08	.95	-.44	.50	.50	70.2	72.5	1.01	g27.5

Tabel 18: Moeilijkheidsgraad van de moeilijkste items

Zes items hebben een hoge measure (>2). Hiertussen zitten geen items met hoge outfit-waarde; deze items werden in de verkennende analyses opgespoord en weggelaten. De discriminerende waarde van de vier moeilijkste items is laag (< 1). Deze items zijn dus niet erg geschikt om de (zeer) sterk taalvaardige kleuters te onderscheiden van de anderen. Dit is niet problematisch, omdat dit ook niet het opzet is van dit instrument. Merk op dat er toch ook enkele items zijn die wel goed discrimineren, en dus goed in staat zijn om verschillen tussen sterk taalvaardige kleuters te registreren.

1.6.2 Gemakkelijkste items (measure < -2)

89	361	385	-2.03	.23	.93	-.34	1.66	1.56	.36	.35	94.5	94.2	1.02	@19.4
36	85	91	-2.04	.45	1.02	.11	1.10	.37	.29	.29	93.4	93.4	.99	@7.5
15	358	376	-2.21	.25	1.05	.29	.76	-.48	.28	.21	95.0	95.0	.99	@3.1
139	377	396	-2.24	.25	.93	-.31	.96	.04	.26	.23	95.0	95.0	1.03	@29.6
135	398	398	-3.19	.37	.92	-.13	.50	-.98	.21	.15	97.9	97.9	1.05	@29.2

Tabel 19: Moeilijkheidsgraad van de gemakkelijkste Items

Er zijn 5 items met een (heel) lage measure (< -2). Deze items zijn voldoende betrouwbaar (geen outfit) en hebben een goede discriminerende waarde. Ze kunnen dus helpen om het precieze niveau van zeer taalzwakke kleuters te bepalen.

1.6.3 Item polarity

We verwachten dat items met een lage measure (gemakkelijke items) correct beantwoord worden door kinderen met een hoge measure (hoge luistervaardigheidsscore). Om dit te verifiëren gaan we na of er gemakkelijke items zijn waarop sterk taalvaardige kleuters, tegen de verwachtingen in, een fout antwoord geven. We gaan daarvoor op zoek naar items met een lage PTMEASURE. Een negatieve waarde of lage waarde (< 0.20) vraagt om nader onderzoek.

Alle items hebben een voldoende hoge PTMEASURE-waarde (>= .20). Dat wijst erop dat voor geen enkel item sterk taalvaardige kleuters vaker foutief antwoorden dan laagtaalvaardige kleuters.

1.7 Relatie tussen moeilijkheidsgraad van de items en eigenschappen van de items

In dit stuk gaan we na of bepaalde eigenschappen of kenmerken van de items een invloed hebben op de measures (moeilijkheid) van de items.

1.7.1 Antwoordopties wel/niet op voorhand te zien

Dit is enkel van toepassing op items in meerkeuzevragen. In de meeste gevallen krijgen de kleuters de antwoordopties te zien terwijl de leraar de instructie, mededeling of verhaal voorleest. Bij enkele taken krijgen de kleuters de vier antwoordmogelijkheden pas daarna te zien. Het gaat om de taken Eenzaam (taak 4), Konijntjes (taak 12) en Naar bad (taak 18).

Wanneer de kleuters moeten luisteren naar wat de leraar vertelt, krijgen ze voor de items van deze drie taken een neutrale prent van het hoofdpersonage te zien i.p.v. de antwoordmogelijkheden.

Het valt op dat er in elk van deze taken een onbetrouwbaar item zat: 4.2, 12.2 en 18.4. Deze onbetrouwbare items werden niet meegenomen in de analyses ter voorbereiding op de cesuurbepaling noch in de verdere rapportering (zie 'sample' hierboven). Voorzichtigheid is dus geboden als we ervoor kiezen om bij een bepaalde taak de antwoordmogelijkheden niet meteen te tonen aan de kleuters.

Er blijven 11 betrouwbare items over (17% van de meerkeuze-items) waarbij de antwoordopties niet meteen worden getoond. De gemiddelde measure van deze set items is 0,65 (SD = 1,04). De gemiddelde measure van de items waarbij de kleuters wel meteen de antwoordopties te zien krijgen is 0,12 (N=54; SD=0,87). De gemiddelde measure van deze items ligt lager, maar een t-test wijst uit dat items waarbij de antwoordopties niet meteen worden getoond, niet significant moeilijker zijn.

1.7.2 Eén of twee items gekoppeld aan een vraag of stukje verhaal

Dit is opnieuw enkel van toepassing op meerkeuzevragen. Meestal volgt er na een instructie, mededeling, vraag of stukje verhaal één item; in enkele gevallen twee items. Dat betekent dat de kleuters twee verschillende stukken informatie moeten halen uit een deeltje informatie, instructie, vragen of een verhaal.

Heel concreet gaat het om 7 items uit 4 taken:

Item nummer	Taak	Vraag	Doelstelling
4.2	Eenzaam	Waarom moet Ana-Lucia huilen/wenen? Waarom is Ana-Lucia verdrietig? Tik op de juiste tekening.	Vragen begrijpen

4.4	Eenzaam	Ana-Lucia kan op het einde van het verhaaltje weer lachen. Hoe komt dat? Waarom kan Ana-Lucia terug lachen? Tik op de juiste tekening.	Vragen begrijpen
9.4	Kabouters	Wat heeft meneer konijn bij/mee? Tik op de juiste tekening.	Verhaal begrijpen
12.2	Konijntjes	Op welke tekening speelt Paultje met al zijn vriendjes? Waar zie je Paultje met al zijn vriendjes? Tik op de juiste tekening.	Verhaal begrijpen
12.6	Konijntjes	Op welke tekening zie je het einde van het verhaaltje? Wat doen de konijntjes op het einde van het verhaaltje? Tik op de tekening die hoort bij het einde van het verhaal.	Verhaal begrijpen
18.2	Naar bad	Wat doet Rosanne in bad? Tik op de juiste tekening.	Verhaal begrijpen
18.4	Naar bad	Wat doet de mama van Rosanne helemaal aan het einde van het verhaaltje? Tik op de juiste tekening.	Verhaal begrijpen

Tabel 20: Vergelijking antwoordopties tonen - niet tonen

Opnieuw vallen hier drie onbetrouwbare items op: 4.2, 12.2 en 18.4. Deze onbetrouwbare items werden niet meegenomen in de analyses ter voorbereiding op de cesuurbepaling en de verdere rapportering (zie 1.2 Dataset van betrouwbare toetsitems). Voorzichtigheid is geboden als we ervoor kiezen om aan één vraag of stukje verhaal twee items te koppelen.

Er zijn nog 4 betrouwbare items die als tweede item voorgelegd worden. We vergelijken de measure van deze items met items die meteen volgen op de vraag of stukje verhaal dat de leraar voorleest. We kijken dus enkel naar taak 4, taak 9, taak 12 en taak 18 (N=16).

De gemiddelde measure van de items die meteen volgen op de vraag of stukje verhaal is 0,05 (N=12; SD=0,92). De gemiddelde measure van de items die als tweede volgen is 1,33 (N=4; SD=0,57). Een t-test wijst uit dat het verschil significant is ($t=-2,58$; $df=14$; $p<0,05$). Als er een tweede item volgt op een vraag of stukje verhaal, is dit dus significant moeilijker voor de kleuters. Voorzichtigheid bij de interpretatie van deze resultaten is geboden omdat het maar om een beperkt aantal items gaat.

1.8 Vaardigheid van de kleuters (n=1955)

Uit de Wright Map konden we afleiden dat de gemiddelde vaardigheid van de kleuters doorgaans hoger ligt dan de gemiddelde moeilijkheidsgraad van de items. Om die reden bestuderen we de

kleuters met een hoge of lage vaardigheidsscore of kleuters met een (zeer) onbetrouwbaar resultaat in meer detail.

1.8.1 Kleuters met een (zeer) hoge taalvaardigheid

69 kleuters (3,5%) beantwoorden alle voorgelegde items correct en behalen daarmee de maximumscore. Daarnaast is er nog een vrij grote groep die een zeer hoge score behalen (>95% correcte antwoorden). Deze groep van zeer goede presteerders telt in totaal 191 kleuters (9,8%).

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< 95% correcte antwoorden	1764	89,0	90,2	90,2
	>= 95% correcte antwoorden	191	9,6	9,8	100,0
	Total	1955	98,6	100,0	
Missing	System	28	1,4		
Total		1983	100,0		

Tabel 21: Kleuters met minstens 95% correcte antwoorden

Het gaat om 191 kleuters (9,8%) met een measure tussen 3,38 en 5,20. De gemiddelde measure van deze groep is 4,14 (SD=0,59). Er zijn geen items die corresponderen met de measures van deze sterke presteerders (hoogste itemmeasure = 2,42). Dit is niet problematisch omdat we de screening vooral willen inzetten om laagtaalvaardige kleuters te onderscheiden van de rest.

We moeten er wel over waken dat de screening ook voor sterk taalvaardige kleuters voldoende boeiend en uitdagend blijft. Adaptief testen kan daartoe eventueel overwogen worden: bijvoorbeeld, als een kleuter x aantal items volledig correct doet, kan de toetsafname stoppen (en wordt de toetsafname dus korter).

1.8.2 Kleuters met een (zeer) lage taalvaardigheid

Er zijn twee kleuters die geen enkele voorgelegde vraag correct beantwoorden; zij behalen een score van 0%. Vijf kleuters behalen een score van 10% of minder. De groep zeer zwakke presteerders is dus zeer klein. 1,2% van de kleuters behaalt een totaalscore van 20% of minder.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	> 20% correcte antwoorden	1932	97,4	98,8	98,8
	<= 20% correcte antwoorden	23	1,2	1,2	100,0
	Total	1955	98,6	100,0	
Missing	System	28	1,4		
Total		1983	100,0		

Tabel 22: Kleuters met maximum 20% correcte antwoorden

De groep zeer zwakke presteerders bestaat uit 23 kleuters met een measure tussen -4,93 en -1,48. De gemiddelde measure van deze groep is -2,31 (SD=0,97). Er zijn vijf betrouwbare items met een measure -2 of lager met goede discriminerende waarde. De screening bevat dus een aantal items die qua moeilijkheidsgraad corresponderen met de measures van deze zeer zwakke presteerders.

Het is niet zo dat deze groep gemiddeld genomen minder items heeft beantwoord (en meer missings heeft). Het is ook niet zo dat ze minder kansen kregen om correct te antwoorden of een minder gevarieerde set aan items voorgelegd kregen. Deze groep zeer zwakke presteerders heeft gemiddeld 26,7 items beantwoord (SD=3,57). De overige kleuters hebben gemiddeld 27,8 items beantwoord (SD=3,06). Dit verschil is niet significant.

We moeten er wel over waken dat de screening voor zeer zwak taalvaardige kleuters niet frustrerend of demotiverend wordt. Adaptief testen kan daartoe eventueel overwogen worden: bijvoorbeeld, als een kleuters x aantal gemakkelijke items niet correct kan beantwoorden, kan de toetsafname stoppen (en wordt de toetsafname dus korter).

1.8.3 Kleuters met zeer onbetrouwbare metingen (outfit MNSQ >=3[7])

De 'outfit' statistiek helpt ons om outliers op te sporen onder de kleuters. Het gaat om kleuters die niet consistent antwoorden. Ze gaan de mist in bij gemakkelijke items en antwoorden (toevallig) correct op moeilijke items. Om deze kleuters te identificeren kijken we naar de outfit MNSQ-waarden. Kleuters met een waarde >=3 beschouwen we als 'outliers'. Zij werden weggelaten uit de analyses voor de cesuurbepaling. Als we uitgaan van de set met 143 betrouwbare items, gaat het om 21 kleuters (1,1%).

[7] Een outfit MNSQ-waarde >= 3 is een minder strenge grenswaarde. We hanteren deze waarde om de 'zware' outliers eruit te halen.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	geen outlier	1934	97,5	98,9	98,9
	outlier	21	1,1	1,1	100,0
	Total	1955	98,6	100,0	
Missing	System	28	1,4		
Total		1983	100,0		

Tabel 23: Kleuters met zeer onbetrouwbare metingen (Outfit MNSQ >=3)

Welke kleuters zorgen voor deze zeer onbetrouwbare metingen? Het zijn iets vaker kleuters met een hogere measure: 11 kleuters hebben een measure >3 (t=-3,36; df=1953; p<,01). Het lijkt dus iets vaker te gaan om goede presteerders die inconsistent antwoorden omdat ze bij een aantal gemakkelijke items niet correct antwoorden.

Als we kijken naar andere kenmerken van deze kleuters, zien we dat ze vaker een hoogopgeleide moeder hebben (18 kleuters) en vaker Nederlands spreken thuis (15 kleuters). Enkel voor de opleiding van de moeder is het verschil significant (t=3,09; df=20,84; p<,01).

Er is geen samenhang met leeftijd (geboortemaand), leermoeilijkheden gesignaleerd door de leraar of afnametype (papier of digitaal).

1.8.4 Kleuters met onbetrouwbare metingen (outfit MNSQ >=2[8])

Vaak wordt bij de outfit MNSQ waarde 2 als grenswaarde genomen om onbetrouwbare cases te identificeren. Als we deze strengere grenswaarde hanteren, stellen we vast dat 69 kleuters (3,5%) beschouwd worden als kleuters met onbetrouwbare metingen (outliers).

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	geen outlier	1886	95,1	96,5	96,5
	outlier	69	3,5	3,5	100,0
	Total	1955	98,6	100,0	
Missing	System	28	1,4		
Total		1983	100,0		

Tabel 24: Aantal kleuters met een niet-betrouwbaar resultaat (outfit MNSQ>=2)

[8] Een outfit MNSQ-waarde >= 2 is een strenge grenswaarde. We hebben deze groep geïdentificeerd om beter zicht te krijgen op welke kleuters een eerder inconsistent antwoordenpatroon vertonen. Kleuters met een outfit MNSQ tussen 2 en 3 zijn geen extreme outliers en nemen we wel mee in de cesuurbepaling.

Wie zijn deze kleuters? Opnieuw zien we dat de kleuters in deze groep vaker een hogere measure hebben: 21 kleuters hebben een measure > 3. De gemiddelde measure van deze groep is 1,89 (SD=1,85) terwijl de gemiddelde measure van de betrouwbare kleuters 1,48 (SD=1,39) bedraagt. Het verschil tussen de twee groepen is echter niet significant.

Als we kijken naar de andere kenmerken van de kleuters in deze groep, zien we dat ze vaker een hoogopgeleide moeder hebben (t=3,77; df=75,68; p<,01) en vaker Nederlands spreken thuis (t=2,07; df=73,88; p<,05). Beide verschillen zijn significant.

Er is geen samenhang met leeftijd (geboortemaand), leermoeilijkheden gesignaleerd door de leraar en afnametype (papier of digitaal).

1.9 Itembias : Differential item functioning (dif-analysis)

Via DIF-analyses identificeren we items die verschillend 'werken' voor verschillende groepen. Door zulke items op te nemen in een toets worden onbedoeld bepaalde groepen benadeeld. Items die zulke 'bias' vertonen zullen we weren voor het definitieve screeningsinstrument.

Via DIF-analyses gaan we na of kleuters die tot een bepaalde groep behoren meer of minder kans hebben om een bepaald item correct te beantwoorden dan kleuters die tot een andere groep behoren. Als het contrast in de item measures tussen de twee groepen hoog is, wijst dit erop dat er verschillen zijn tussen de twee groepen. Een t-test gaat vervolgens na of het verschil significant is. Wij hanteren de volgende grenswaarden om bias te identificeren: DIF-contrastwaarde >.50 en een p-waarde <.01. In totaal zijn er 25 van de 143 items (17,5%) waarbij deze grenswaarden overschreden worden.

De DIF-analyses worden uitgevoerd voor afnametype, gezinstaal en opleiding van de moeder. Items die voldoen aan deze voorwaarden, worden in principe niet opgenomen in de selectie voor de cesuurbepaling en de finale versie van het screeningsinstrument.

1.9.1 Afnamemodaliteit

We gaan na of de modaliteit van de afname leidt tot verschillen in moeilijkheidsgraad van de items. Een afname op papier heeft code 0, een digitale afname code 1. De waarden voor het DIF-contrast moeten dan als volgt geïnterpreteerd worden:

- Positieve DIF-contrast: item measure in papieren versie is significant hoger dan item measure in digitale versie. De papieren versie is moeilijker.
- Negatieve DIF-contrast: item measure in digitale versie is significant hoger dan item measure in papieren versie. De digitale versie is moeilijker.

Item nummer	Taak	Vraag	DIF-contrast	Moeilijker voor?

8.2	Juf is jarig	Bij de volgende taart zegt de juf: 'Leg enkele schijfjes banaan op de taart. En leg daarna op elk schijfje banaan een kers.' Tik op de juiste tekening.	-1.13	Digitaal
10.2	Klasafspraken	Dit betekent dat er drie kinderen met de poppen mogen spelen. Als er drie kinderen met de poppen spelen, dan is de poppenhoek vol. Dan mogen er geen andere kindjes meer bij. Op welke tekening doen de kinderen het goed? Tik op de juiste tekening.	-0.81	Digitaal
13.3	Lievelingsboeken	Zita vindt verhalen over kabouters, draken of feeën niet leuk. Ze houdt het meest van verhalen over kinderen. Verhalen die écht gebeurd zijn. Met welk verhaal zou Zita blij zijn? Tik op het boek voor Zita.	-0.73	Digitaal
15.2	Mona's hoeken	De leeshoek in de klas van Mona is erg gezellig. Er staat een grote zetel waarin je kan zitten als je een boekje leest. Je kan ook op één van de grote kussens op de grond zitten. En er zijn natuurlijk veel boekjes. Ze staan allemaal in een boekenrek. Tik op de leeshoek in de klas van Mona.	0.75	Papier
15.4	Mona's hoeken	De beweeghoek is de lievelingshoek van Mona. In de hoek staan twee fietsjes. Op de grond ligt een dikke mat. Naast de mat staat een glijbaan. Daar wil Mona graag een keertje op. Tik op de beweeghoek in de klas van Mona.	-0.80	Digitaal
17.4	Myriam	De juf zegt: 's Morgens moet je je fruit in de gele bak op de tafel leggen. Je schriftje leg je op de kast.' Op welke tekening doet Myriam het goed? Tik op de tekening waar Myriam het goed doet.	-0.88	Digitaal
19.4	Park	Juf Lotte zegt: 'Tijdens de wandeling mag je niet eten maar wel drinken.' Op welke tekening doen alle kinderen het goed? Tik op de juiste tekening.	-1.54	Digitaal

26.1	Varken en Rups	Wat vonden Varken en Rups het allerleukste om samen te doen? Tik op de juiste tekening.	-0.99	Digitaal
26.2	Varken en Rups	Waar zou Rups gaan wonen? Tik op de juiste tekening.	0.72	Papier
26.3	Varken en Rups	Hoe voelt Rups zich als ze hoort dat ze moet verhuizen? Tik op de juiste tekening.	0.99	Papier
30.3	Zandtafel	Nikola heeft een flesje gevuld met water. Hij giet het uit en het water spettert op Anaïs. Het water is koud en daar moet Anaïs mee lachen. Waar is Anaïs? Tik op Anaïs.	-1.13	Digitaal

Tabel 25: Resultaten DIF-analyses papier-digitaal

Bij 11 items is er sprake van een significant verschil in moeilijkheid tussen de papieren versie en de digitale versie. Het gaat voornamelijk om items uit meerkeuzevragen (30.3 is een uitzondering). In de meeste gevallen is de digitale versie moeilijker dan de papieren. In bepaalde gevallen zou dit te maken kunnen hebben met kleine details die moeilijker zichtbaar zouden zijn op het iPadscherm.

Concreet denken we hierbij aan volgende items:

- 8.2: details op de taart (schijfjes banaan)
- 10.2: speelgoed dat kinderen vast hebben
- 17.4: voorwerpen die Myriam vast houdt
- 19.4: dingen die de kinderen in de hand houden (eten en drinken)

De verklaring voor de resultaten van deze DIF-analyse is echter niet eenduidig. In taak 15 (Mona's hoeken) bijvoorbeeld is er ook sprake van kleine details. Maar item 15.2 (leeshoek in Mona's hoeken) is volgens de output moeilijker in de papieren versie. Item 15.4 (beweeghoek in Mona's hoeken) zou dan weer moeilijker zijn in de digitale versie, maar bevat minder kleine details dan 15.2. De hypothese dat kleine details moeilijker zichtbaar zijn op een tabletscherm gaat in het geval van taak 15 dus niet op.

Daarnaast zijn er enkele items waarbij details op de tekening niet doorslaggevend zijn om een correct antwoord te kunnen geven. Concreet gaat het om item 13.3 (lievelingsboeken) en de items bij taak 26 (Varken en Rups). Bij taak 26 (Varken en Rups) stellen we bovendien vast dat één item (26.1) moeilijker is in de digitale versie, terwijl de twee andere items (26.2 en 26.3) dan weer moeilijker zijn in de papieren versie. Hier is niet meteen een verklaring voor te vinden.

1.9.2 Gezinstaal

We gaan na of de taal die thuis gesproken wordt leidt tot een verschillende inschaling van de moeilijkheidsgraad van de items. Een kleuter die thuis enkel Nederlands spreekt krijgt code 0, een kleuter die thuis minstens één andere taal spreekt krijgt code 1.

- Positieve DIF-contrast: item measure ligt hoger voor kinderen die thuis enkel Nederlands spreken. Het item is moeilijker voor Nederlandstalige kleuters.
- Negatieve DIF-contrast: item measure ligt hoger voor meertalige kinderen. Het item is moeilijker voor meertalige kleuters.

Item nummer	Taak	Vraag	DIF-contrast	Moeilijker voor?
1.2	Bewegen	Spring nu over de knuffel.	-0.85	Meertaligen
1.5	Bewegen	Plaats je twee handen in je zij.	-0.73	Meertaligen
9.2	Kabouters	Wat hebben Hamza en Mo bij/mee voor kabouter Kobus? Tik op de juiste tekening.	1.05	Nederlandstaligen
11.2	Klastaakjes	Wat doet juf Noura? Hoe kijkt juf Noura naar Fiene? Tik op de juiste tekening.	0.76	Nederlandstaligen
12.6	Konijntjes	Op welke tekening zie je het einde van het verhaaltje? Wat doen de konijntjes op het einde van het verhaaltje? Tik op de tekening die hoort bij het einde van het verhaal.	1.11	Nederlandstaligen
16.4	Mug en Olifant	Olifant voelde zich warm vanbinnen. Hij begon te blozen. Zijn wangen werden helemaal rood. Hij voelde voorzichtig met zijn slurf op zijn rug. Hij vond Mug en hij gaf haar een zachte aai met zijn slurf. Sindsdien zijn Mug en Olifant de dikste vrienden. Welke tekening past goed bij het verhaal?	0.86	Nederlandstaligen
17.3	Myriam	De juf zegt: 'Je moet in de rij staan en je beurt afwachten als ik flesjes water uitdeel.' Op welke tekening doet Myriam het goed?	0.91	Nederlandstaligen

		Tik op de tekening waar Myriam het goed doet.		
17.4	Myriam	De juf zegt: 's Morgens moet je je fruit in de gele bak op de tafel leggen. Je schriftje leg je op de kast.' Op welke tekening doet Myriam het goed? Tik op de tekening waar Myriam het goed doet.	0.88	Nederlandstaligen
17.6	Myriam	De juf zegt: 'De poppen moeten in de kast. De blokken en de auto's moeten op de mat blijven liggen.' Op welke tekening doet Myriam het goed? Tik op de tekening waar Myriam het goed doet.	0.91	Nederlandstaligen
20.2	Rommel in de eetzaal	Hang de jas die aan de stoel hangt, aan de kapstok in de gang. Trek een lijn van de jas naar de juiste plaats.	0.68	Nederlandstaligen
20.3	Rommel in de eetzaal	Leg het mes dat op de grond is gevallen, op de kar. Pas op: leg het in de bovenste bak van de kar. Trek een lijn van het mes naar de juiste plaats.	-0.99	Meertaligen
22.4	Speeltijd	In de hoek van de speelplaats staan fietsen. Aan twee fietsen hangen vlaggetjes. één vlag hangt hoog, de andere laag. Kleur de hoogste vlag.	-1.13	Meertaligen
26.3	Varken en Rups	Hoe voelt Rups zich als ze hoort dat ze moet verhuizen? Tik op de juiste tekening.	1.16	Nederlandstaligen
27.1	Verjaardagsfeest	Op de tafel staat een taart. Op de taart staan vijf kaarsjes. Teken één kaarsje bij op de taart.	-0.86	Meertaligen
27.5	Verjaardagsfeest	Op de tafel staan vijf glazen cola. Sommige glazen zijn leeg, andere zijn halfvol. Het middelste glas moet helemaal vol cola. Kleur het	-0.72	Meertaligen

		middelste glas tot het helemaal vol is.		
29.1	Waar is	Waar is het kind met een trui met strepen? Tik op het kind met de trui met streepjes.	-0.98	Meertaligen

Tabel 26: Resultaten DIF-analyses gezinstaal Nederlands-meertalig/anderstalig

Er zijn 16 items die significant moeilijker zijn voor één van de twee taalgroepen: 9 items zijn moeilijker voor kleuters die thuis enkel Nederlands spreken en 7 items zijn moeilijker voor kleuters die thuis (ook) een andere taal spreken.

De items die voor leerlingen met een meertalige achtergrond een hogere moeilijkheidsgraad hebben dan voor leerlingen met een eentalige achtergrond, komen vaker uit taken die gemiddeld genomen gemakkelijker zijn (bewegen, verjaardagsfeest, waar is), maar de specifieke items zijn daarom niet gemakkelijk. Verder valt het op dat er in deze items vaak sprake is van ruimtebegrippen, rangordes en getallen (o.a. over, twee, bovenste, hoogste, vijf, één, middelste, vol). De items die een hogere moeilijkheidsgraad hebben voor eentalige leerlingen, lijken vaker te maken hebben met gevoelens (11.2, 16.4, 26.3) en met het opvolgen van regels/afspraken (17.3, 17.4, 17.6, 20.2). Een duidelijke verklaring hiervoor is er niet meteen.

1.9.3 Opleidingsniveau van de moeder

We gaan na of het opleidingsniveau van de moeder leidt tot een verschillende inschaling van de moeilijkheidsgraad van de items. Een kleuter wiens moeder minstens een diploma secundair onderwijs heeft, krijgt code 0, een kleuter met een moeder die lager opgeleid is, krijgt code 1.

- Positieve DIF-contrast: item measure ligt hoger voor kinderen met een hoogopgeleide moeder. Het item is moeilijker als de moeder hoogopgeleid is.
- Negatieve DIF-contrast: item measure ligt hoger voor kinderen met een laagopgeleide moeder. Het item is moeilijker als de moeder laagopgeleid is.

Item nummer	Taak	Vraag	DIF-contrast	Moeilijker voor?
9.2	Kabouters	Wat hebben Hamza en Mo bij/mee voor kabouter Kobus? Tik op de juiste tekening.	0.92	Moeder hoogopgeleid
27.1	Verjaardagsfeest	Op de tafel staat een taart. Op de taart staan vijf kaarsjes. Teken één kaarsje bij op de taart.	-0.81	Moeder laagopgeleid

Tabel 27: Resultaten DIF-analyses opleidingsniveau moeder hoog-laag

Er zijn 2 items waarbij de moeilijkheid significant anders ingeschaald wordt naargelang de opleiding van de moeder. In één geval gaat het om het item dat moeilijker is voor kleuters met een hoogopgeleide moeder en in het andere geval is het item moeilijker voor kleuters met een laagopgeleide moeder. Deze twee items hebben ook een significante DIF-contrast voor gezinstaal.

1.10 Relatie tussen de vaardigheid van kleuters en hun persoonskenmerken

In dit stuk gaan we na of bepaalde persoonskenmerken van de kleuters een invloed hebben op de measure (luistervaardigheid) van de kleuters. We testen deze invloeden aan de hand van meerdere regressie-analyses.

1.10.1 Leeftijd (geboortemaand)

We hebben leeftijdsgegevens van 1920 kleuters. De gemiddelde leeftijd van de kleuters op 1 januari 2021 is 65,66 maanden (SD=3,45). Gemiddeld worden de kleuters uit onze steekproef 6 jaar rond 20 juni. De mediaan ligt op 66 maanden. De helft van de kleuters is dus geboren tussen januari en juni. De andere helft tussen juli en december.

We gaan na of kleuters die later op het jaar geboren zijn een lagere measure hebben dan kleuters die in het begin van het jaar geboren zijn. Enkel de kleuters van wie we zowel leeftijdsgegevens als taalscreeningsgegevens hebben, zijn mee opgenomen in onderstaande tabellen en grafieken (N=1896).

Tabel 28 geeft een overzicht van het aantal kleuters per geboortemaand, de gemiddelde measure en de standaardafwijking per geboortemaand.

Geboortemaand	Aantal kleuters	Gemiddelde measure	SD
Januari	190	1,69	1,38
Februari	161	1,99	1,42
Maart	144	1,67	1,46
April	162	1,54	1,52
Mei	139	1,60	1,44
Juni	164	1,62	1,50
Juli	177	1,53	1,39
Augustus	178	1,21	1,29
September	152	1,45	1,20
Oktober	143	1,27	1,44

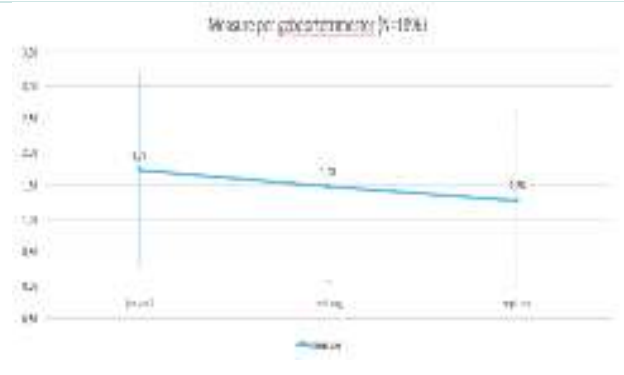
November	138	1,12	1,27
December	148	1,26	1,30
Totaal	1896	1,50	1,40

Tabel 28: Gemiddelde measure per geboortemaand

In Figuur 15 wordt deze informatie visueel voorgesteld.

Figuur 15: Measure per geboortemaand met standaarddeviatie

De grafiek en de tabel tonen aan dat de measure in dalende lijn gaat naarmate kinderen later op het jaar geboren zijn. Er zijn weliswaar enkele 'sprongen' in de trendlijn, maar de tendens lijkt dalend te zijn. Een ANOVA-test wijst uit dat de verschillen in measure tussen de geboortemaanden significant zijn (F=5,00 df=11; p<,001). De test voor lineariteit wijst uit dat er geen sprake is van 'afwijking van lineariteit' (F=1,39; df=10; p>,05). Bijgevolg concluderen we dat de trend lineair is.



Geboren tussen	Gemiddelde measure	SD	Aantal
Jan- April	1,73	1,45	657
Mei - Aug	1,48	1,41	658
Sept- Dec	1,28	1,31	581
Totaal	1,50	1,40	1896

Figuur 16: Measure per geboortetrimester met standaarddeviatie

Als we de kleuters groeperen per geboortetrimester, is er duidelijk sprake van een lineair en dalend verband tussen geboortemaand en de measure. Een ANOVA-test wijst uit dat de verschillen in measure tussen de geboortetrimesters significant zijn (F=16,05; df=2; p<,001). De test voor lineariteit wijst uit dat er geen sprake is van 'afwijking van lineariteit' (F=0,10; df=1; p>,05). Bijgevolg concluderen we dat de trend lineair is.

1.10.2 Gezinstaal en opleidingsniveau van de moeder

We gaan na of de gezinstaal en het opleidingsniveau van de moeder een effect hebben op de measures van de kleuters. We gaan ook na of er sprake is van een interactie-effect, en controleren voor leeftijd. We doen dit a.d.h.v. een univariate GLM. We hebben geldige data voor 1888 kleuters.

Descriptive Statistics

Dependent Variable: measure (vaardigheidsscore)

taal_j_1	opl_moeder_j_1	Mean	Std. Deviation	N
enkel NDL thuis	hoogopgeleide moeder	2,2198	1,28816	816
	laagopgeleide moeder	1,3920	1,08659	254
	Total	2,0233	1,29177	1070
ook andere taal thuis	hoogopgeleide moeder	1,0070	1,29935	347
	laagopgeleide moeder	,7056	1,19742	471
	Total	,8334	1,24983	818
Total	hoogopgeleide moeder	1,8579	1,40524	1163
	laagopgeleide moeder	,9461	1,20448	725
	Total	1,5078	1,40336	1888

Tabel 29: Measure gezinstaal – opleidingsniveau moeder

We zien dat de gemiddelde measure van kleuters met een hoogopgeleide moeder hoger ligt, en dit zowel voor kleuters die thuis enkel Nederlands spreken als voor kleuters die thuis (ook) een andere taal spreken. Nederlandstalige kleuters hebben gemiddeld genomen een hogere measure dan meertalige kleuters.

Tests of Between-Subjects Effects

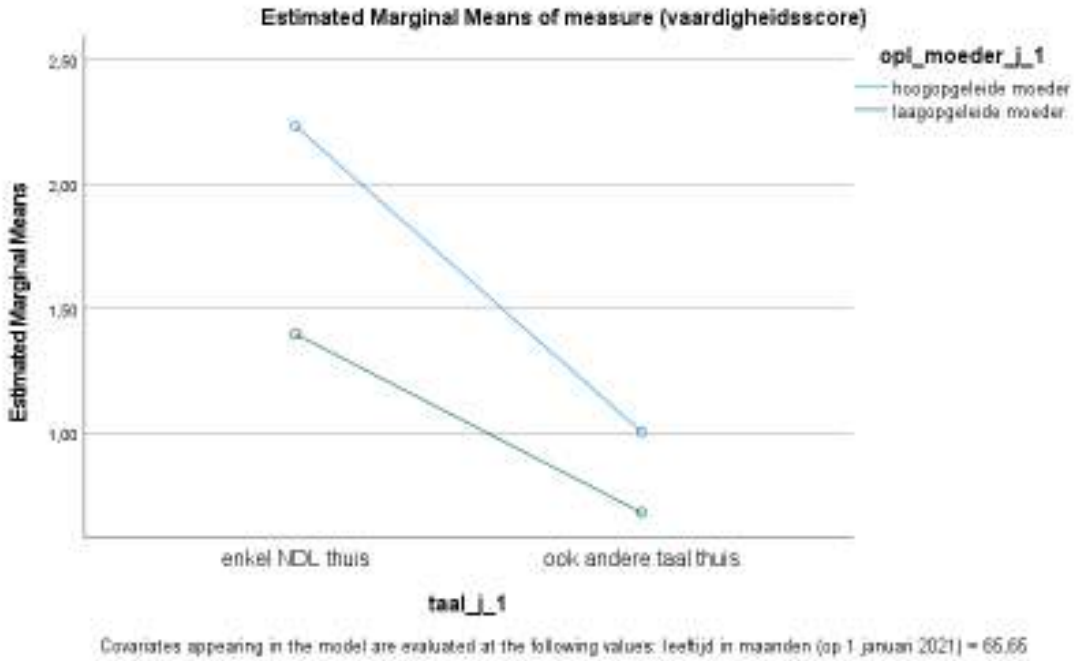
Dependent Variable: measure (vaardigheidsscore)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	922,384[a]	4	230,596	155,413	,000
Intercept	59,160	1	59,160	39,871	,000
leeftijd_maanden	115,205	1	115,205	77,644	,000
taal_j_1	371,588	1	371,588	250,436	,000
opl_moeder_j_1	132,300	1	132,300	89,165	,000
taal_j_1 * opl_moeder_j_1	25,645	1	25,645	17,284	,000
Error	2793,925	1883	1,484		
Total	8008,403	1888			
Corrected Total	3716,310	1887			

a. R Squared = ,248 (Adjusted R Squared = ,247)

Tabel 30: Interactie-effect gezinstaal – opleidingsniveau moeder

Zowel de gezinstaal als de opleiding van de moeder hebben een significant effect op de measure van het kind. We zien dat ook het interactie-effect significant is: het effect van de opleiding van de moeder is groter bij Nederlandstalige kleuters. In Nederlandstalige gezinnen is het verschil in measure tussen kleuters met een hoogopgeleide en laagopgeleide moeder groter dan in anders- of meertalige gezinnen. De leeftijd van het kind heeft ook een significante impact op de measure, maar wordt in onderstaande grafiek constant gehouden.



Figuur 17: Interactie-effect gezinstaal – opleidingsniveau moeder met controle voor leeftijd

1.10.3 Gezinstaal en leeftijd

We gaan na of de gezinstaal en leeftijd in maanden een effect hebben op de measure van de kleuters. We gaan ook na of er sprake is van een interactie-effect. We willen namelijk weten of het effect van leeftijd (geboortemaand) sterker of zwakker is in gezinnen waar (ook) een andere taal wordt gesproken.

We doen dit aan den hand van een meervoudige regressie en kiezen voor de forward-methode. Dit omdat we willen zien wat de meerwaarde is van elke toegevoegde voorspeller in het model. We maken de interactievariabele 'gezinstaal x leeftijd in maanden' aan. Deze voegen we toe aan het regressiemodel.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,420[a]	,177	,176	1,27377	,177	404,500	1	1885	,000
2	,454[b]	,206	,205	1,25103	,030	70,196	1	1885	,000

a. Predictors: (Constant), taal recode

b. Predictors: (Constant), taal recode, leeftijd in maanden (op 1 januari 2021)

Tabel 31: Significantie regressiemodel gezinstaal en leeftijd in maanden

Het finale regressiemodel bevat 'gezinstaal' en 'leeftijd in maanden' als voorspellers en is significant: F=244,77; df=2; p<0,001 (N=1888).

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	leeftijd in maanden (op 1 januari 2021)	,172[b]	8,378	,000	,189	,997
	interactie gezinstaal x leeftijd	1,057[b]	8,045	,000	,182	,024
2	interactie gezinstaal x leeftijd	,172[c]	,429	,668	,010	,003

a. Dependent Variable: measure (vaardigheidsscore)

b. Predictors in the Model: (Constant), taal recode

c. Predictors in the Model: (Constant), taal recode, leeftijd in maanden (op 1 januari 2021)

Tabel 32: Interactievariabele gezinstaal - leeftijd: geen significante voorspeller voor de measure

Enkel gezinstaal en leeftijd dragen significant bij aan het regressiemodel. Deze twee voorspellers verklaren samen 20,6% van de variantie in de measure. De interactievariabele maakt geen deel uit van het finale model en is dus geen significante voorspeller voor de measure. Het effect van leeftijd op de measure verschilt dus niet in Nederlandstalige of meertalige gezinnen.

1.10.4 Relatie tussen de vaardigheid van kleuters en hun gecombineerde persoonskenmerken

We weten dat de gemiddelde measure van kleuters hoger ligt als:

- het kind ouder is (leeftijd in maanden);
- het kind een hoogopgeleide moeder heeft (minstens secundair onderwijs);
- het kind thuis enkel Nederlands spreekt.

Via een regressiemodel gaan we na op welke manier en hoe sterk elk persoonskenmerk bijdraagt aan de measure. We kiezen voor de 'stepwise' methode, waarbij we eerst het persoonskenmerk 'gezinstaal' invoeren, vervolgens 'opleidingsniveau van de moeder' en als laatste 'leeftijd in maanden'. Op die manier achterhalen we meteen welk persoonskenmerk de sterkste impact heeft op de measure. Bij toevoeging van een persoonskenmerk aan het model wordt ook telkens nagegaan of het kenmerk een significante bijdrage levert aan het regressiemodel of niet.

Het regressiemodel bevat de drie persoonskenmerken en is significant: F=199,73; df=3; p<,001 (N=1888).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,420[a]	,177	,175	1,27377	,177	404,500	1	1885	,000
2	,458[b]	,210	,209	1,24810	,033	79,371	1	1885	,000
3	,491[c]	,241	,240	1,22335	,031	78,046	1	1884	,000

a. Predictors: (Constant), taal_j_1

b. Predictors: (Constant), taal_j_1, opl_moeder_j_1

c. Predictors: (Constant), taal_j_1, opl_moeder_j_1, leeftijd in maanden (op 1 januari 2021)

Tabel 33: Significantie regressiemodel leeftijd-opleidingsniveau moeder-gezinstaal Nederlands

Elk van de drie persoonskenmerken dragen significant bij aan het regressiemodel. Gezinstaal, opleiding van de moeder en leeftijd in maanden verklaren tesamen 24,1% van de variantie in de measure. Gezinstaal verklaart het grootste deel van de variantie (17,7%). De opleiding van de moeder verklaart 3,3% en de leeftijd van de kleuter 3,1%.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,023	,039		51,958	,000
	taal_j_1	-1,190	,059	-,420	-20,112	,000
2	(Constant)	2,156	,041		52,626	,000
	taal_j_1	-1,009	,062	-,359	-16,195	,000
	opl_moeder_j_1	-,561	,063	-,194	-8,909	,000
3	(Constant)	-2,575	,537		-4,794	,000
	taal_j_1	-1,022	,061	-,361	-16,869	,000
	opl_moeder_j_1	-,576	,062	-,200	-9,341	,000
	leeftijd in maanden (op 1 januari 2021)	,072	,008	,178	8,834	,000

a. Dependent Variable: measure (vaardigheidsscore)

Tabel 34: Leeftijd-opleidingsniveau moeder-gezinstaal Nederlands: significante voorspellers voor de measure

De beta-waarde laat toe om de invloed van de persoonskenmerken met elkaar te vergelijken. De invloed van gezinstaal is met voorsprong het grootst (β=-0,36). De impact van de opleiding van de moeder (β=-0,20) en van leeftijd in maanden (β=0,18) is iets kleiner, maar nog steeds significant. De

B-waarde geeft ons meer informatie over de precieze daling in measure tussen de verschillende groepen:

- anders- en/of meertalig: measure ligt 1,022 lager;
- laagopgeleide moeder: measure ligt 0,576 lager;
- één maand ouder: measure ligt 0,072 hoger.

Omdat regressiemodellen erg gevoelig zijn voor outliers, doen we deze analyses nog eens opnieuw zonder de kleuters die geïdentificeerd werden als outliers. We hanteren de strenge grens: kleuters met outfit MNSQ-waarde >= 2 worden nu uit de regressieanalyse gelaten.

We zien geen grote veranderingen. Het regressiemodel bevat eveneens de drie persoonskenmerken en is significant: F=183,83; df=3; p<,001 (N=1819).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics			
					R Square Change	F Change	df1	df2	Sig. F Change
1	,410[a]	,168	,168	1,26062	,168	367,528	1	1817	,000
2	,448[b]	,201	,200	1,23583	,033	74,612	1	1816	,000
3	,483[c]	,233	,232	1,21118	,032	75,677	1	1815	,000

a. Predictors: (Constant), taal_j_1

b. Predictors: (Constant), taal_j_1, opl_moeder_j_1

c. Predictors: (Constant), taal_j_1, opl_moeder_j_1, leeftijd in maanden (op 1 januari 2021)

Tabel 35: Significantie regressiemodel leeftijd-opleidingsniveau moeder-gezinstaal Nederlands zonder kleuters met outfit

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,993	,039		50,563	,000
	taal_j_1	-1,142	,060	-,410	-19,171	,000
2	(Constant)	2,126	,042		51,102	,000
	taal_j_1	-,960	,062	-,345	-15,448	,000
	opl_moeder_j_1	-,546	,063	-,193	-8,638	,000
3	(Constant)	-2,572	,542		-4,749	,000
	taal_j_1	-,983	,061	-,353	-16,124	,000
	opl_moeder_j_1	-,564	,062	-,199	-9,103	,000
	leeftijd in maanden (op 1 januari 2021)	,072	,008	,179	8,699	,000

a. Dependent Variable: measure (vaardigheidsscore)

Tabel 36: Leeftijd-opleidingsniveau moeder-gezinstaal Nederlands: significante voorspellers voor de measure zonder kleuters met outfit

De conclusies houden stand als we de outliers eruit laten. De drie persoonskenmerken dragen significant bij aan het regressiemodel. Gezinstaal, opleiding van de moeder als leeftijd in maanden verklaren in dit model tesamen 23,3% van de variantie in de measure. Gezinstaal verklaart opnieuw het grootste deel van de variantie (16,8%). De opleiding van de moeder verklaart 3,3% en de leeftijd van de kleuter 3,2%.

1.11 Inschatting taalvaardigheid door de leraar

1.11.1 Werkwijze

Voorafgaand aan de screening vroegen we aan de leraren om de luistervaardigheid van hun kleuters in te schatten. Op basis van die inschatting kennen we een kleurcode toe aan elk kind. Merk op dat deze kleurcode niet overeenstemt met de uiteindelijke cesuren (zie Hoofdstuk 7: Cesuren bij koala), die op dat moment nog niet waren bepaald.

- Groen: kleuters die volgens de leraar hoog zullen scoren en geen problemen zullen ervaren met de taalscreening.
- Geel: kleuters die volgens de leraar net niet voldoende zullen scoren en bij sommige taken of items moeilijkheden zullen ervaren.
- Rood: kleuters die volgens de leraar zwak zullen scoren en bij veel taken of items moeilijkheden zullen ervaren.

Naast een algemene inschatting vroegen we de leraren ook om hun kleuters te rangschikken van meest naar minst taalvaardig. Deze inschatting van de leraren van 'meest luistervaardig' naar 'minst luistervaardig' werd als volgt gehercodeerd.

- In een klas van 20 kleuters krijgt de meest taalvaardige kleuter score 1, de minst taalvaardige kleuter score 20.
- Door de toegekende score te delen door het aantal kinderen in de klas en deze vervolgens af te trekken van 1, krijgen we een score tussen 0 en 1 die de relatieve positie van het kind in de groep uitdrukt. Bijvoorbeeld: 1-(1/20)=0,95.
- Een hogere waarde wijst op een hogere rangschikking binnen de klasgroep voor taalvaardigheid.

We kijken nu naar de relatie tussen deze beide inschattingen van de leraar en de werkelijke measure van de kleuters.

1.11.2 Inschatting van leraren op basis van categorisering van kleuters ('hoog', 'net niet voldoende', 'zwak')

We bekijken eerst de taalinschatting van de leraar op basis van een categorising in de drie bovenvermelde groepen.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	groen	785	39,6	47,5	47,5
	geel	505	25,5	30,6	78,1
	rood	361	18,2	21,9	100,0
	Total	1651	83,3	100,0	
Missing	System	332	16,7		
Total		1983	100,0		

Tabel 37: Taalinschatting van de leraar aan de hand van een kleurcode

Voor 1651 kleuters werd een geldige kleurcode toegekend. De leraren verwachten dat iets minder dan de helft (47,5%) geen problemen zal ervaren met de taalscreening. Voor 21,9% van de kleuters verwachten ze grote moeilijkheden.

Met een correlatietest (Spearman's rho) gaan we na of er samenhang is tussen de inschatting van de leraar met een kleurcode en de werkelijke measure.

Correlations

			measure (vaardigheids score)	taalinschattin g LK kleurcode (recode)
Spearman's rho	measure (vaardigheidsscore)	Correlation Coefficient	1,000	,644**
		Sig. (2-tailed)	.	,000
		N	1955	1624
	taalinschatting LK kleurcode (recode)	Correlation Coefficient	,644**	1,000
		Sig. (2-tailed)	,000	.
		N	1624	1651

**. Correlation is significant at the 0.01 level (2-tailed).

Tabel 38: Correlatie tussen inschatting van de leraar aan de hand van een kleurcode en de measure van de kleuter

De hoogste categorie is 'groen' en de laagste is 'rood'. We zien een sterke en significante positieve correlatie tussen de taalinschatting van de leraar en de werkelijke measure van de kleuters (r=0,64;

p<,001). Een hogere inschatting van de leraar hangt samen met een hogere measure. Dit toont aan dat leraren het algemene taalvaardigheidsniveau van hun kleuters op een prestatieschaal met algemene taalvaardigheidscategorieën goed kunnen inschatten als hen daar expliciet naar wordt gevraagd.

1.11.3 Inschatting van leraren op basis van rangschikking

We bekijken vervolgens de inschatting als leraren de kleuters ordenen van meer naar minder taalvaardig.

Correlations

			measure (vaardigheids score)	taalinschattin g LK rangorde (recode)
Spearman's rho	measure (vaardigheidsscore)	Correlation Coefficient	1,000	,591**
		Sig. (2-tailed)	.	,000
		N	1955	1091
	taalinschatting LK rangorde (recode)	Correlation Coefficient	,591**	1,000
		Sig. (2-tailed)	,000	.
		N	1091	1112

**. Correlation is significant at the 0.01 level (2-tailed).

Tabel 39: Correlatie tussen inschatting van de leraar aan de hand van een rangschikking en de measure van de kleuter

Er werd voor 1112 kleuters een geldige rangschikking geregistreerd. We zien opnieuw een sterke en significante positieve correlatie tussen de inschatting van de leraar en de werkelijke measure van de kleuters (r=0,59; p<,001). Een hogere positie in de toegekende rangorde door de leraar hangt samen met een hogere measure. Dit toont aan dat leraren het algemene taalvaardigheidsniveau van hun kleuters op een prestatieschaal waarin kleuters gerangschikt worden volgens taalvaardigheid goed kunnen inschatten.

1.11.4 Inschatting van leraren op basis van categorisering én rangschikking van kleuters

Een regressieanalyse (stepwise multiple regression analysis) laat toe om na te gaan welke inschattingsmethode de sterkste voorspeller is van de uiteindelijke measure.

Het regressiemodel met de twee inschattingsscores als voorspellers voor de measure van de kleuter is significant: F=404,64; df=2; p<.001 (N=1050).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics			
					R Square Change	F Change	df1	df2	Sig. F Change
1	,642[a]	,412	,411	1,08866	,412	732,943	1	1048	,000
2	,660[b]	,436	,435	1,06534	,024	45,334	1	1047	,000

a. Predictors: (Constant), taalinschatting LK kleurcode (recode)

b. Predictors: (Constant), taalinschatting LK kleurcode (recode), taalinschatting LK rangorde (recode)

Tabel 40: Regressiemodel inschatting aan de hand van kleurcode en rangschikking

De twee inschattingsscores dragen significant bij aan het regressiemodel. Inschatting aan de hand van een kleurcode en rangschikking binnen de klasgroep verklaren tesamen 43,6% van de variantie in de measure. De kleurcode verklaart het grootste deel van de variantie (41,2%). De rangschikking voegt daar nog 2,4% verklarende variantie aan toe.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1,016	,102		-9,921	,000
	taalinschatting LK kleurcode (recode)	1,151	,043	,642	27,073	,000
2	(Constant)	-,847	,103		-8,188	,000
	taalinschatting LK kleurcode (recode)	,843	,062	,470	13,627	,000
	taalinschatting LK rangorde (recode)	1,138	,169	,232	6,733	,000

a. Dependent Variable: measure (vaardigheidsscore)

Tabel 41: Taalinschatting aan de hand van een kleurcode = sterkste voorspeller voor de measure van een kleuter

De beta-waarde laat toe om de bijdrage van de twee inschattingsscores met elkaar te vergelijken. We zien dat de taalinschatting met een kleurcode de sterkste voorspeller is van de uiteindelijke measure (β=0,47). Een inschatting door de leraar op basis van een kleurcode (of indeling in drie groepen) lijkt dus een efficiënte en betrouwbare manier om in het kader van brede beeldvorming een eerste indicatie te krijgen over de taalvaardigheid van een kleuter.

1.11.5 Inschatting van leraren op basis van rangschikking en controle voor leeftijd

We doen opnieuw een regressie-analyse, maar willen nu controleren voor leeftijd. We kiezen daarom voor een 'hierarchical multiple regression analysis', waarbij we de variabele 'leeftijd' in een apart blok zetten. Dit laat toe om de impact van (1) leeftijd en (2) inschattingen van de leraar afzonderlijk na te gaan.

Het volledige regressiemodel (met zowel leeftijd als de inschattingen door de leraar als voorspellers voor de measure) is significant: F=258,66; df=3; p<.001 (N=1023).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,136[a]	,019	,018	1,39897	,019	19,251	1	1021	,000
2	,657[b]	,432	,431	1,06500	,414	371,375	2	1019	,000

a. Predictors: (Constant), leeftijd in maanden (op 1 januari 2021)

b. Predictors: (Constant), leeftijd in maanden (op 1 januari 2021), taalinschatting LK kleurcode (recode), taalinschatting LK rangorde (recode)

Tabel 42: Significantie regressiemodel inschatting aan de hand van kleurcode en inschatting aan de hand van rangschikking (met controle voor leeftijd)

We zien dat leeftijd slechts 1,9% van de variantie in de measure verklaart en dat de twee inschattingsscores van de leraar (kleurcode en rangschikking) 41,4% van de variantie in de measure verklaren. Tesamen verklaren de voorspellers 43,2% van de variantie in de measure.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2,127	,855		-2,488	,013
	leeftijd in maanden (op 1 januari 2021)	,057	,013	,136	4,388	,000
2	(Constant)	-1,925	,656		-2,932	,003
	leeftijd in maanden (op 1 januari 2021)	,017	,010	,040	1,683	,093
	taalinschatting LK kleurcode (recode)	,835	,062	,465	13,371	,000
	taalinschatting LK rangorde (recode)	1,111	,171	,227	6,500	,000

a. Dependent Variable: measure (vaardigheidsscore)

Tabel 43: Leeftijd levert geen significante bijdrage aan het model en beïnvloedt de inschatting niet

Als we de bijdrage van elke variabele afzonderlijk bekijken, zien we dat leeftijd geen significante bijdrage levert aan het regressiemodel. Het is dus niet zo dat leeftijd in maanden de inschatting van de leraar vertekent. De inschattingsscores van de leraar dragen wel significant bij aan het model (zowel kleurcode als rangschikking). Zoals we hierboven al hadden vastgesteld, is de inschatting op basis van kleurcode (of met andere woorden de indeling in drie groepen) de sterkste voorspeller (β=0,46).

1.11.6 Inschatting van leraren en voor verschillende groepen: gezinstaal

We gaan na of de taalinschatting van de leraar de measure voor bepaalde groepen kleuters beter voorspelt. We kijken eerst naar gezinstaal.

- **Groep A: kleuters die thuis enkel Nederlands spreken**

Het regressiemodel met de twee inschattingsscores als voorspellers voor de measure van kleuters die thuis enkel Nederlands spreken is significant: F=113,04; df=2; p<0,001 (N=586).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
	taal_1 = enkel NDL thuis (Selected)				R Square Change	F Change	df1	df2	Sig. F Change
1	,506[a]	,256	,255	1,07927	,256	201,400	1	584	,000
2	,529[b]	,279	,277	1,06336	,023	18,610	1	583	,000

a. Predictors: (Constant), taalinschatting LK kleurcode (recode)

b. Predictors: (Constant), taalinschatting LK kleurcode (recode), taalinschatting LK rangcode (recode)

Tabel 44: Significantie regressiemodel inschatting aan de hand van kleurcode en inschatting aan de hand van rangschikking voor kleuters die thuis enkel Nederlands spreken

De twee inschattingsscores dragen significant bij aan het regressiemodel en verklaren voor deze groep tesamen 27,9% van de variantie in de measure. Dit is nog steeds significant, maar ligt wel een stuk lager dan het percentage voor de volledige groep. De kleurcode blijft de grootste bijdrage leveren aan de verklaring van de variantie in de measure (25,6%). De rangschikking voegt daar 2,3% verklarende variantie aan toe.

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-,280	,177		-1,583	,114
	taalinschatting LK kleurcode (recode)	,946	,067	,506	14,192	,000
2	(Constant)	-,092	,180		-,511	,610
	taalinschatting LK kleurcode (recode)	,656	,094	,351	6,969	,000
	taalinschatting LK rangorde (recode)	,988	,229	,217	4,314	,000

a. Dependent Variable: measure (vaardigheidsscore)

b. Selecting only cases for which taal_j_1 = enkel NDL thuis

Tabel 45: Inschatting aan de hand van kleurcode is sterkste voorspeller voor kleuters die thuis enkel Nederlands spreken

Als we kijken naar de beta-waarden, zien we dat de taalinschatting met een kleurcode ook voor de groep Nederlandstalige kleuters de sterkste voorspeller is van de uiteindelijke measure (β=0,35).

- Groep B: meertalige kleuters

Het regressiemodel met de twee inschattingsscores als voorspellers voor de measure van meertalige kleuters is significant: F=149,11; df=2; p<0,001 (N=436).

Model Summary

Model	R taal_j_1 = ook andere taal thuis (Selected)	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics			
					R Square Change	F Change	df1	df2	Sig. F Change
1	,618[a]	,381	,380	1,02226	,381	267,489	1	434	,000
2	,639[b]	,408	,405	1,00125	,027	19,398	1	433	,000

a. Predictors: (Constant), taalinschatting LK kleurcode (recode)

b. Predictors: (Constant), taalinschatting LK kleurcode (recode), taalinschatting LK rangorde (recode)

Tabel 46: Significantie regressiemodel inschatting aan de hand van kleurcode en inschatting aan de hand van rangschikking voor anders- en meertalige kleuters

De twee inschattingsscores dragen significant bij aan het regressiemodel en verklaren voor deze groep tesamen 40,8% van de variantie in de measure. Dit ligt een stuk hoger dan voor de groep Nederlandstalige kleuters. Dit wijst erop dat de taalinschatting van de leraar een betere indicatie is

voor de measure van meertalige kleuters dan deze van Nederlandstalige kleuters. Opnieuw is het de kleurcode die het grootste deel van de variantie verklaart (38,1%). De rangschikking voegt daar 2,7% aan toe.

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-1,064	,130		-8,170	,000
	taalinschatting LK kleurcode (recode)	1,029	,063	,618	16,355	,000
2	(Constant)	-,988	,129		-7,674	,000
	taalinschatting LK kleurcode (recode)	,789	,082	,474	9,596	,000
	taalinschatting LK rangorde (recode)	1,098	,249	,217	4,404	,000

a. Dependent Variable: measure (vaardigheidsscore)

b. Selecting only cases for which taal_j_1 = ook andere taal thuis

Tabel 47: Inschatting aan de hand van kleurcode is sterkste voorspeller voor kleuters die thuis enkel Nederlands spreken

Als we kijken naar de beta-waarden, zien we dat de taalinschatting met een kleurcode (of m.a.w. de indeling in drie groepen) ook voor de groep meertalige kleuters de sterkste voorspeller is van de uiteindelijke measure (β=0,47). De beta-waarde ligt een stuk hoger in deze groep. De taalinschatting met kleurcode (of met andere woorden de indeling in drie groepen) levert dus op een snelle manier betrouwbare informatie op over de taalvaardigheid van meertalige kleuters (meer dan voor Nederlandstalige kleuters).

1.11.7 Inschatting van leraren en verschillen tussen groepen: opleiding moeder

○ Groep A: kleuters met hoogopgeleide moeder

Het regressiemodel met de twee taalinschattingsscores als voorspellers voor de measure van kleuters met een hoogopgeleide moeder is significant: F=195,06; df=2; p<0,001 (N=663).

Model Summary

Model	R opl_moeder_j_1 = hoogopgeleide moeder (Selected)	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	,588[a]	,346	,345	1,12699	,346	350,160	1	661	,000
2	,610[b]	,371	,370	1,10491	,025	26,465	1	660	,000

a. Predictors: (Constant), taalinschatting LK kleurcode (recode)

b. Predictors: (Constant), taalinschatting LK kleurcode (recode), taalinschatting LK rangorde (recode)

Tabel 48: Significantie regressiemodel inschatting aan de hand van kleurcode en inschatting aan de hand van rangschikking voor kleuters met een hoogopgeleide moeder

De twee taalinschattingsscores dragen significant bij aan het regressiemodel en verklaren voor deze groep samen 37,1% van de variantie in de measure. Ook in dit model is het de kleurcode die het grootste deel van de variantie verklaart (34,6%). De rangschikking voegt daar 2,5% aan toe.

Coefficients[a,b]

Model		Unstandardized Coefficients: B	Unstandardized Coefficients: Std. Error	Standardized Coefficients: Beta	t	Sig.
1	(Constant)	-,849	,155		-5,465	,000
	taalinschatting LK kleurcode (recode)	1,135	,061	,588	18,713	,000
2	(Constant)	-,723	,154		-4,682	,000
	taalinschatting LK kleurcode (recode)	,845	,082	,438	10,311	,000
	taalinschatting LK rangorde (recode)	1,094	,213	,219	5,144	,000

a. Dependent Variable: measure (vaardigheidsscore)

b. Selecting only cases for which opl_moeder_j_1 = hoogopgeleide moeder

Tabel 49: Inschatting aan de hand van kleurcode is sterkste voorspeller voor kleuters met een hoogopgeleide moeder

Als we kijken naar de beta-waarden, zien we dat de taalinschatting met een kleurcode (of met andere woorden de indeling in drie groepen) ook voor de groep kleuters met een hoogopgeleide moeder de sterkste voorspeller is van de uiteindelijke measure (β=0,44).

- Groep B: kleuters met laagopgeleide moeder

Het regressiemodel met de twee taalinschattingsscores als voorspellers voor de measure van kleuters met een laagopgeleide moeder spreken is significant: F=125,67; df=2; p<0,001 (N=359).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics			
					R Square Change	F Change	df1	df2	Sig. F Change
1	,632[a]	,399	,397	,86703	,399	237,001	1	357	,000
2	,643[b]	,414	,411	,85636	,015	9,013	1	356	,003

a. Predictors: (Constant), taalinschatting LK kleurcode (recode)

b. Predictors: (Constant), taalinschatting LK kleurcode (recode), taalinschatting LK rangorde (recode)

Tabel 50: Significantie regressiemodel inschatting aan de hand van kleurcode en inschatting aan de hand van rangschikking voor kleuters met ene laagopgeleide moeder

De twee taalinschattingsscores dragen significant bij aan het regressiemodel en verklaren voor deze groep samen 41,4% van de variantie in de measure. De kleurcode verklaart 39,9% van de variantie in de measure en de rangschikking voegt daar 1,5% aan toe. De taalinschatting van de leraar verklaart voor kleuters met een laagopgeleide moeder dus iets beter de measure dan voor kleuters met een hoogopgeleide moeder, al is het verschil hier minder uitgesproken dan voor de opdeling op basis van gezinstaal.

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-,904	,135		-6,671	,000
	taalinschatting LK kleurcode (recode)	,976	,063	,632	15,395	,000
2	(Constant)	-,786	,140		-5,626	,000
	taalinschatting LK kleurcode (recode)	,762	,095	,493	8,038	,000
	taalinschatting LK rangorde (recode)	,856	,285	,184	3,002	,003

a. Dependent Variable: measure (vaardigheidsscore)

b. Selecting only cases for which opl_moeder_j_1 = laagopgeleide moeder

Tabel 51: Inschatting aan de hand van kleurcode is sterkste voorspeller voor kleuters met een laagopgeleide moeder

Als we kijken naar de beta-waarden, zien we dat de taalinschatting met een kleurcode (of met andere woorden de indeling in drie groepen) ook voor de groep kleuters met een laagopgeleide moeder de sterkste voorspeller is van de uiteindelijke measure (β=0,49). De beta-waarde ligt iets hoger in deze groep, maar minder groot dan bij de vergelijking op basis van gezinstaal. We kunnen wel concluderen dat taalinschatting met een kleurcode (of m.a.w. de indeling in drie groepen) correcte en betrouwbare informatie oplevert over de taalvaardigheid van kleuters, zelfs nog iets meer voor kleuters met een laagopgeleide moeder.

1.11.8 Inschatting van leraren en verschillen tussen groepen: leeftijd in maanden

We delen de kleuters op in drie geboorteperiodes: januari-april, mei-augustus en september-december. Voor elk van deze groepen kijken we naar de voorspellende waarde van de taalinschatting van de leraar op de measure van de kleuters.

Er werden geen opvallende verschillen tussen de drie geboorteperiodes vastgesteld. Over de gehele lijn voorspellen de taalinschattingen van de leraar de uiteindelijke measure van de kleuters goed en is de inschatting van de leraar met een kleurcode (of met andere woorden de indeling in drie groepen) de sterkste voorspeller.

1.12 Voorspellers van luistervaardigheid op school- en kindniveau (multilevel analyses)

In de voorgaande paragrafen werd vastgesteld dat leeftijd van een kind, thuistaalsituatie en opleidingsniveau van de moeder significante voorspellers zijn van de luistervaardigheid van een kleuter zoals gemeten door KOALA. In deze komende paragrafen gaan we na in welke mate het schoolniveau, naast het individueel kindniveau bijdraagt aan de verklaring van de luistervaardigheid van een kleuter. Daarvoor testen we in de eerste plaats of het schoolniveau significant de luistervaardigheid kan voorspellen; als dat het geval is, is het aangewezen om een model te testen waarbij niet alleen variabelen op kindniveau maar ook deze op schoolniveau worden opgenomen en getest.

1.12.1 Null model, random intercept op schoolniveau

Als eerste stap gaan we na of er significante verschillen zijn op schoolniveau (verschillen tussen scholen), en of het dus zinvol is om een onderscheid te maken tussen factoren die op kindniveau (level 1) en op schoolniveau (level 2) de variantie in de measure kunnen verklaren. Om te testen of er verschillen zijn tussen scholen testen we het random intercept model. De output van dit model wordt weergegeven in Tabel 52 en Tabel 53.

Tabel 52 geeft de informatiewaarde van het model dat het schoolniveau meeneemt. De fit van dit multilevel model wordt getest met een chi-square likelihood ratio test en weergegeven via -2 log-

likelihood (2LL). Hoe kleiner de waarde, hoe beter het model. De 2LL voor dit model is 6735,75, en zullen we hierna testen tegenover de 2LL van andere modellen.

Information Criteria[a]

-2 Log Likelihood	6735,753
Akaike's Information Criterion (AIC)	6741,753
Hurvich and Tsai's Criterion (AICC)	6741,765
Bozdogan's Criterion (CAIC)	6761,487
Schwarz's Bayesian Criterion (BIC)	6758,487

The information criteria are displayed in smaller-is-better form.

a. Dependent Variable: measure (vaardigheidsscore).

Tabel 52: Informatiewaarde van het multilevel model

Tabel 38 geeft de output voor het schoolniveau weer.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Residual		1,731224	,056498	30,642	,000	1,623958	1,845576
Intercept [subject = schoolcode_NUM]	Variance	,259151	,054736	4,735	,000	,171304	,392046

a. Dependent Variable: measure (vaardigheidsscore).

Tabel 53: parameterschattingen van covariantie

De significante p-waarde van de intercept (p<0.05) duidt op significante verschillen op schoolniveau. Het meenemen van het schoolniveau kan met andere woorden een deel van de variantie in luistervaardigheid van kleuters verklaren. We concluderen dat het zinvol is om zowel variantie op kindniveau als op schoolniveau te modelleren.

1.12.2 Drie multilevel modellen getest

Nu we weten dat het schoolniveau een significant deel van de variantie verklaart, kunnen we nagaan welke andere variabelen, op kindniveau en op schoolniveau, relevant zijn om mee te nemen in het model.

Tabel 54 geeft een overzicht van de verschillende uitgeprobeerde modellen, de schattingen van de opgenomen parameters en hun standaardfout. Significante voorspellers (p<.05 werden aangeduid met *).

In Tabel 54 lezen we de resultaten van het Nulmodel en van de modellen die we erna nog hebben getest. In Model 1 voegden we de kindkenmerken op niveau 1 toe, in Model 2 de schoolkenmerken op niveau 2. In Model 3A testten we voor de significante achtergrondvariabelen op kindniveau ook de bijdrage van de interactie tussen deze kenmerken. Model 3B is een vereenvoudiging van Model 3A, waarbij enkel de significante voorspellers werden opgenomen. Dit Model vormt de beste voorspeller van de luistervaardigheid van een kleuter in het kalibratie-onderzoek.

De kolom 'estimate' (schatting) geeft inzicht in de mate waarin een variabele de measure beïnvloedt, en de richting waarin deze beïnvloedt (positief of negatief effect). Een significant effect werd aangeduid met een * en in vet gepresenteerd. Voor onderwijscontexten wordt meestal een grenswaarde van p<.05 gehanteerd.

Op kindniveau werden de variabelen taal, opleidingsniveau van de moeder, schooltoelage en buurt getest. De verschillende modellen geven de invloed aan van deze variabelen indien de kleuter 'aantikt' op de bijbehorende OKI-indicator. Elk van deze variabelen kan ook op schoolniveau een rol spelen. De schatting geeft dan aan in welke mate het schoolgemiddelde van deze OKI-indicator, en dus het percentage kleuters dat aantikt voor deze indicator, van invloed is op de luistervaardigheidsmeasure.

Het finale model - Model 3B – geeft het beste inzicht in de variabelen die op school of kindniveau de measure beïnvloeden, en de mate en richting waarin deze invloed uitoefenen (positief of negatief effect). De kolom met de significantiewaarden geeft aan of er sprake is van een significant effect voor elke individuele variabele.

	Nulmodel		Model 1		Model 2		Model 3A		Model 3B	
	Est	SE	**Est**	SE	**Est**	SE	**Est**	SE	**Est**	SE
Kindniveau Intercept	**1,51***	(0,07)	**-2,40***	(0,51)	**-2,11***	(0,53)	**-2,02***	(0,53)	**-2,05***	(0,52)
Taal			**-0,90***	(0,06)	**-0,88***	(0,06)	**-1,25***	(0,12)	**-1,17***	(0,10)
Oplmoeder			**-0,47***	(0,06)	**-0,47***	(0,06)	**-0,94***	(0,16)	**-0,76***	(0,11)
Toelage			**-0,40***	(0,06)	**-0,38***	(0,06)	**-0,73***	(0,09)	**-0,73***	(0,08)
Buurt			0,06	(0,06)	0,06	(0,07)	0,06	(0,07)		
leeftijd_maanden			**0,07***	(0.01)	**0,07***	(0,01)	**0,07***	(0,01)	**0,07***	(0,01)
taal*oplmoeder							0,32	(0,22)		
taal*toelage							**0,51***	(0,15)	**0,48***	(0,12)

oplmoeder*toelage							**0,59***	(0,19)	**0,48***	(0,13)
taal*oplmoe*toelage							-0,22	(0,26)		
oplmoeder_school					0,34	(0,48)	0,29	(0,47)		
taal_school					-0,47	(0,31)	-0,42	(0,30)		
toelage_school					**-1,01***	(0,42)	**-0,90***	(0,41)	**-0,64***	(0,30)
buurt_school					0,32	(0,21)	0,29	(0,20)		
Schoolniveau Intercept	**0,26***	(0,05)	**0,13***	(0,03)	**0,11***	(0,03)	**0,10***	(0,03)	**0,11***	(0,03)
Residual	**1,73***	(0,06)	**1,32***	(0,04)	**1,32***	(0,04)	**1,29***	(0,04)	**1,29***	(0,04)
2LL	6735,75		5976,29		5962,21		5917,01		5924,50	
DF	3		8		12		16		10	

Tabel 54: Vergelijking van Multilevelmodellen

We bestuderen hieronder de bijdrage van de verschillende variabelen aan de individuele measure van een kind, zoals deze uit het Model 3A en 3B blijkt.

- Intercept: het significante intercept duidt op het gegeven dat er in de verschillende modellen, ook het laatste Model 3B nog significante verschillen zijn tussen kleuters.
- Gezinstaal: negatief significant effect op de measure. De measure van kleuters die aantikken op de indicator gezinstaal – kleuters met een meertalige thuiscontext- ligt 1,17 lager dan die van kleuters die niet aantikken en dus in een Nederlandstalige thuisomgeving opgroeien.
- Opleiding moeder: negatief significant effect op de measure. De measure van kleuters die aantikken op de indicator opleiding moeder, en dus een relatief laagopgeleide moeder hebben, ligt 0,76 lager dan die van kleuters met een hogeropgeleide moeder.
- Schooltoelage: negatief significant effect op de measure. De measure van kleuters die een schooltoelage ontvangen ligt 0,73 lager dan die van kleuters die geen toelage ontvangen.
- Buurt: geen significant effect op de measure. Kleuters die in een buurt met veel schoolse vertraging wonen hebben geen significant lagere of hogere measure dan kleuters in een buurt met beperkte schoolse vertraging.
- Leeftijd: positief significant effect op de measure. Voor elke maand die een kleuter ouder is, neemt de measure toe met 0,07. Dit betekent dat kleuters van januari versus december binnen dit model een verschil kennen van 0.84

We bekijken ook even op welke manier de interactie tussen de kindkenmerken de luistervaardigheidsmeasure beïnvloedt. Hiervoor hebben we in Model 3 enkel de variabelen getest die al een significante bijdrage leverden aan Model 2: de indicatoren ‘taal’, ‘opleiding moeder’ en ‘toelage’.

- Taal en toelage: De indicator 'Taal' en 'toelage' hebben elk een negatief effect op de measure van een kind (respectievelijk -1,17 en -0,73). De interactievariabele taal*toelage heeft daarbovenop een lichtjes positief effect van 0,48. Dit geeft aan dat aantikken voor de indicator 'taal' én voor de indicator' 'toelage' niet volledig cumulatief werkt: een kleuter die voor beide indicatoren aantikt, heeft een significant lagere measure dan een kleuter die voor geen enkele indicator aantikt. Het model voorspelt echter geen verschil van 1,90 (1,17 + 0,73), maar een kleiner verschil van 1,52 (1,90-0,48).
- Opleiding en toelage: De indicator 'opleiding' en 'toelage' hebben elk een negatief effect op de measure van een kind. De interactievariabele opleiding*toelage heeft daarbovenop een lichtjes positief effect van 0,48. Dit geeft aan dat aantikken voor de indicator 'opleiding' én voor de indicator 'toelage' niet volledig cumulatief werkt: Een kleuter die voor beide indicatoren aantikt, heeft een significane lagere measure dan een kleuter een kleuter die voor geen enkele indicator aantikt. Het model voorspelt echter geen verschil van 1,49 (0,76 + 0,73), maar een kleiner verschil van 1,01 (1,49-0,48).

Vervolgens kijken we naar de output van de variantie op level 2 van het model. Uit Model 1 en de significante verschillen op schoolniveau ($p<.05$) leidden we af dat er ook in het model waarin variabelen op kindniveau werden opgenomen, nog steeds significante variantie overblijft tussen kleuters van verschillende scholen.

We concluderen dat er nog steeds sprake is van onverklaarde variantie op niveau twee (niveau van de school) en dat het met andere woorden zinvol is om op zoek te gaan naar factoren op schoolniveau die deze variantie verder kunnen verklaren. In Model 2 werden al enkele verklarende factoren op schoolniveau opgenomen: het percentage aantikkers op school voor de indicatoren taal, opleiding van de moeder, toelage en buurt. Dit zijn ook de variabelen op schoolniveau waarover we beschikken

- % aantikkers voor Gezinstaal op schoolniveau: We zien geen significant effect op de measure. Kleuters in een school met veel kinderen die opgroeien in een meertalige context scoren niet significant verschillend van kleuters in een school met een overwegend Nederlandstalige thuiscontext.
- % aantikkers voor Opleiding moeder op schoolniveau: We zien geen significant effect op de measure. Kleuters in een school met veel kinderen met een laagopgeleide moeder scoren niet significant anders dan scholen waar dit niet het geval is.
- % aantikkers voor Schooltoelage op schoolniveau: We zien een significant effect op de measure. Kleuters in een school met veel kinderen die een schooltoelage krijgen hebben een measure die .64 lager is dan kleuters in een school met weinig kinderen die een schooltoelage krijgen.
- % aantikkers voor Buurt op schoolniveau We zien geen significant effect op de measure. Kleuters in een school in een buurt met veel kinderen met schoolse vertraging scoren niet significant anders dan scholen waar dit niet het geval is.

2 ▪ Kwalitatief: analyse van de verzamelde feedback

Gedurende het volledige kalibratieonderzoek hebben wij via verschillende kanalen feedback op de verschillende elementen van het taalscreeningsinstrument en de afnamecondities gekregen. In een eerste stap werd alle feedback verzameld. Na afronding van het kalibratieonderzoek hebben we alle feedback bij elkaar gelegd en gesynthetiseerd.

Deze feedback was voor het onderzoek zeer waardevol. Eerst en vooral gaf de mogelijkheid om feedback te geven een stem aan alle partijen die bij het onderzoek betrokken waren (bv. leraren, zorgcoördinatoren, jobstudenten of andere vrijwilligers die bij de toetsafname geassisteerd hebben, en bovendien ook de kleuters zelf). De feedback liet ons toe om een algemeen beeld te krijgen van hoe de scholen en toetsafnemers het screeningsinstrument hebben ervaren. Zo kwamen wij te weten welke aspecten gewaardeerd werden, over welke aspecten er kritiek of bedenkingen bestonden, en we kregen bovendien ook suggesties om bepaalde aspecten aan te passen of te verbeteren. Daarnaast liet de feedback ons toe om zeer gericht te gaan kijken of bepaalde taken nog aangepast moesten worden.

2.1 Kanalen voor het verzamelen van de feedback

Feedback op het screeningsinstrument en de afnamecondities werd op formele en informele wijze verzameld via de volgende kanalen:

- Registratieformulieren: Dit zijn de formulieren waarop de toetsafnemers tijdens de afname de scores van de concrete doe-opdrachten hoorden in te geven; op deze formulieren was er telkens ruimte voorzien voor opmerkingen bij de taken.
- Spontane feedback: Daarnaast kregen we ook spontane feedback. Hierbij gaat het bijvoorbeeld over opmerkingen van de leraren die tijdens de afnames meteen werden doorgegeven aan de onderzoekers; notities van de toetsafnemers en toetsassistenten; spontane feedback via mail…
- Feedbackformulieren: Wanneer de testafnames waren afgerond hebben we naar alle deelnemende scholen via mail een feedbackformulier (formaat: Word document) verstuurd. Daarop werden respondenten expliciet uitgenodigd om feedback te geven op onderstaande aspecten:

    - Praktische organisatie:

        - Hoe verliep het praktische aspect van de afnames (bv. de voorbereiding, het lokaal, de materialen…)?
        - Welke praktische tips heb je voor de afnames van de definitieve taalscreening (schooljaar 2020-2021)? Wat zou jij anders aanpakken?
        - Zijn de instructies voor toetsafnemers duidelijk? Hoe kan het nog beter?

- Taken en clusters:
    - Wat vind je van de verschillende soorten taken (doe-opdrachten concreet, doe-opdrachten op papier, meerkeuzevragen op papier)? Wat werkte goed/minder goed?
    - Welke opmerkingen (positief/negatief) heb je nog bij taken algemeen?
    - Welke opmerkingen (positief/negatief) heb je nog bij specifieke taken? Noteer de titel van de taak en je opmerkingen.
    - Welke bedenkingen (positief/negatief) heb je nog bij afbeeldingen algemeen?
    - Welke bedenkingen (positief/negatief) heb je nog bij specifieke afbeeldingen? Noteer de titel van de taak en je opmerkingen.
- Reacties van de kleuters:
    - Hoe reageerden de kleuters? Deden ze uitspraken over bepaalde taken of over de afname algemeen?
- Algemene feedback op het screeningsinstrument
    - Vind je de taalscreening bij 5-jarige kleuters zinvol?
    - Welke inhouden/hoofdstukken moeten we zeker voorzien in de handleiding van de definitieve taalscreening (schooljaar 2020-2021)?
    - Hoe kunnen we ervoor zorgen dat scholen ook echt aan de slag gaan met de resultaten?
    - Hoe kunnen we ervoor zorgen dat de taalscreening leidt tot meer gelijke onderwijskansen?

2.2 Samenvatting van de feedback

2.2.1 Feedback op praktische aspecten van de screening

Heel wat van de ontvangen feedback had betrekking op praktische aspecten van de organisatie en afname van de taalscreening:

- De afnames van het screeningsinstrument zijn in het kalibratieonderzoek goed verlopen. Voldoende voorbereiding bleek hierbij wel een vereiste te zijn, en deze werd vaak als redelijk intensief ervaren: er kwam heel wat voorbereidend werk bij kijken (een geschikt lokaal zoeken; alle instructies op voorhand doornemen; materiaal klaarzetten…), en dit nam redelijk veel tijd in beslag (bijvoorbeeld 1 voormiddag).

  Het vraagt toch wel wat voorbereidend werk om alles goed door te lezen over hoe het in elkaar zit. Eens je dit weet, valt het werk wat betreft voorbereiding goed mee. (leraar)

- Velen pleiten ervoor om de afnames door twee toetsafnemers te laten gebeuren. Op die manier kan een toetsafnemer zich concentreren op het voorlezen van de instructies, en de andere

toetsafnemer kan bijvoorbeeld helpen met het omdraaien van de blaadjes, er mee voor zorgen dat de kleuters aandachtig blijven, en er mee op letten dat de kleuters niet spieken. Bovendien kan de tweede toetsafnemer zich bezighouden met de kleuters die hun beurt afwachten tijdens de individuele afnames van de concrete doe-opdrachten. Bij de digitale afnames was de aanwezigheid van minstens twee toetsafnemers een must.

Een testafname met twee leiders is noodzakelijk om een zo goed mogelijk klimaat te creëren. (leraren van deelnemende school)

Zorg dat je dit niet alleen afneemt. Het is altijd handig om nog een extra juf te hebben die de kleuters ook nog even kan helpen. Of gerust stellen. (leraar)

- De duur van de afnames werd ervaren als te lang voor de kleuters (bijna unaniem). Als gevolg hiervan werd er vaak een concentratieverlies bij de kleuters gerapporteerd. Heel wat deelnemers pleiten daarom voor een reducering van het takenpakket (minder taken en/of items) of voor het opsplitsen van de afname in meerdere delen (bijvoorbeeld: meerkeuzevragen op dag 1; individuele, concrete doe-opdrachten op dag 2).

  De test zelf duurde te lang om in 1 keer af te leggen. Beter in 2 blokken verdelen en verspreiden over 2 dagen. Op die manier blijven de kleuters meer gefocust. Nu was er ook geen pauze voorzien tijdens de test. Dit is echt te lang! (leraren van deelnemende school)

- De voorbereidende online sessies en de helpdesk werden zeer goed onthaald. Een school pleit er zelfs voor om gelijkaardige hulpplatformen aan te bieden bij de definitieve afname van de screening, bijvoorbeeld door middel van webinars.

  De digitale infoworkshop, PowerPoint en afvinklijst waren zeer handige tools. (toetsassistent)

- De instructies voor de toetsafnemers werden als zeer duidelijk en als voldoende ervaren (bijna unaniem). Hier en daar zouden bepaalde aspecten nog explicieter aangegeven mogen worden (bv. wanneer kleuters hun blad moeten omdraaien, of wanneer kleuters hun potlood moeten neerleggen en naar de leraar moeten luisteren).

  Voor de rest moet ik wel zeggen dat alles super duidelijk was! Je kan volgens mij niets fout doen, omdat alles zo goed is uitgelegd. (toetsassistent)

  We vonden de uitleg van de bundel klaar en duidelijk. (zorgcoördinator)

- Het materiaal om op voorhand klaar te leggen (bv. potloden en stiften, poppetjes en andere materialen voor bepaalde individuele doe-opdrachten) bleek gemakkelijk te vinden te zijn op school.

  Het materiaal dat gevraagd werd om klaar te zetten was makkelijk te verzamelen. Dit is zeker aanwezig in een school. (leraren van deelnemende school)

- Wat betreft de papieren bundels van taken en instructies gaven meerdere afnemers aan dat ze in de toekomst liefst een zo gemakkelijk mogelijk hanteerbaar formaat zouden krijgen. Het actuele formaat zou nog aan de onhandige kant zijn omdat er onder andere te veel losse bladen tussen zitten.
- Over het algemeen werd positief gereageerd op het werken met groepjes van 4 of 5 kleuters. (Twee zorgjuffen raden wel aan om afnames van zeer taalzwakke kleuters per twee of zelfs één op één te doen).

  Door deze screening in groepjes van 5 af te nemen kan je elke kleuter beter opvolgen. In de klas is daar niet altijd tijd voor. (leraar)

- Over het algemeen werd de digitale versie van de afname als een meerwaarde gezien omwille van meerdere voordelen (bv. geen papierwerk; rechtstreekse registratie van de scores; indirect met de kleuters werken rond media-doelen). Ook de kleuters vonden het leuk om op de tablets te werken. Het tikken op het scherm was voor de kleuters een gemakkelijkere respons dan het tekenen van een kring; meerdere kleuters leken namelijk nog moeite te hebben met schrijfmotoriek. Over de digitale afname werd verder door de toetsassistenten nog meegegeven dat de mogelijkheid om zowel offline als online af te nemen een grote meerwaarde was. Desalniettemin blokkeerde het systeem al wel eens, waardoor af en toe data verloren gingen. Toetsassistenten wijzen daarom op het belang van degelijke een omgeving en programma's die een degelijke registratie toelaten.

2.2.2 Feedback op de taken en taaktypes

Verder ontvingen we ook feedback op individuele taken.

- Algemene opmerkingen:
    - Over het algemeen werd er positief op de taken gereageerd. Deze waren 'uitnodigend' voor de kleuters. Ook de instructies waren 'goed begrijpbaar'" voor de kleuters geformuleerd. Een leraar gaf ook expliciet aan dat de 'inbedding in een verhaal en betekenisvolle situaties' de taken 'aangenaam' maakt voor de kleuters.
    - De variatie in de opdrachten werd over het algemeen als een grote meerwaarde gezien.

  De verschillende taken is een goede keuze. Hier kan je de kleuters op verschillende vlakken observeren. (zorgjuf)

  De afwisseling tussen de verschillende opdrachten vind ik een pluspunt. Door de variatie blijft het aangenaam voor de kinderen. (zorgleraar)

    - Ook het gebruik van interculturele voornamen in de taken (bv. Hamza, Fiene, Mo, Rosanne, Myriam…) werd vaak aangehaald en als een pluspunt gezien.
    - De afbeeldingen werden zeer positief onthaald: 'kindvriendelijk, duidelijk, spraken de kinderen aan, mooi, hedendaags, helder, leuk, aantrekkelijk'. Enkel bij de digitale versie van de screening vielen enkele tekeningen met veel details wat klein uit.

- Individuele doe-opdrachten:
    - Deze typetaken werden ervaren als zeer leuk voor de kleuters; volgens de feedback waren deze 'natuurlijk' en 'op kindniveau'. Daarnaast waren de fysieke doe-opdrachten een welgekomen afwisseling na de meerkeuze-opdrachten waarbij de kleuters moesten stilzitten.

De doe-opdrachten waren de leukste voor de kinderen. De opdrachten waren duidelijk en op niveau van de kinderen.(klasleraar)

De doe-opdrachten individueel bij de juf waren het fijnst voor hen en mij. (klasleraar)

Doe-opdrachten 'met eigen lichaam' verlopen vlotter dan de andere doe-opdrachten. Deze afwisseling is ook welkom voor de kleuters; ze zijn enthousiast om deze opdracht uit te voeren. (pedagogische begeleiders)

Anderzijds waren de doe-opdrachten het meest natuurlijke en konden de kleuters, die het moeilijk hadden met zitten en focussen, dan wel heel goed tonen dat ze wel luistervaardig waren. (toetsassistent)

    - Sommige toetsafnemers gaven aan dat ze zich onzeker voelden bij de scoring van individuele doe-opdrachten.

Sommige doe-opdrachten concreet (vooral dan die waarbij de kleuters moeten vertellen), vind ik persoonlijk soms moeilijk om daar een score aan te geven. Hoe objectief kan je hierin zijn? (leraar)

    - Uit de feedback blijkt ook dat er bij deze taken voldoende aandacht aan de afnamecondities moet worden besteed zodat de kleuters op hun gemak zijn, ze niet worden gestoord door kleuters die hun beurt moeten afwachten, en er niet gespiekt kan worden.
    - Bijna unaniem werd er aangegeven dat de waarom-vragen (bv. Waarom vind je iets lekker? Waarom vind je iets leuk?) problematisch waren. Op deze vragen bleek een zinvol antwoord niet altijd gemakkelijk te geven; daarom waren ze dan ook moeilijk te beantwoorden voor de kleuters en moeilijk te scoren voor de toetsafnemers.

Doe-opdrachten verliepen vrij vlot, de waarom vragen (taken Lievelingseten, Spelen, Dieren uitbeelden) zijn voor de kleuters heel erg moeilijk om te beantwoorden en moeilijk te scoren door de leraar. (taal- en beleidscoach)

De waarom-vragen vond ik moeilijk voor de kinderen, ik beeldde me in wat ikzelf zou antwoorden en dat kwam ook niet zo vlot. (zorgcoördinator)

- Doe-opdrachten op papier:
    - Deze werden algemeen als positief ervaren.

        - *De eerste opdracht waar meestal een afbeelding aan te pas kwam vond ik persoonlijk een duidelijke en goede luisteropdracht waar je meteen een goed zicht krijgt op de luistervaardigheden van de kleuters. (leraren van deelnemende school)*

    - De opdrachten waarbij kleuters iets moeten inkleuren of een kruisje moeten zetten worden als gemakkelijker ervaren dan de opdracht waarbij de kleuters zelf moesten tekenen (taak Verjaardagsfeest). Blijkbaar vormde zelf tekenen voor sommige kleuters hierbij een drempel.

De taken waarbij vormgeving aan bod komt, zorgen voor drempels bij de kleuters. Wanneer ze gewoon onderdelen op een prent moeten kleuren (de speeltuin) gaat dat vlot. Als het gaat om iets 'tekenen' op een prent (de verjaardag) zorgt dat voor stress, onzekerheid en bij sommige kleuters voor wat faalangst. 'Ik kan geen bloem tekenen', 'ik weet niet hoe je snoepjes moet tekenen'. (leraren van deelnemende school)

- Meerkeuze taken:
    - De taken die zeer dicht bij de leefwereld van de kleuters aansluiten, werden bijzonder geprezen (bv. taken die te maken hebben met de verschillende hoeken in een kleuterklas, of waarbij het over typische klasafspraken gaat…).
    - Vooral bij de meerkeuzetaken was het vaak een uitdaging om de aandacht van de kleuters te houden.

Bij de meerkeuzevragen moest je als leraar goed opletten dat de kinderen de 'gehele' opdracht beluisterden vooraleer ze een keuze maakten! Het kwam toch wel vaker voor dat kinderen al een tekening aankruisten vooraleer de volledige opdracht gelezen werd. Het was dus echt wel nodig om met twee leraren aanwezig te zijn bij de afname van de toets. Indien één leraar dit zou doen, zou de controle bij de leerlingen minimaal zijn en voor een vertekend beeld zorgen. Vaak was er toch wel even gebrek aan concentratie bij de kinderen. (leraren van deelnemende school)

    - Verhaaltaken: Over het algemeen werden de verhalen goed onthaald en als kindvriendelijk en aansluitend bij de leefwereld van de kleuters gezien. De kortere verhalen waren voor de kleuters gemakkelijker dan de langere verhalen. Het werd als zeer moeilijk voor de kleuters ervaren wanneer de kleuters naar een (stuk van een) verhaal moesten luisteren en een vraag kregen, en dan pas de antwoordopties te zien kregen.

Verhaaltjes ook aansluitend bij leefwereld van de kleuters en vragen aangepast aan wat een kleuter zou moeten begrijpen. (zorgcoördinator)

    - Sommige deelnemers vrezen bij deze taken ook voor de impact van het feit dat de kleuters deze manier van werken (meerkeuze, een kring tekenen rond de juiste afbeelding…) nog niet kennen.

Niet alle kleuters kennen de opdracht 'trek een kring rond…'. (pedagogische begeleiders)

Hier zien wij een gevaar voor 'teach the testing'. (leraren van deelnemende school)

2.2.3 Reacties van de kleuters

Toetsafnemers rapporteerden ook over de reacties van de kleuters.

- Over het algemeen waren de kleuters zeer enthousiast om deel te nemen. Ze vonden het leuk en spannend, en ze waren trots dat ze even in de schoenen mochten staan van 'kindjes uit het eerste leerjaar'. Ze vonden ook leuk dat ze in kleine groepjes met hun leraar mochten werken. Maar het duurde voor de meeste kleuters wel te lang (zie hoger); hierdoor verloren ze hun aandacht en concentratie.

  Ze zeiden achteraf dat ze het leuk vonden. (leraar)

  De kleuters vonden het leuk maar wel veel. (zorgcoördinator)

  Eigenlijk vonden de meeste kleuters het best wel leuk om eens in een boekje te werken. (Leraar)

  Ze waren wel fier en vonden het leuk om te doen. (leraren van deelnemende scholen)

  Over het algemeen reageerden de kleuters positief op het feit dat ze eens mochten tonen wat ze kunnen. (leraren van deelnemende school)

  We hadden niet de indruk dat kinderen het niet graag deden. Ze genoten van het werken in een klein groepje samen met de juf. Ze voelden zich 'groot' zoals de kindjes in het eerste leerjaar aan een bankje met een bundeltje en een potlood. (leraren van deelnemende school)

- Over het algemeen zoeken kleuters wel de bevestiging dat ze het goed doen. Sommige kleuters waren wel onzeker en bang om fouten te maken. Dit toont aan hoe belangrijk het is om tijdens de afnames de kleuters aan te moedigen, hen complimentjes te geven en hen te zeggen dat ze niets fout kunnen doen.

  Zowel ik als de leraar gaven veel complimentjes, we merkten dat de kleuters bevestiging zochten of wouden tonen wat ze aangeduid hadden. Onze aanmoedigingen waren trouwens niet gelinkt aan of de kleuter juist of fout antwoordde. (toetsassistent)

- Digitale afnames: de kleuters vonden het enorm leuk om op de tablets te mogen werken.
- Bladeren omdraaien en cirkels tekenen was niet evident voor alle kleuters. Op het scherm tikken bij digitale afnames was gemakkelijker.

2.2.4 Algemene feedback op het screeningsinstrument

Tot slot reflecteerden een aantal afnemers en leraren over het screeningsinstrument in het algemeen. De volgende punten kwamen hierbij aan bod.

- Is een taalscreening bij 5-jarige kleuters zinvol?
    - Over het algemeen werd de taalscreening als zeer zinvol gezien. Sommige leraren gaven aan dat een screening kan zorgen voor een objectiever, veelzijdig beeld van de kleuters, en dat ze het normeringsaspect van de screening waarderen. Het zou ook helpen om gerichter te kunnen ingaan op tekortkomingen bij de kleuters. Sommige leraren gaven bijvoorbeeld ook aan dat er door de testafnames "verrassingen" aan het licht kwamen.

Zeker zinvol, ben heel blij dat er eindelijk een taalscreening komt die wat meer aanleunt bij de leefwereld van de kleuters. De taaltesten waren echt aan vernieuwing toe. Pluim hiervoor. (zorgcoördinator)

Taalscreening van 5-jarige kleuters is en blijft ontzettend belangrijk en is dan ook zeer zinvol. (zorgcoördinator)

Als leraar krijg je een meer objectief beeld hoe ver kleuters staan in vergelijking met grote groep kleuters van dezelfde leeftijd. (leraar)

We vinden het zinvol omdat je wel een mooi beeld krijgt waar je als school van kan leren. Bepaalde dingen vielen echt op tijdens de afname, de juf is geprikkeld om ermee aan de slag te gaan. Niet om de toets te oefenen, maar om de kinderen een nog beter aanbod te geven. (leraren van deelnemende school)

Kleuter H komt bijna niet tot spreken op school en thuis. Juf A vermoedt een taalstoornis of een spreekstoornis. Bij de opdrachten blijkt dat H bijzonder snel en steeds correct de meerkeuzevragen kan beantwoorden. Juf A geeft aan dat ze nu zeker weet dat H een sterke talige kleuter is die logopedisch een probleem heeft. Ze zal hierover in gesprek met de ouders gaan. (pedagogisch begeleider)

    - Sommige leraren gaven aan dat ze de resultaten van de screening in de toekomst graag zouden gebruiken tijdens oudercontacten om duidelijk aan ouders te kunnen tonen wanneer een kind niet voldoende taalvaardig in het Nederlands is.
    - Verder hebben heel wat deelnemers aan de screening ook bedenkingen geuit. Zo werd vastgesteld dat het bij de screening over een momentopname gaat en de betrouwbaarheid daarvan werd in twijfel getrokken. Verder zou de screening ook voor stress bij de kleuters kunnen zorgen en was niet voor iedereen duidelijk wat de meerwaarde van de screening tegenover de eigen observaties van de leraar zou zijn. Sommigen vrezen ook dat de screening ingezet zou kunnen worden om de toelating van de kleuters voor het eerste leerjaar te bepalen.

Weinig nieuwe informatie rond de taalvaardigheid van kleuters (naast eigen observaties). (pedagogische begeleiders)

Persoonlijk vind ik dit een momentopname. Een momentopname is nooit representatief omdat het kind een slechte dag kan hebben of andere factoren die een rol kunnen spelen. (leraren van deelnemende school)

Ook levert de test stress op voor kinderen. Je probeert dit wel in een jasje te steken alsof het geen test is, maar ze voelen dit aan. Kleuters die anders spontaan zijn, kruipen nu in hun schulp. (leraren van deelnemende school)

Gaan de resultaten ons iets meer zeggen dan wij al weten is de vraag en wat kunnen wij dan nog meer doen? (leraren van deelnemende school)

Ik ben bang dat de resultaten van de screening gebruikt zullen worden als 'stok achter de deur' om kleuters niet naar het eerste leerjaar te laten gaan, of naar het buitengewoon onderwijs te laten doorstromen. Dat zou de gelijke onderwijskansen geen deugd doen. Veel van deze vaardigheden – taal of andere – leren kinderen tijdens hun verdere schoolloopbaan.(toetsassistent)

- Meerdere keren werd ook de wens geuit om de taalscreening in te kunnen zetten om vooruitgang te meten.

We zouden na de screening moeten in kaart kunnen brengen of een leerling vooruit gaat of niet. (zorgcoördinator)

Wij vinden een opvolgscreening op het einde van de derde kleuterklas bij de kinderen die uitvielen ook zinvol zodat er al dan niet leerwinst kan vastgesteld worden. (leraren van deelnemende school)

- Hoe kunnen we ervoor zorgen dat scholen aan de slag gaan met de resultaten?
    - Bijna unaniem werd er gepleit voor een gemakkelijk hanteerbaar systeem dat de resultaten van screening onmiddellijk en op een gemakkelijk interpreteerbare manier weergeeft, op kleuter- en op klasniveau.
    - Uit de feedback bleek bovendien dat het zeer belangrijk is om voldoende toe te lichten waarom de taalscreening een meerwaarde is ten opzichte van observaties door de leraren zelf, en wat de taalscreening voor de leraren kan betekenen.

Door een visie erbij te geven waarom deze taaltesten zo belangrijk zijn. (zorgcoördinator)

- Er is een grote vraag naar een duidelijk hoofdstuk in de handleiding 'Wat na de screening?' dat voorzien moet zijn van zeer concrete tips en aanbevelingen over stappen die ondernomen kunnen worden om met de resultaten aan de slag te gaan en de taalontwikkeling van de kleuters extra te stimuleren.

In de handleiding zeker een hoofdstuk voorzien met 'wat na deze screening – tips voor de verdere klaspraktijk'; bv. bij voorlezen meer ingaan op de emoties, achteraf reflecteren op waargenomen emoties. (pedagogische begeleiders)

Uit de resultaten kunnen afleiden welke acties kunnen opgezet worden op kind- en klasniveau. (leraren van deelnemende school)

Door voldoende tips en aanbevelingen te doen over hoe er met de kinderen die zwak scoren kan gewerkt worden. En hierbij zeker ook rekening houden met de verschillen tussen de scholen. Wij zijn een school met reeds heel veel leerlingen waarbij de thuistaal niet Nederlands is, hierdoor moeten we reeds goed inzetten op taalstimulering. (leraren van deelnemende school)

Hoe we aan de slag kunnen met de resultaten weten we nog niet concreet. We hebben reeds een erg uitgebouwd taalbeleid, waarin we veel acties ondernemen om de taalarme kleuters te ondersteunen op verschillende vlakken. Echter al onze kleuters krijgen dit aanbod, ook taalsterke worden gestimuleerd om taal te horen, te begrijpen en te gebruiken. Het zit reeds verwerven in ons dagelijks schoolaanbod. (leraren van deelnemende school)

Wanneer we uit de testanalyse kunnen afleiden aan welke ondersteuning de kleuters nood hebben. Eventueel ook verwijzing naar materialen waarmee gewerkt kan worden. (Taal- en beleidscoach)

Remediëringsbundel voorzien, tips geven hoe hiaten kunnen worden aangepakt. (leraren van deelnemende school)

We hopen dan ook dat deze test van jullie ons nog meer tips brengt om in de praktijk aan de slag te gaan met de resultaten... (zorgleraar)

Ik denk dat het ook sterk zou zijn als we enkele tips toevoegen van wat je als leraar kan doen om resultaten te verbeteren. Daarin zou ik niet alleen focussen op individuele remediëring van kleuters die uitvallen, maar ook op taalvaardigheidsonderwijs in de brede basiszorg. Een kind dat bijvoorbeeld sterk scoort op naar een instructie luisteren, maar niet op naar een verhaal luisteren, kan sterker worden door vaker naar verhalen te luisteren, expliciet betrokken te worden tijdens het voorlezen, in interactie gaan over verhalen met leraar of andere kinderen … Dat kan je zeker in de klas doen! (toetsassistent)

- Verder hebben sommigen ook de wens geuit om de ouders van de kleuters meer te kunnen betrekken om het taalontwikkelingsproces in goede banen te kunnen leiden.

Voorzien van een inspiratiegids van hoe we ouders kunnen sensibiliseren (belang van taal) en hun aanzetten om actief met taal bezig te zijn thuis en buiten een schoolse context. (zorgcoördinator)

- Meerdere deelnemers gaven ook aan dat ze meer middelen (bv. extra personeel, lesuren, logopedisten toegankelijk maken voor ouders die het op financieel vlak niet

breed hebben) nodig hebben om in de praktijk met de resultaten aan de slag te gaan en de taalverwerving van de kleuters extra te kunnen stimuleren. Dit zou ook helpen om meer gelijke onderwijskansen te scheppen.

Een taalbad is nodig, maar is in de realiteit niet altijd haalbaar wegens een te kort aan middelen. (leraren van deelnemende school)

Onze leraren werken nu al kindgericht. Ze hebben vaak handen en tijd te kort om dit allemaal waar te maken daarom: Extra uren zorg i.v.m. taal.; extra ondersteuning voor de leraren. (zorgcoördinator)

2.3 Samenvattende conclusies uit het kalibratieonderzoek

Hieronder formuleren we een aantal conclusies die uit de analyses van de data van het kalibratieonderzoek kunnen worden getrokken.

- De items van het kalibratie-onderzoek laten toe om de taalvaardigheid Nederlands van 5-jarige kleuters op een betrouwbare en valide manier in kaart te brengen.
- Het is mogelijk om valide en betrouwbaar te meten met dit instrument, zeker bij middel- tot laagtaalvaardige leerlingen. Voor een goed beeld van sommige kleuters is inbedding in een bredere beeldvorming belangrijk. Ook hier: hoe beter de afnamecondities, hoe betrouwbaarder het resultaat (zie punt 5).
- Het is op basis van het instrument mogelijk om laagtaalvaardige kleuters te detecteren. Het is tevens mogelijk om met behulp van het taalscreeningsinstrument te differentiëren tussen kleuters met een verschillend niveau van taalvaardigheid.
- De taalscreening heeft een plafondeffect: Het is op basis van de items in het instrument niet mogelijk om genuanceerde uitspraken te doen over het taalvaardigheidsniveau van hoogtaalvaardige kleuters. Dat is echter niet de bedoeling van het instrument.
- De screening afnemen is haalbaar, maar de 'afnamecondities' zijn een belangrijk aandachtspunt. Een implementatietraject kan zinvol zijn. Daarin kunnen o.a. de setting voor de afname en de voorbereiding en coaching van de toetsafnemers aan bod komen.
- De toetstaken op zich zijn goed en werken goed (ecologische validiteit). Gebruikers zijn positief over de tekeningen, de verschillende types taken, de hedendaagse context en afbeeldingen, de aansluiting bij de kleuterleeftijd… Zowel de papieren als de digitale versie leveren valide en betrouwbare resultaten op.
- De inschatting van leraren en van het instrument komen sterk overeen. Dat geldt zeker voor meertalige kleuters. Voor veel kleuters is de taalscreening dus een bevestiging (maar dat vinden leraren fijn). Toch is deze overeenstemming er niet voor sommige kleuters; Deze discrepantie kan een uitnodiging vormen om verder, breder, dieper te kijken).
- Uit de data kunnen we afleiden dat een vorm van adaptieve toetsing mogelijk en zelfs zinvol zou kunnen zijn: dit zou frustratie bij leraren en leerlingen vermijden, het is efficiënt (in die zin dat het tijdswinst kan opleveren) en levert zinvolle output aan.
- Analyses wijzen uit dat verschillen in leeftijd, in socio-economische gezinssituatie, in de mate waarin kleuters worden blootgesteld aan schooltaal, in de mate waarin ze Nederlands aan het verwerven zijn elementen zijn die meespelen in het vaardigheidsniveau van een kleuter. Dit zijn echter factoren die een kleuter - noch zijn ouders - onder controle hebben. Om die reden mogen de resultaten uit de taalscreening niet gebruikt worden om kleuters of hun leraren kansen tot verdere ontwikkeling te ontnemen. Wanneer de taalscreening gebruikt wordt als bedoeld (m.a.w. als knipperlichtfunctie) vormt dit geen probleem.
- De aangeleverde output moet leiden tot 'iets doen met de taalscreening'. Dat kan alleen als uit de afname een realistisch en volledig beeld van de resultaten wordt meegegeven. In dit beeld

moeten zeker elementen worden meegenomen als gezinstaal en opleidingsniveau, de behaalde verschillende doelstellingen, de verschillende typetaken…

- Gebruikers vinden het belangrijk concrete tips voor de klaspraktijk te krijgen om aan de slag te kunnen met de resultaten van de taalscreening. Daarnaast vragen ze extra middelen en willen ze graag adviezen om ouders te betrekken.

HOOFDSTUK 7: CESUREN BIJ KOALA

1 ▪ Cesuurbepaling met cesuurcommissie

Voor het bepalen van de cesuren werd gebruik gemaakt van de Bookmarkmethode (Janssen e.a., 2003; 2004; Cizek & Bunch, 2006) bij een groep van beoordelaars. Bij de Bookmarkmethode krijgen de beoordelaars de toetsitems in stijgende graad van moeilijkheid voorgelegd. De taak van de beoordelaars bestaat erin om (figuurlijk gesproken) een bladwijzer te plaatsen tussen de twee items die de overgang vormen tussen items die wel beheerst moeten worden en items die nog niet beheerst moeten worden. Dat doen ze in verschillende rondes waarin ze geleidelijk naar een (minstens redelijk) gemeenschappelijk standpunt toe groeien.

Alle methoden voor cesuurbepaling, dus ook de Bookmarkmethode, steunen op het gecombineerde oordeel van een groep van deskundigen. Deze deskundigen kunnen uit meerdere domeinen afkomstig zijn. In een cesuurcommissie zitten daarom niet alleen leraren, maar ook andere deskundigen die belang hebben bij de uitkomst van de cesuurbepaling en impact van de cesuur. Het gaat erom dat er vanuit verschillende invalshoeken gekeken wordt naar de cesuur en de impact daarvan op kind- en schoolniveau.

Om dit proces goed te doorlopen moet de cesuurcommissie voldoende groot zijn (maar niet té groot zodat dialoog mogelijk is in de groep) en een brede vertegenwoordiging van het onderwijsveld bevatten. Wij kozen voor een cesuurcommissie van maximaal 20 personen met daarin personeelsleden van scholen, onderwijsondersteuners, lerarenopleiders, vertegenwoordigers van het beleid en onderwijskwaliteit, en onderzoekers. Alle aangeschreven personen hebben de nodige expertise als het gaat over kleuters en kleuteronderwijs.

De initiële cesuurcommissie telt 19 leden en is als volgt samengesteld. De namenlijst staat in Bijlage 5: leden van de cesuurcommissie.

Personeelsleden van scholen	3 leraren kleuteronderwijs 2 zorgcoördinatoren basisonderwijs, 1 pedagogisch directeur van een basisschool 1 directeur van een kleuterschool
Onderwijsondersteuners	3 pedagogisch begeleiders/adviseurs kleuteronderwijs 1 vormer met expertise voorschools, kleuter en lager
Lerarenopleiders	2 lerarenopleiders kleuteronderwijs[9] 1 lerarenopleider zorg

[9] Eén lerarenopleider kleuter heeft afgehaakt na de eerste online samenkomst. Deze persoon werd tijdens de tweede online samenkomst vervangen door een beleidsondersteuner.

Vertegenwoordigers beleid en onderwijskwaliteit	2 leden van de inspectie (werkgroep kleuter) 1 lid van AHOVOKS met expertise basisonderwijs
Onderzoekers	1 onderzoeker met expertise 'meertaligheid' en 'evaluatie' 1 onderzoeker met expertise 'denkontwikkeling' bij kleuters

Tabel 55: Overzicht leden cesuurcommissie

Na de eerste online samenkomst met de cesuurcommissie heeft één lerarenopleider afgehaakt. We hebben dit in eerste instantie proberen om te vangen door een extra persoon uit het werkveld te rekruteren. We hebben een taalcoach van een scholengemeenschap aangeschreven met de vraag of één van hun personeelsleden beschikbaar was om deel te nemen aan de cesuurbepaling. Dit was helaas niet het geval. In tweede instantie hebben we een beleidsondersteuner van een kleuterschool aangesproken. Daarop kregen we een positieve reactie.

2 ▪ Procedure

2.1 Contact en communicatie met de cesuurcommissie

Nadat de leden van de cesuurcommissie bevestigend hadden geantwoord op de algemene uitnodigingsmail, ontvingen zij de nodige informatie over het verloop van de cesuurbepaling.

De cesuurbepaling volgens de Bookmarkmethode volgt de volgende procedure:

- Kadering van de taalscreening, het opzet, het instrument.
- Inzage in het instrument en de items.
- Sortering van de items van gemakkelijk naar moeilijk op basis van de 'measure'.
- Cesuur plaatsen in verschillende rondes > Experten plaatsen een 'bookmark' (bladwijzer): ze geven m.a.w. aan welke opgave nog wel en welke niet meer beheerst hoeft te worden door een kleuter wiens taalontwikkeling vlot verloopt en geen extra ondersteuning nodig heeft. Deze individuele plaatsing van een cesuur wordt gevolgd door discussie in grote of kleine groep.
- Finale cesuur plaatsen.

Omwille van de coronamaatregelen vonden de samenkomsten online plaats (via Microsoft Teams). Dit had als voordeel dat de drempel om deel te nemen lager was voor personen die anders een (grote) verplaatsing zouden moeten maken. Een nadeel is dat de kennismaking met elkaar minder spontaan verloopt en dat het moeilijker is voor de leden om een goed beeld te krijgen van de (praktijk)ervaring en expertise van de anderen. Ook informele uitwisselingsmomenten, bijvoorbeeld tijdens pauzes, vallen hierdoor weg. Toch hebben we ervaren dat het ook via de online weg mogelijk is om een degelijke cesuurbepaling te doen. De leden werden aangemoedigd om zowel via mondelinge tussenkomsten als via de chat hun indrukken te delen en vragen te stellen. Na afloop van de online samenkomst konden ze nog steeds via e-mail vragen stellen.

2.2 Selectie van items voor de cesuurbepaling

Voor de berekening van de item-measures op basis waarvan de itemselectie voor de cesuurbepaling gebeurt, werd enkel rekening gehouden met:

- betrouwbare items, d.w.z. geen items met hoge misfit waarden. Items met een MNSQ > 2 werden niet meegenomen in de berekening van de measures. Het gaat om 5 items. We laten deze items uit de verdere analyses en gaan verder met de 143 overblijvende betrouwbare items.

- kleuters met een betrouwbare meting, d.w.z. geen kleuters met hoge misfit waarden. De grens ligt hier op een outfit MNSQ-waarde >3[10]. Het gaat om kleuters die inconsistent antwoorden, bijvoorbeeld correcte antwoorden geven op moeilijke items maar opeens de mist in gaan bij gemakkelijke items. Met deze kleuters werd geen rekening gehouden bij de berekening van de measures voor de cesuurbepaling. Het gaat om 21 kleuters (1,1%). We namen deze kleuters wel mee in de analyses voor de rapportering omdat zij deel uitmaken van de werkelijke populatie en er ook rekening met hen gehouden moet worden wanneer scholen gebruik maken van het instrument.

In totaal bestaat de screening uit 143 betrouwbare items en 30 taken. Voor de cesuurbepaling werden 43 items geselecteerd uit 23 taken, die goed gespreid zijn over de schaal. Deze 43 items staan gerangschikt van gemakkelijk en naar moeilijk in een digitale 'booklet' voor de leden van de cesuurcommissie. Zij nemen deze booklet verschillende keren door en bepalen vervolgens de cesuur door een virtuele 'bookmark' (bladwijzer) te leggen tussen twee items.

Selectiecriteria voor de items:

- Het verschil in moeilijkheid moet voldoende duidelijk zijn: we zorgen daarom voor een voldoende grote en zo gelijk mogelijke afstand in measure tussen de geselecteerde items.
- Discriminerende waarde van de items: we geven de voorkeur aan items met een grote discriminerende waarde (>1), d.w.z. items die een goed onderscheid maken tussen 'sterke' en 'zwakke' presteerders.
- Spreiding in moeilijkheid van items binnen een taak: waar mogelijk selecteren we items die deel uitmaken van eenzelfde taak, maar duidelijk verschillen in moeilijkheidsgraad. Dit geeft de leden van de cesuurcommissie een goed beeld van hoe binnen eenzelfde context zowel moeilijke als gemakkelijke items aangeboden worden.
- Spreiding over verschillende doelstellingen en typetaken: we waken erover dat er in de selectie voldoende items zitten voor elke doelstelling en typetaak (zie Tabel 56: Spreiding van items over doelstellingen voor booklet cesuurbepaling en Tabel 57: Spreiding van items over typetaak voor booklet cesuurbepaling hieronder).
- Feedback vanuit de scholen: we vermijden de selectie van items waarbij de scholen tijdens het kalibratieonderzoek (veel) twijfels of bedenkingen hebben geuit.
- DIF-contrasten: we hebben gekeken of de items verschillen in moeilijkheidsgraad naargelang het afnametype (papier of digitaal), de gezinstaal (Nederlands of niet-Nederlands) en opleiding van de moeder (hoog- of laagopgeleid). Items waarbij er sprake is van een significant DIF-contrast voor één van deze drie kenmerken werden niet opgenomen in de selectie. Dit deden we om de vergelijkbaarheid van de items over alle condities en populaties te bewaren.

[10] We wilden enkel de kleuters eruit halen die zeer inconsistent hebben geantwoord (de echte outliers). Daarom hebben we hier de minder strenge grenswaarde (outfit MNSQ >3) gehanteerd.

Tabel 56 geeft een overzicht van de verschillende doelstellingen, zoals deze opgenomen werden in de booklet van de cesuurbepaling.

Doelstelling	Aantal items in selectie	Aantal items in screening	% in selectie	% in screening
Instructies begrijpen	13	35	30,2%	24,4%
Informatieve mededelingen begrijpen	11	43	25,6%	30,1%
Vragen begrijpen	11	44	25,6%	30,8%
Verhalen begrijpen	8	21	18,6%	14,7%

Tabel 56: Spreiding van items over doelstellingen voor booklet cesuurbepaling

De spreiding over de verschillende doelstellingen komt overeen met de zwaartes zoals deze opgenomen waren in de toetsmatrijs bij deze toets (zie 2.2. TOETSMATRIJS).

Dit is ook zo voor de verschillende typetaken, zoals blijkt uit het overzicht in Tabel 57: Spreiding van items over typetaak voor booklet cesuurbepaling.

Typetaak	Aantal items in selectie	Aantal items in screening	% in selectie	% in screening
Gestandaardiseerde observatie	15	46	34,8%	32,2%
Doe- en zoek-opdracht	11	36	25,6%	25,2%
Meerkeuze-opdracht	17	61	39,6%	42,7%

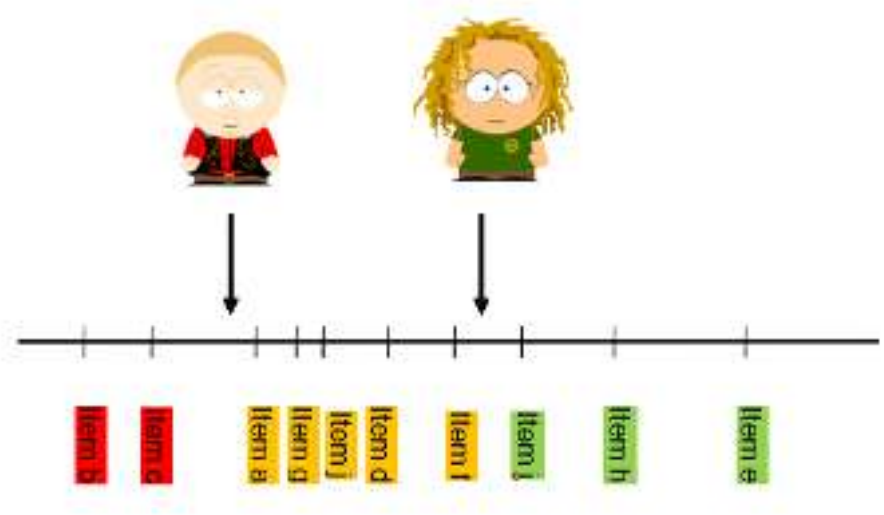
Tabel 57: Spreiding van items over typetaak voor booklet cesuurbepaling

2.3 Twee cesuren, drie groepen

Na een verkennende analyse namen we de beslissing om twee cesuren te bepalen. De item-measures kennen een normaalverdeling met breed plateau. De items concentreren zich dus niet rond één measure, maar verspreiden zich goed over een voldoende breedte op de moeilijkheidsschaal. Dit maakt het mogelijk om (net zoals bij SALTO) twee cesuren vast te leggen,

waarbij deze zich wellicht (eerder) aan de bovenkant en (eerder) aan de onderkant van dit plateau zullen situeren. Door twee cesuren vast te leggen ontstaan er drie groepen kleuters.

- Cesuur A maakt onderscheid tussen enerzijds kleuters die voldoende hebben aan de basisaanpak van de leraar (in de groene zone) en anderzijds kleuters voor wie de leraar alert tot zeer alert moet zijn (in de oranje of rode zone).
- Cesuur B maakt onderscheid tussen enerzijds kleuters waarvoor een leraar zeer alert moet zijn, die zeer waarschijnlijk intensieve extra ondersteuning nodig hebben (in de rode zone) en anderzijds kleuters waarvoor waarschijnlijk extra ondersteuning of de basisaanpak van de leraar volstaat (in de oranje respectievelijk de groene zone).



Figuur 18: Visuele voorstelling van de groepen en cesuren

Deze cesuren werden als volgt gepresenteerd aan de commissieleden:

Groene zone	Boven cesuur A
•Doe zo verder: deze kleuters kunnen met jouw basisaanpak voldoende evolutie doormaken op vlak van taalontwikkeling doorheen de derde kleuterklas.	

Oranje zone	Tussen cesuur A en B
•Wees alert: deze kleuters hebben mogelijk extra ondersteuningsmaatregelen nodig om voldoende evolutie door te maken op vlak van taalontwikkeling in de derde kleuterklas.	

Rode zone	Onder cesuur B
•Wees zeer alert: deze kleuters hebben waarschijnlijk intensieve extra ondersteuningsmaatregelen nodig om voldoende evolutie door te maken op vlak van taalontwikkeling in de derde kleuterklas.	

Figuur 19: Beschrijving van de groepen en cesuren

3 ▪ Cesuurbepaling Dag 1

De eerste online samenkomst vond plaats op woensdag 31 maart 2021. De bijeenkomst bestond uit twee delen.

3.1 Deel 1: In groep

De onderzoekers schetsen de context, en geven informatie over de achtergrond en ontwikkeling van het taalscreeningsinstrument. Ze gaan in op de kansen en valkuilen van een taalscreening bij kleuters.

De onderzoekers geven ook toelichting bij de opbouw van de screening, de afnamecondities, de verschillende typetaken en verschillende doelstellingen die aan bod komen in het instrument. Ze situeren de screening binnen het kader van brede beeldvorming.

De onderzoekers presenteren vervolgens de twee vragen die het vertrekpunt vormen voor het bepalen van de twee cesuren. De deelnemers kijken in detail naar de vraagformulering en gaan in op de overeenkomsten en verschilpunten tussen de twee vragen. De onderzoekers verwijzen naar de metafoor 'hoogspringen': we zoeken naar het moment (het item) waarbij de laat gaat trillen maar de kleuter er nog net over geraakt.

Vraagformulering CESUUR A

Neem een kleuter voor ogen die mogelijk extra ondersteuningsmaatregelen nodig heeft om voldoende evolutie door te maken op vlak van taalontwikkeling in de derde kleuterklas.

Wat is het moeilijkste item dat de beste kleuter met dit profiel nog correct kan beantwoorden?

- Kleuters die boven deze cesuur liggen, zitten in de groene zone
- Kleuters die onder deze cesuur liggen, zitten in de oranje of rode zone

Vraagformulering CESUUR B

Neem een kleuter voor ogen die waarschijnlijk intensieve extra ondersteuningsmaatregelen nodig heeft om voldoende evolutie door te maken op vlak van taalontwikkeling in de derde kleuterklas.

Wat is het moeilijkste item dat de beste kleuter met dit profiel nog correct kan beantwoorden?

- Kleuters die boven deze cesuur liggen, zitten in de oranje of groene zone
- Kleuters die onder deze cesuur liggen, zitten in de rode zone

De onderzoekers presenteren een aantal belangrijke bevindingen uit het kalibratie-onderzoek. Deze inzichten kunnen helpen bij het maken van afwegingen tijdens de cesuurbepaling.

De onderzoekers lichten de individuele opdracht toe: uitleg bij de selectiecriteria voor de items en voorstelling van de booklet met 43 geselecteerde en gerangschikte items.

3.2 Deel 2: Individueel

Vervolgens kregen de commissieleden de mogelijkheid om individueel een bookmark te plaatsen en zich hierop voor te bereiden.

3.2.1 Individuele familiarisatie

De deelnemers maken zich vertrouwd met de geselecteerde items en de taken waaruit deze afkomstig zijn. De focus ligt op kennismaking met het instrument en de items. Er zijn in totaal 23 taken waaruit item geselecteerd werden. Dit is een behoorlijk aantal taken en om ballast te vermijden, hebben we ervoor gekozen om de volledige context van de taak niet mee te geven wanneer deze zeer eenvoudig is en voor zich spreekt. Voor bepaalde items is het wel noodzakelijk om zicht te hebben op de context waarin deze zijn ingebed. In dat geval wordt de volledige context wel meegegeven in de booklet.

3.2.2 Mogelijkheid tot vragen stellen

Het KOALA-team blijft beschikbaar in de online ruimte voor vragen over de items en taken en/of terugkoppeling over de informatie die gegeven werd tijdens deel 1. Deelnemers die ervoor kiezen om op een later tijdstip hun bookmarks te leggen, kunnen hun vragen via e-mail sturen. We kregen een aantal vragen binnen rond de context van taken. Over de concrete doe-opdrachten kwamen ook een enkele vragen rond beoordelingscriteria.

3.2.3 Twee individuele ‘bookmarks’ of cesuren leggen

De deelnemers kunnen de cesuren meteen leggen, of ervoor kiezen om de individuele opdracht op een later tijdstip uit te voeren. In dat geval is het van belang om de nodige notities te nemen. Er is hiervoor plaats voorzien in de booklet. De commissieleden geven hun cesuren door via een online formulier.

4 ▪ Cesuurbepaling Dag 2

De tweede online samenkomst vond plaats op woensdag 21 april van 13u tot 17u.

4.1 Deel 1: terugkoppeling na ronde 1

4.1.1 Korte opfrissing context en vraagstelling

De onderzoekers herhalen kort wat de bedoeling is van de cesuurbepaling en kijken opnieuw naar de belangrijkste contextuele elementen die de cesuurcommissie goed in het achterhoofd moet houden bij het leggen van de cesuur:

- Steekproef kalibratieonderzoek: oververtegenwoordiging van kleuters die aantikken op OKI-indicatoren. Gemiddelde OKI-score in steekproef is 1,82 (tegenover 1,10 in ganse populatie). Op deze manier wilden we voldoende 'risicokleuters' laten deelnemen aan het kalibratieonderzoek.
- Brede beeldvorming: de screening staat niet op zichzelf en is één element dat leraren helpt om een totaalbeeld te krijgen van de ontwikkeling van een kleuter.
- Valkuilen en kansen: de cesuren gaan over het scheppen van kansen voor kleuters. We hopen dat de scores positief worden benaderd en dat het voor extra taalverwervingskansen kan zorgen. We mogen daarbij uiteraard niet blind zijn voor de valkuilen (o.a. teaching to the test, labelen van kleuters en screening als 'toegangsticket' gebruiken voor het 1e leerjaar).
- Vraagformulering: we kijken opnieuw naar de vraagformuleringen die het vertrekpunt vormen voor het leggen van cesuur A en cesuur B, en de kenmerken van de drie groepen die vervolgens ontstaan. Enkele leden van de cesuurcommissie vonden het moeilijk om twee cesuren te leggen en/of vragen zich af waarom het twee cesuren moeten zijn, bijvoorbeeld omdat dit de deur kan openen voor het installeren van niveaugroepen in het onderwijs. Naar analogie met SALTO werd in overleg met de opdrachtgever bepaald dat we ook voor KOALA twee cesuren leggen. We geven aan dat er ook bij één cesuur niveaugroepen kunnen ontstaan. Alles hangt af van de manier waarop de school met de resultaten aan de slag gaat, maar dat is iets wat buiten de controle van deze cesuurcommissie ligt.

4.1.2 Bedenkingen en reflecties bij het plaatsen van de cesuren in de eerste ronde (op dag 1)

Via Mentimeter peilen de onderzoekers naar de bedenkingen en reflecties die naar boven zijn gekomen tijdens het leggen van de eerste individuele cesuur. We doen dit aan de hand van twee vragen. Nadat de deelnemers hun input hebben gegeven, volgt een plenaire discussie gemodereerd door een onderzoeker.

In de eerste vraag ligt de focus op overwegingen die een rol hebben gespeeld bij het leggen van de cesuren. We vragen aan de deelnemers om dit in drie kernwoorden weer te geven.

Vraag 1: *Wat waren voor jou belangrijke overwegingen om cesuur A en cesuur B op die plek te leggen? Wat zijn volgens jou belangrijke elementen om mee te nemen naar de volgende rondes?*

Figuur 20: Resultaten menti-meter bij vraag 1 ingevuld door de leden van de cesuurcommissie

Het valt op dat 'context' vooraan staat bij een aantal mensen:

- Voor sommigen gaat dit over de context van de vraag of taak, m.a.w. wat een kleuter ziet en hoort in aanloop naar de vraagstelling. Het gaat dus over een aspect van de screening. De context die een kleuter meekrijgt is belangrijk om te kunnen inschatten of het item moeilijk of gemakkelijk is.
- Voor anderen gaat context eerder over het al dan niet aansluiten van de vraag of taak bij de belevingswereld van de kleuters en over achtergrondkennis die al dan niet een rol kan spelen. De meeste commissieleden vinden dat er veel herkenbare elementen zitten in de taken die aansluiten bij wat kleuters dagdagelijks tegenkomen en ervaren in de klas en op school. Enkelen geven aan dat de taken waarbij de natuur en het platteland een belangrijke rol spelen, minder herkenbaar zijn voor kleuters die opgroeien in een grootstad.
- Voor kleuters die enkel op school Nederlands horen, zijn items waarin (voornamelijk) schooltaal aan bod komt veel eenvoudiger. Schooltaal in een item kan het gemakkelijker maken, omdat ze deze woorden dan herkennen (bv. 'in de rij gaan staan').

We gaan even dieper in op de rol van concentratie, aandacht en werkgeheugen:

- Verschillende leden geven aan dat de score van een kleuter (sterk) beïnvloed kan worden door de mate van concentratie. Taalvaardige kleuters die zich moeilijk kunnen concentreren, zullen niet goed scoren bij taken die wat langer duren en/of waarbij veel informatie moet verwerkt worden. Dit illustreert het belang van brede beeldvorming: de resultaten op de screening moeten altijd in verband gebracht worden met andere observaties.

- Binnen eenzelfde (verhaal)context krijgt de kleuter verschillende vragen (of items) voorgelegd. Als dingen een aantal keer terugkomen, wordt er ook beroep gedaan op het leervermogen van de kleuter. Sommigen vragen zich af of items die later aan bod komen in een bepaalde taak gemakkelijker zijn dan de eerste items, omdat kleuters dan al meer vertrouwd zijn met de taak en de context ervan. Daar zijn vanuit de analyses geen aanwijzingen voor.

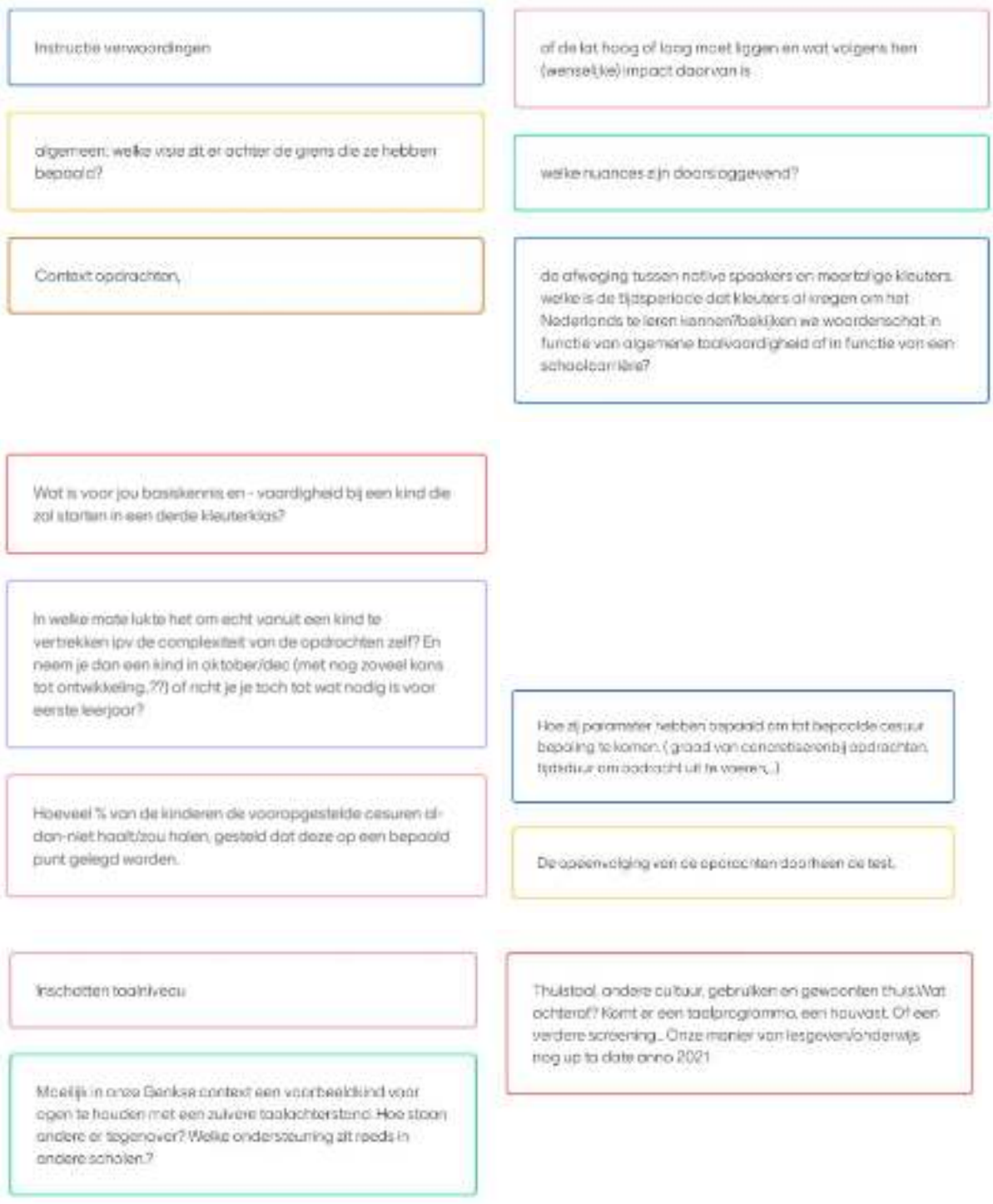
De afnamecondities hebben ook een rol gespeeld tijdens het leggen van de cesuur:

- Kleuters mogen vragen om een vraag te herhalen. Als dat gebeurt, kan dit wel een verschil maken in hoe moeilijk een vraag is voor de kleuter. Het zou zinvol zijn om deze instructie (‘je mag vragen aan de juf/meester om het nog eens voor te lezen’) vooraf te geven aan de kleuters; zo weten ze dat het oké is om te vragen om iets te herhalen en gaan ze dat misschien ook sneller doen.

- Zeker bij de concrete doe-opdrachten is de beoordeling niet altijd zwart-wit. Soms stelt de kleuter een handeling die gedeeltelijk juist is. Tijdens het kalibratieonderzoek werd hiervoor een aparte categorie voorzien in de beoordeling. Maar voor de cesuurbepaling werd enkel rekening gehouden met een volledig correcte reactie of handeling.

In de tweede vraag ligt de focus op de uitwisseling van gedachten, overwegingen en eventuele twijfels. Er zijn wellicht verschillende overwegingen die hebben meegespeeld bij het leggen van de cesuren en we willen weten wat de deelnemers graag willen aftoetsen bij de groep. We leggen de deelnemers hiervoor een open vraag voor.

Vraag 2: *Ik ben benieuwd wat anderen denken over…*



Figuur 21: Resultaten menti-meter bij vraag 2 ingevuld door de leden van de cesuurcommissie

Tijdens de discussie focussen de deelnemers op de vraag 'van wat voor kind vertrek je om de cesuur te bepalen?':

- Sommige deelnemers hebben tijdens de eerste cesuur voornamelijk gekeken naar de algemene taalvaardigheid, anderen hebben het accent meer gelegd op schoolse

taalvaardigheid. Een commissielid geeft aan dat hij geprobeerd heeft om te focussen op wat echt relevant is voor kleuters in een derde kleuterklas, en dat schooltaal daarin een belangrijke rol speelt, zonder daarom voorbij te gaan aan de leefwereld (thuissituatie) van de kleuters. Voor stadskinderen kan dit anders zijn dan voor kinderen in een landelijke omgeving.

- Het was voor sommigen niet gemakkelijk om tijdens het leggen van de cesuur een bepaalde kleuter voor ogen te blijven houden. Ze hadden het gevoel dat ze het beeld van deze kleuter niet konden vasthouden, waardoor ze meer vanuit de moeilijkheidsgraad van de items dan vanuit de kleuter vertrokken om de cesuur te leggen.

- Een commissielid geeft aan dat het haar geholpen heeft om voor elke groep (cesuur A en cesuur B) een specifieke kleuter voor ogen te houden. Ze heeft geprobeerd om echt in de schoenen van de kleuter te gaan staan en zichzelf (als leraar) aan de zijlijn te houden.

- Kleuters evolueren erg snel. Gaat het om een kind dat in oktober in de derde kleuterklas zit? Of een kind dat op de drempel naar het eerste leerjaar staat? Kinderen lopen op die jonge leeftijd cognitief sterk uiteen. Ook dat maakte het moeilijk.

De deelnemers informeren bij elkaar naar de gepercipieerde moeilijkheidsgraad van de taken.

- Heel wat commissieleden hebben een conflict ervaren tussen de voorgelegde volgorde van de items en de inschattingen die ze zelf maken. Verschillende deelnemers geven aan dat ze zich niet helemaal konden vinden in de vastgelegde moeilijkheidsgraad van de items. Soms vonden ze een 'moeilijke' vraag eerder gemakkelijk, of omgekeerd.

- Onderzoek geeft aan dat zowel onderzoekers als leraren moeilijk kunnen voorspellen of een item moeilijk is of niet voor een kleuter.

Alvorens de impactdata voorgesteld worden, geven we nog een aantal bijkomende inzichten mee uit het kalibratieonderzoek:

- Er zijn 9,8% extreem goede presteerders (>= 95% correcte antwoorden). De groep extreem zwakke presteerders (<20% correcte antwoorden) is zeer klein: 1,2%.

- Wie zijn de 'onbetrouwbare kleuters', d.w.z. kleuters die inconsistent antwoorden? Er is geen samenhang met leeftijd. Deze kleuters hebben vaker een hoogopgeleide moeder en spreken vaker enkel Nederlands thuis. Het gaat om een beperkte groep[11].

- Verschillende factoren hebben een invloed op de measure. De invloed van gezinstaal is het grootst, daarna volgt de opleiding van de moeder. Leeftijd heeft de minst sterke impact.

- Interactie-effect tussen opleiding van de moeder en gezinstaal: het positieve effect van de opleiding van de moeder (hogere measure als moeder hoogopgeleid is), is nog sterker in gezinnen waarin enkel Nederlands gesproken wordt.

- Geen interactie-effect tussen leeftijd en gezinstaal: het effect van leeftijd op de measure verschilt dus niet naargelang de taal die thuis gesproken wordt.

[11] Als we de een minder strenge grenswaarde hanteren (outfit MNSQ > 2) gaat het om 3,5% van de kleuters (N=69). Als we een strenge grenswaarde hanteren (outfit MNSQ > 3) gaat het om 1,1% van de kleuters (N=21).

- Er is een negatief lineair verband tussen leeftijd (in maanden) en de measure.

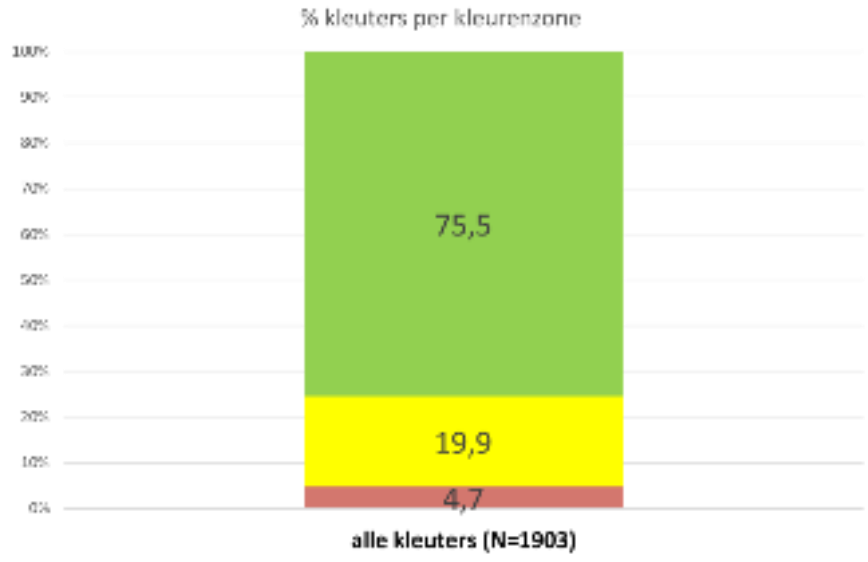
4.1.3 Presentatie cesuren na de eerste ronde

Cesuur A		Cesuur B		
29	0,520	15	-0,720	mediaan
32,25	0,800	21	-0,210	p75
25,75	0,215	14	-0,820	p25
16	-0,630	4	-1,800	min
35	1,110	22	-0,110	max

Tabel 58: Resultaten cesuurbepaling ronde 1

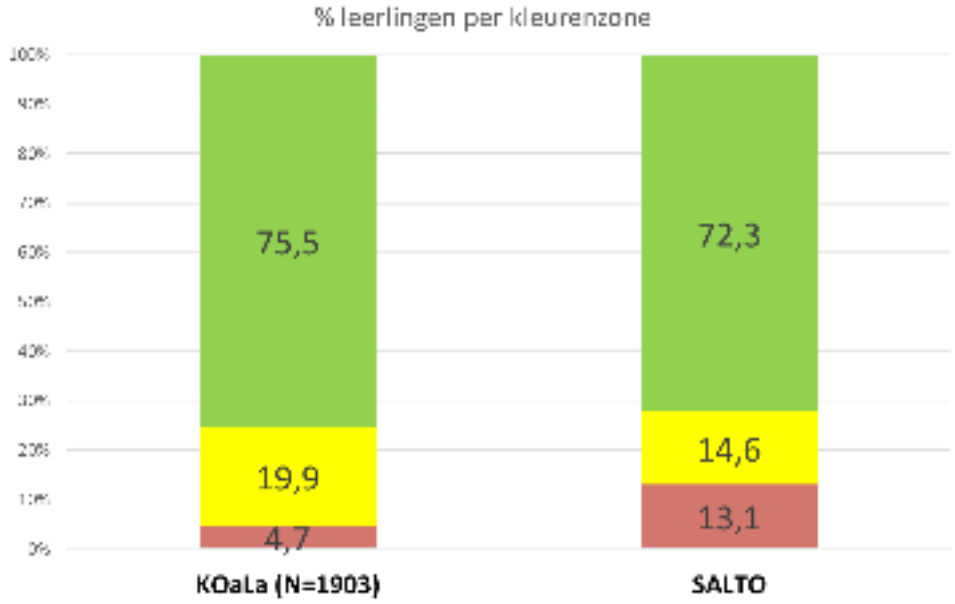
- Cesuur A ligt op item 29 van 43. Cesuur B op item 15 van 43.
- De afstand tussen p25 en p75 (= middelste helft van de deelnemers) is ongeveer even groot voor beide cesuren, met name een afstand van ongeveer 7 items. De afstand tussen de measures is voor beide cesuren ook gelijkaardig (ongeveer 0,60).
- Er zijn voor beide cesuren enkele uitschieters naar onder: item 16 voor cesuur A en item 4 voor cesuur B.

4.1.4 Presentatie van de impactdata op kindniveau na de eerste ronde



Figuur 22: Impactdata op kindniveau cesuurbepaling ronde 1

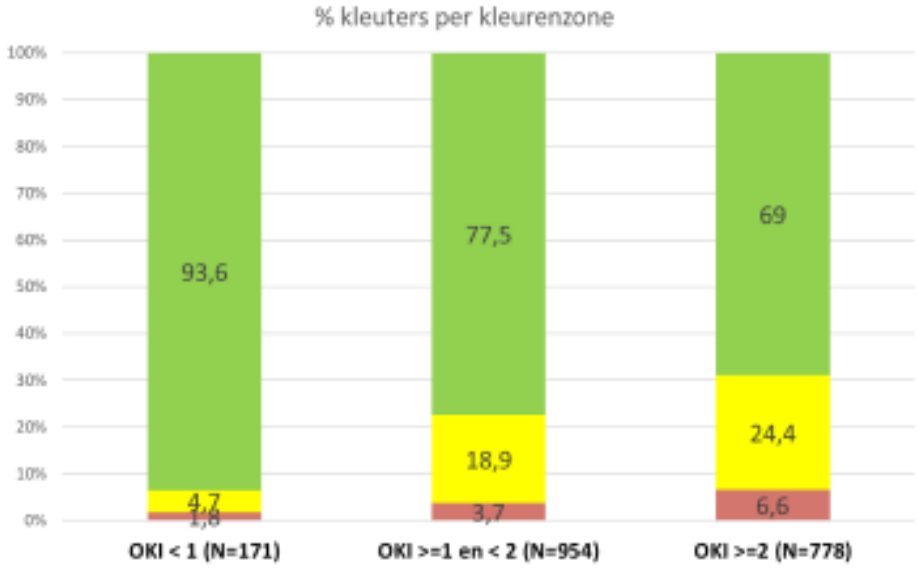
We zien een grote groene zone en heel kleine rode zone. 75,5% van de kleuters zit boven cesuur A. 4,7% van de kleuters zit onder cesuur B.



Figuur 23: Impactdata op kindniveau cesuurbepaling ronde 1 in vergelijking met SALTO

We zien dat de groene zone bij SALTO gelijkaardig is. Maar de rode zone is bij SALTO een stuk groter.

4.1.5 Presentatie van de impactdata op schoolniveau (OKI-categorie van de school) na de eerste ronde



Figuur 24: Impactdata op schoolniveau cesuurbepaling ronde 1

Er is een duidelijk verschil tussen de OKI-categorieën. Scholen met meer indicatorleerlingen hebben een kleinere groene zone. Vooral de gele zone neemt sterk toe bij een OKI-waarde >=1. De rode zone blijft klein, ook bij scholen met gemiddelde OKI >=2.

4.1.6 Plenaire discussie over cesuren van de eerste ronde en de bijhorende impactdata

In deze discussie ligt de focus op de betekenis van de impactdata. We vragen eerst aan de leden om voor zichzelf enkele notities te maken aan de hand van volgende richtvragen:

- *Vind je de plaats van de cesuren verrassend? Waarom?*
- *Zijn de impactdata voor jou een verrassing? Waarom?*
- *Vind je de verschillen tussen de OKI-categorieën opvallend? Helpen deze gegevens scholen om een goede inschatting te maken van waar hun kleuters staan? Waarom wel/niet?*

Vervolgens vragen we hen om op basis van deze notities drie kernwoorden in te geven via Mentimeter. Nadat de deelnemers hun input hebben gegeven, volgt een plenaire discussie gemodereerd door een onderzoeker.

Vraag: *Selecteer o.b.v. je notities drie kernwoorden die jouw redenering of (achterliggende) gedachten bij de betekenis van deze impactdata goed weergeven.*

Figuur 25: Resultaten menti-meter dag 2 ingevuld door leden van de cesuurcommissie

Heel wat commissieleden geven aan ze de impactdata niet verrassend vinden. Toch wijken sommigen daarvan af, zij geven aan dat de groene zone (te) groot is en/of de rode zone (te) klein.

We krijgen dus twee signalen: enerzijds is deze uitkomst niet verrassend, anderzijds weerklinken signalen dat de cesuren wat hoger mogen liggen. Tijdens de discussie is er ruimte om indrukken bij deze impactdata met elkaar te delen en vragen we naar mogelijke redenen voor de eerder lage cesuur.

- Een aantal commissieleden zijn werkzaam op een school met gemiddelde OKI >2. Enkele leraren geven aan dat de impactdata aansluiten bij de taalvaardigheidsinschatting die zij zouden maken van de kleuters op hun school. Ze hebben het gevoel dat de uitkomst van de eerste ronde overeenkomt met hun verwachtingen.
- Andere leraren zijn toch verrast. Zij hadden de lat een stuk hoger gelegd en kunnen zich niet helemaal vinden in de grote groene zone. Deze sluit volgens hen immers niet aan bij de realiteit. Naar hun gevoel zijn er heel wat meer kleuters die moeilijkheden ervaren met taalvaardigheid Nederlands. De lat mag voor hen een stuk hoger omdat ze het belangrijk vinden dat kleuters in de loop van de derde kleuterklas voldoende schooltaal verwerven omdat ze deze begrippen functioneel moeten kunnen inzetten in het eerste leerjaar (bv. begrippen i.v.m. ruimte en plaats).
- Bij de interpretatie van de resultaten is brede beeldvorming heel belangrijk. Een kleuter kan een laag taalniveau hebben, maar kan dat (gedeeltelijk) compenseren met een sterke werkhouding.

Verschillende commissieleden uiten hun bezorgdheid over de manier waarop de screening zal worden ingezet en welke acties of interventies eraan gekoppeld zullen worden. Sommige leraren geven het signaal dat zij zonder extra middelen niet nog meer extra ondersteuning kunnen geven aan kleuters in de oranje of rode zone in de dagelijkse klaspraktijk. Ze vragen zich bijgevolg af of er middelen gekoppeld zullen worden aan de resultaten van de kleuters, en op welke manier dat dan zal gebeuren.

4.2 Deel 2: Tweede ronde

De deelnemers krijgen 15 minuten de tijd om hun tweede individuele cesuur door te geven via het online invulformulier.

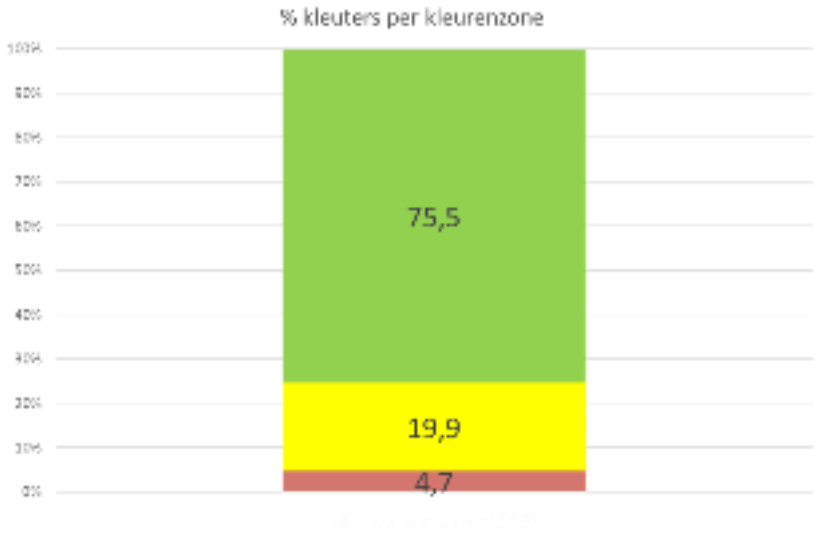
4.2.1 Presentatie van de cesuren na de tweede ronde

Cesuur A		Cesuur B		
29	**0,520**	15	**-0,720**	mediaan
32	0,780	17	-0,570	p75
26	0,240	14	-0,820	p25
19	-0,400	4	-1,800	min
36	1,170	22	-0,110	max

Figuur 26: Resultaten cesuurbepaling ronde 2

- De mediaanwaarden van beide cesuren zijn onveranderd gebleven (cesuur A op item 29, cesuur B op item 15).
- De afstand tussen p25 en p75 (= middelste helft van de deelnemers) is vooral bij cesuur B kleiner geworden. Dit weerspiegelt zich in een kleinere afstand tussen de measures (nog maar 0,25 bij cesuur B). Over cesuur B lijkt de eensgezindheid toegenomen.
- Voor cesuur A is het minimum opgeschoven naar boven (19 i.p.v. 16). Ook dit wijst op een toename in de eensgezindheid. Voor cesuur B is er nog steeds een sterke uitschieter naar onder.

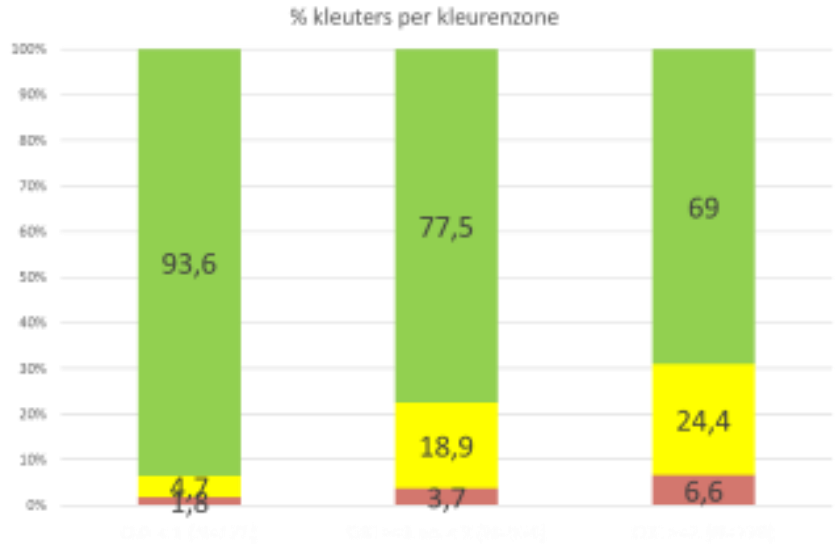
4.2.2 Presentatie van de impactdata op kindniveau na de tweede ronde



Figuur 27: impactdata op kindniveau cesuurbepaling ronde 2

De cesuren zijn gelijk gebleven, waardoor er geen verschuivingen zijn gebeurd in de impactdata. We zien nog steeds een grote groene zone en heel kleine rode zone. 75,5% van de kleuters zit boven cesuur A. 4,7% van de kleuters zit onder cesuur B.

4.2.3 Presentatie van de impactdata op schoolniveau (OKI-categorie van de school) na de tweede ronde



Figuur 28: Impactdata op schoolniveau cesuurbepaling ronde 2

De cesuren zijn gelijk gebleven, waardoor er ook op schoolniveau geen verschuivingen zijn gebeurd in de impactdata. Er is wel een duidelijk verschil tussen de OKI-categorieën. Scholen met meer indicatorleerlingen hebben een kleinere groene zone. Vooral de gele zone neemt sterk toe bij een OKI-waarde >=1. De rode zone blijft klein, ook bij scholen met gemiddelde OKI >=2.

4.2.4 Discussies in groepjes + plenaire terugkoppeling over de cesuren en impactdata van de tweede ronde

Groep 1

- De groep vindt het frappant dat er niets veranderd is. De groene zone blijft erg groot, de lat mag volgens hen hoger gelegd worden. Vooral cesuur B vindt deze groep erg laag. Ze delen de mening dat de rode zone groter mag zijn, vooral voor scholen met een OKI >2.
- Als je uitgaat van deze percentages, zou er slechts (ongeveer) één kleuter per klas in de rode zone zitten en ongeveer vijf in de gele zone. De deelnemers hebben de indruk dat dit in realiteit toch meer kleuters zijn. Deze screening heeft een signaalfunctie, maar met de huidige cesuur wordt er geen krachtig signaal gegeven.
- Er is bezorgdheid over de oranje zone. De deelnemers vrezen dat er (te) weinig aandacht zal gaan naar deze zone en dat er dan uiteindelijk niets gebeurt met deze groep kleuters. Ze verwijzen hierbij naar SALTO. Ook voor kleuters in de oranje zone is het noodzakelijk om extra taalstimulering te voorzien. Ze mogen niet over het hoofd gezien worden.

Groep 2

- De groep maakt zich de bedenking dat scholen op een heel verschillende manier een interpretatie kunnen geven aan de grote groene zone. Zo zullen er scholen zijn die al veel inspanningen hebben geleverd en zich zullen herkennen in deze cijfers. Maar wat dan met scholen die nog niet veel inspanningen hebben gedaan? Gaan zij met een grote groene zone nog wel geprikkeld worden om extra inspanningen te doen?

- De rode zone blijft verrassend klein. Scholen krijgen op die manier het signaal dat ze goed bezig zijn, en dat is weinig motiverend. Mogelijks heeft dit ook een impact op de middelen die een school al dan niet zou krijgen. Dat is volgens de groep geen correct en geen wenselijk signaal.

- De groep merkt op dat er een discrepantie is met de taalinschattingen van de leraren die deelnamen aan het kalibratieonderzoek. De cesuur sluit dan weer wel goed aan bij de inschattingen van (een aantal) leraren/praktijkmensen uit de cesuurcommissie.

Groep 3

- De meeste mensen lijken bij hun oorspronkelijke idee te blijven. Ze geven aan dat ze goed hebben nagedacht over hun cesuur en rekening hebben gehouden met veel factoren.

- Sommige groepsleden hadden toch graag een aantal opdrachten meer in detail willen bekijken (volledige context van de taak). De groep heeft de tijd genomen om dit te doen voor een aantal items.

- Er is ook in deze groep bezorgdheid over het signaal dat met deze cesuur gegeven wordt. Ze vragen zich af of je met zo'n grote groene zone wel voor voldoende alertheid zorgt. Daarnaast wordt de bedenking gemaakt dat deze indeling niet overeenstemt met het beeld dat leraren hebben.

Groep 4

- De groep vindt dat de lat zeer laag ligt en is hier bezorgd over. Het lijkt alsof we snel tevreden zijn, en dat kan toch niet de bedoeling zijn. Dat lat mag wat hen betreft dus een stuk hoger komen te liggen.

- Met deze cesuur geven we het signaal dat er weinig of geen problemen zijn, terwijl dat in realiteit wel zo is. De taalinschattingen door de leraren die deelnamen aan het kalibratie-onderzoek komen volgens deze groep beter overeen met de realiteit. De screening moet dienen om problemen te signaleren, maar met de huidige cesuur is dat niet het geval. De groep is van mening dat je op die manier de problemen doorschuift naar het eerste leerjaar.

- Brede beeldvorming blijft belangrijk: kleuters met concentratieproblemen zullen moeilijkheden ervaren met deze screening, maar dit wil nog niet meteen zeggen dat er een groot probleem is met hun taalvaardigheid, al kunnen concentratieproblemen ook weer een impact hebben op de taal.

- De screening is een momentopname: we weten niet hoe stabiel de resultaten zijn. Als de screening gebeurt aan het begin van het schooljaar, weten we niet waar de kleuter op het einde van het schooljaar zal staan. Dit is een oproep om altijd voorzichtig te zijn met de interpretatie van de resultaten.

Groep 5

- De groep vraagt zich af welk signaal we willen geven aan de scholen. Als de groene groep te groot is, kan dit tot gevolg hebben dat scholen zich weinig gemotiveerd voelen om (extra) in te zetten op taal. Als de rode groep te groot is, kan dit ook demotiverend zijn. De groep zou daarom graag te oranje groep wat groter zien. Ze zijn van mening dat een grote oranje groep een goede wake-up call kan zijn voor scholen.
- Belangrijk dat scholen realiseren dat ze ook bij een grote groene zone inspanningen moeten blijven doen op vlak van taal. Als school moet je voor de rest van het schooljaar nog altijd kwaliteitsvol (taal)onderwijs geven, zelfs al zitten (bijna) alle kleuters in de groene zone. Ook kleuters in de groene zone moeten voldoende kansen krijgen om te blijven groeien.
- Voor scholen is het belangrijk om te weten hoe ze verder ondersteund zullen worden in hun werking voor de kleuters in de verschillende kleurenzones.
- Je kan op verschillende manieren 'strenger zijn' en de lat hoger leggen. Je kan de rode zone groter maken. Je kan de lat ook hoger leggen door de oranje groep groter te maken. Maar je kan ook de bedenking maken dat met de huidige cesuur voor 24% van de kleuters een knipperlicht gaat branden dat aangeeft dat we alert moeten zijn. Zo klein is deze groep niet.
- Iemand maakt de opmerking dat het tijdens het kalibratieonderzoek opviel dat de meeste kleuters een behoorlijk goede prestatie leverden. Ook kleuters waarvan zij het niet verwacht had. Je mag kleuters niet onderschatten, de lat mag best hoog liggen.

5 ▪ Resultaat (Finale Cesuur)

De deelnemers krijgen opnieuw 15 minuten de tijd om hun derde en finale individuele cesuur door te geven via het online invulformulier.

5.1 Presentatie van de cesuren na de derde ronde

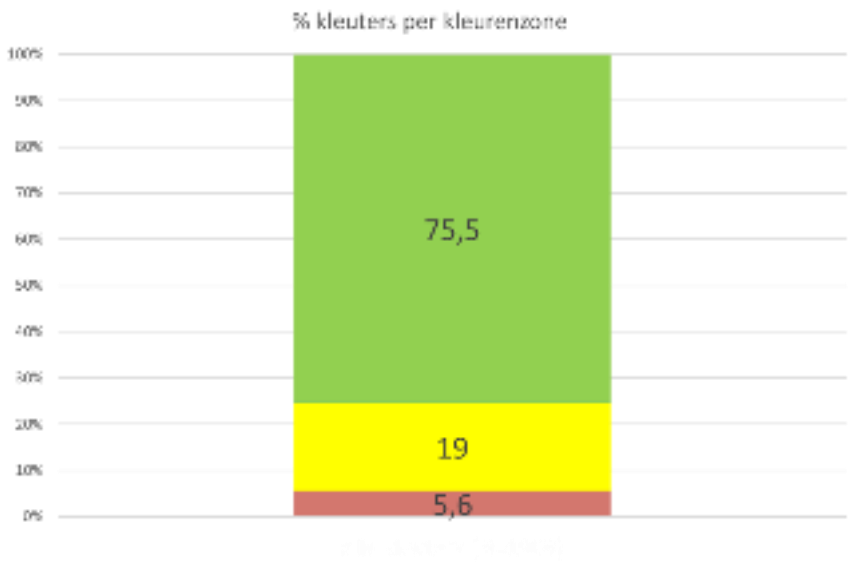
Na de derde ronde werden opnieuw de resultaten van de cesuren aan de cesuurbepalingscommussie gepresenteerd zoals voorgesteld in Tabel 58: Resultaten cesuurbepaling ronde 1.

Cesuur A		Cesuur B		
29	**0,520**	16	**-0,630**	mediaan
32	0,780	19	-0,400	p75
27	0,330	14	-0,820	p25
19	-0,400	9	-1,380	min
34	0,950	22	-0,110	max

Tabel 59: Resultaten cesuurbepaling ronde 3

- De mediaanwaarde van cesuur A is onveranderd gebleven. Cesuur B is een beetje opgeschoven naar boven (item 16 i.p.v. item 15).
- De afstand tussen p25 en p75 (= middelste helft van de deelnemers) is tijdens deze ronde bij cesuur A kleiner geworden. Bij cesuur B is ze opnieuw iets groter worden. In vergelijking met de eerste cesuur is de eensgezindheid wel groter geworden, zowel voor cesuur A als voor cesuur B.
- Zowel voor cesuur A als voor cesuur B zijn de minima en maxima verder naar elkaar toegekomen. Ook dit wijst op een toename in de eensgezindheid binnen de cesuurcommissie.

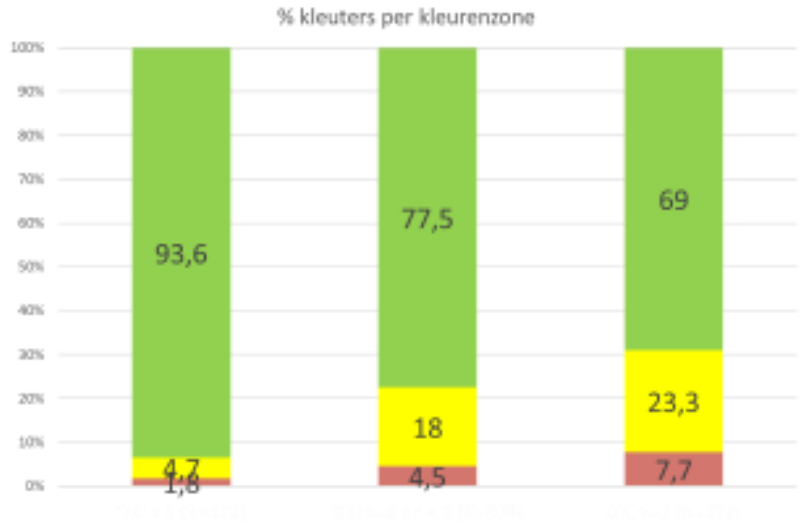
5.2 Presentatie van de impactdata op kindniveau na de derde ronde



Figuur 29: Impactdata op kindniveau cesuurbepaling ronde 3

Cesuur A is niet veranderd, waardoor de groene zone even groot is gebleven. Er is een kleine verschuiving gebeurd in cesuur B, waardoor de rode zone een beetje groter is geworden en de oranje zone iets kleiner.

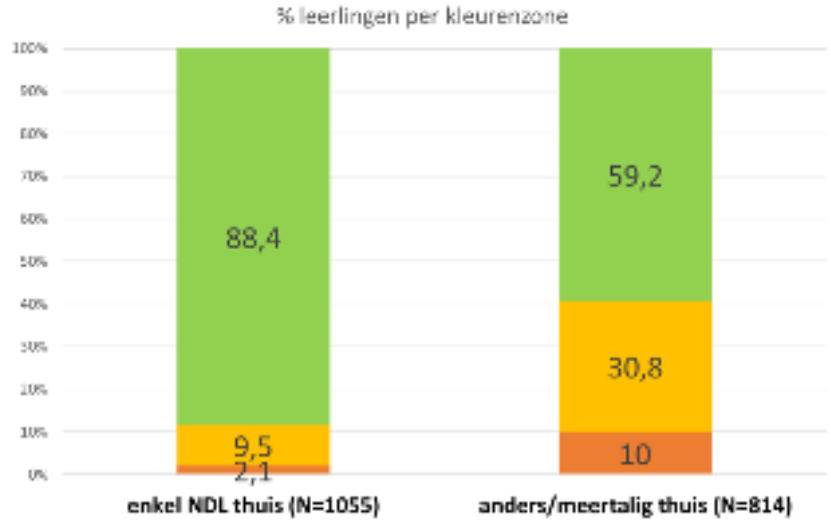
5.3 Presentatie van de impactdata op schoolniveau (OKI-categorie van de school) na de derde ronde



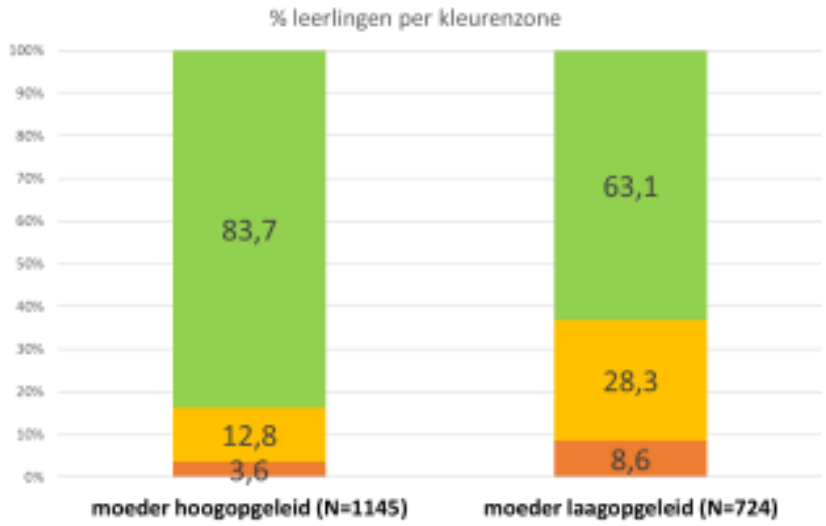
Figuur 30: Impactdata op schoolniveau cesuurbepaling ronde 3

Cesuur A is gelijk gebleven, waardoor de groene zone ook op schoolniveau niet veranderd is. Cesuur B werd iets hoger gelegd. De rode zone is daardoor iets groter geworden voor scholen met een gemiddelde OKI >=1 en OKI >=2. Voor scholen met OKI <1 zijn er geen verschuivingen.

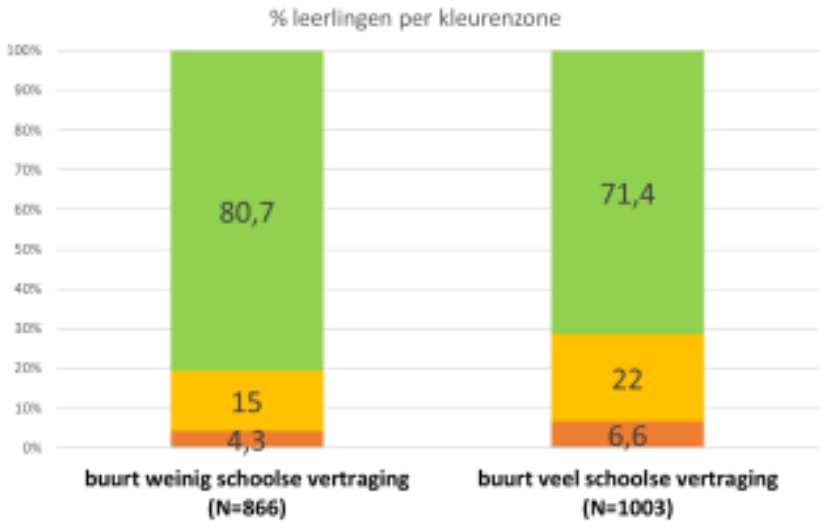
5.4 Presentatie van de impactdata voor OKI-indicatoren na de derde ronde



Figuur 31: impactdata voor OKI-indicator gezinstaal cesuurbepaling ronde 3



Figuur 32: Impactdata voor OKI-indicator opleidingsniveau moeder cesuurbepaling ronde 3



Figuur 33: Impactdata voor OKI-indicator buurt cesuurbepaling ronde 3

Het is opvallend dat bij de kleuters met een andere gezinstaal 1 op 4 kleuters zich in de rode of oranje zone bevindt.

6 ▪ Conclusie

De commissie was evenwichtig samengesteld en de leden stelden zich constructief-kritisch op, ondanks de ongewone omstandigheden (online bijeenkomst). Alle commissieleden waren bijzonder aandachtig tijdens de plenaire toelichting en voerden de opdrachten met grote toewijding en precisie uit, zoals we konden vaststellen in de discussiemomenten.

Doorheen de verschillende rondes groeiden de commissieleden dichter naar elkaar toe en bereikten een vrij goede consensus. Het is wel opvallend dat ze zich niet meer sterk lieten beïnvloeden na de eerste individuele cesuurbepaling. Mogelijk heeft dit te maken met de ruime tijd die commissieleden hadden om zich te buigen over hun beslissing als gevolg van de digitale organisatie van de cesuurbepaling: de leden hadden mogelijk de overwegingen die een rol kunnen spelen bij het leggen van een cesuur goed doordacht, waardoor er weinig ruimte was voor aanpassingen.

Met deze cesuren kan het onderzoeksteam aan de slag om het definitieve instrument samen te stellen en de handleiding vorm te geven.

HOOFDSTUK 8: GEVOLGEN VAN DE VERSCHILLENDE ONDERZOEKSSTAPPEN VOOR TOETSONTWIKKELING

In dit hoofdstuk beschrijven we de stappen die gezet werden om te komen tot het definitieve instrument voor het kalibratie-onderzoek. Door triangulatie van informatie van experten en ervaringsdeskundigen en kwantitatieve en kwalitatieve informatie over de werking van de screening uit een afname bij kleuters., kunnen we na het kalibratie-onderzoek een kwalitatief taalscreeningsinstrument samenstellen.

1 ▪ Vertrekpunt: de SALTO-toetsbatterij

Zoals eerder vermeld werd voor de ontwikkeling van het instrument vertrokken van de toetsbatterij die ontwikkeld werd in het kader van SALTO. Voor de ontwikkeling van SALTO werden immers meer taken en items ontwikkeld dan de 8 taken en 39 items die momenteel in gebruik zijn om leerlingen aan het begin van het eerste leerjaar te screenen. Concreet bestond de SALTO-toetsbatterij uit 21 taken en 123 items die potentieel hergebruikt konden worden in de taalscreening voor kleuters, mits deze taken geschikt bleken om de taalvaardigheid van kleuters te screenen.

1.1 Screening van de psychometrische informatie op itemniveau

Het oorspronkelijke SALTO-rapport en de oorspronkelijke analyses die werden uitgevoerd vormen een eerste waardevolle bron van informatie om te bepalen welke items uit de SALTO-toetsbatterij geschikt zijn voor een taalscreening van kleuters. Voor een eerste selectie van geschikte items bekeken we de moeilijkheidsgraad, standaardfout en fitstatistieken van de verschillende items. Omdat we selecteerden in functie van een andere populatie, met een lagere taalvaardigheid, was vooral de moeilijkheidsgraad van een item richtinggevend voor selectie.

- We vertrokken van de 21 taken en 123 items die ontwikkeld werden voor SALTO. Anders dan voor het ontwikkelproces bij SALTO elimineerden we niet meteen de items met een beperkte moeilijkheidsgraad. Integendeel, net de items die uit de analyse kwamen als eenvoudig voor leerlingen van het lager onderwijs. werden weerhouden als mogelijk interessant voor een taalscreening kleuters. (N= 123, 21 taken) .
- De behouden items werden opnieuw gescreend voor discriminatiegraad en itemfit. Items met een slechte itemfit werden verwijderd (N= 100, 19 taken).
- De oorspronkelijke analyses, waarbij de te gemakkelijke en slecht fittende items van bij aanvang werden verwijderd, werden nogmaals uitgevoerd om ook een zicht te krijgen op de moeilijkheidsgraad en fit van een verzameling van items binnen een taak. Deze analyse gaf bijkomende, niet-doorslaggevende informatie voor de selectie van taken en bijbehorende items en maakte vooral de gerichte spreiding van moeilijkheidsgraad eenvoudiger: we selecteerden de meest eenvoudige taken uit SALTO, en vulden deze aan met taken met een gemiddelde en hoge moeilijkheidsgraad.

1.2 Inhoudelijke screening van de items

De 85 items die overbleven op basis van de screening van psychometrische data werden verder gescreend op basis van inhoudelijke criteria:

- spreiding over de verschillende doelstellingen van het referentiekader (zie 2.4 Aanpassingen toetsconstruct, toetsmatrijs en typetaken);
- spreiding over de verschillende vormen van complexiteit en abstractieniveau (zie 2.3 Variatie in complexiteit).

Daarnaast voegden we een efficiëntiecriterium (zie ook 2.1 Vier criteria voor het taalscreeningsinstrument voor kleuters) toe:

- minstens drie items per taak > Een taak bestaat steeds uit het beschrijven of creëren van een context. Binnen die context krijgen kleuters verschillende opdrachten. Dat is efficiënter dan steeds nieuwe contexten oproepen.

De inhoudelijke screening bevestigde de gemaakte keuzes op basis van de psychometrische screening: de geselecteerde items bleken alle vooropgestelde doelstellingen te omvatten, alsook te verschillen in complexiteit en abstractieniveau. Ook het efficiëntiecriterium werd vervuld voor alle overblijvende taken.

Tegelijkertijd liet overlap wat betreft doelstellingen, complexiteit en abstractieniveau toe om enkele taken te schrappen. Uiteindelijk bleven in totaal 17 taken, bestaande uit 85 items over. Die werden voorgelegd aan de resonansgroep (zie 2.1 Feedback op geselecteerde SALTO-taken).

2 ▪ Hergebruikte SALTO-taken

2.1 Feedback op geselecteerde SALTO-taken

De geselecteerde taken en items werden enerzijds door het ontwikkelteam en anderzijds door de resonansgroep van feedback voorzien.

De feedback van de resonansgroepleden richtte zich op de mate waarin het SALTO-construct en de bijbehorende taken en items ingezet kunnen worden in een taalscreening voor kleuters. De resonansgroepleden antwoordden positief op de vraag of de SALTO een goede basis kon vormen, mits openheid naar aanpassingen aan toetsconstruct, toetsmatrijs en typetaken.

Zo schuiven de resonansgroepleden de volgende algemene aandachtpunten naar voren voor het betrouwbaar evalueren van taalvaardigheid van kleuters:

- rechtstreekse observatie van kleuterhandelingen;
- afstemming op de leefwereld van kleuters;
- directe toetsing;
- rekening houden met geschikte afnamecondities (locatie, afnemer…).
- de mogelijkheid voorzien voor ondersteuning en interactie met een volwassene;
- spelsituaties integreren, waar kleuters mogelijk meer vaardigheid tonen dan op een test.

Met betrekking tot specifieke taken en items wijzen de resonansgroepleden op het belang van een correcte koppeling van taken en doelstellingen, de duidelijkheid en aantrekkelijkheid van illustraties en helder en uniform taalgebruik. Ook waren er opmerkingen bij enkele minder goed gekozen afleiders.

Tot slot wijzen de resonansgroepleden op het belang van flankerende informatie bij de toets door

- de screening aan te vullen met (gestandaardiseerde) observaties;
- ouders en kinderen te betrekken bij de beeldvorming;
- mogelijkheden tot verbreding en verdieping van de beeldvorming te voorzien;
- …

2.2 Vaststellingen uit het vooronderzoek

Via het vooronderzoek verzamelden we verdere informatie over de afnamecondities en de werking van de taken en items. We testten 20 individuele kleuters en daarnaast een klas 5-jarigen in een multiculturele school in Gent. Acht onderzoekers van het CTO namen de screening af in de thuissituatie en in één schoolcontext.

De onderzoekers namen notities en maakten opnames. De afname gebeurde op gestandaardiseerde wijze, maar de kleuters mochten spontaan reageren tijdens de testafname, en de onderzoekers stelden gerichte vragen om de redenering achter een onverwachte of foutieve antwoordkeuze te achterhalen.

Afnameconditie

Met betrekking tot de afname condities deden we op basis van het vooronderzoek de volgende vaststellingen:

- de meeste kleuters kunnen zich makkelijk twintig minuten tot een half uur concentreren mits voldoende afwisseling in taken (maar die concentratiespanne is zeer wisselend naar gelang de rijpheid en de taalvaardigheid van de kleuter);
- kleuters hebben baat bij het opnemen van doe-opdrachten (omwille van concentratie, motivatie en testrijpheid);
- digitale afname is mogelijk en motiverend;
- alle kleuters (behalve enkele kleuters met weinig schoolse ervaring) zijn in staat de opdrachten motorisch uit te voeren indien we kiezen voor het trekken van een kring eerder dan het zetten van een kruisje;
- afname in groepjes van 6 is het absolute maximum: in scholen met veel taalzwakkere kleuters en/of veel kleuters met minder schoolse ervaring of rijpheid zijn groepjes van 4 à 5 kleuters aan te raden;
- een grote hoeveelheid meerkeuze-opdrachten voor (taal)zwakkere kleuters kan moeilijk en demotiverend zijn.

2.2.1 Taken en items

Het vooronderzoek gaf daarnaast ook informatie over de concrete werking van typetaken, taken, afleiders en instructies, en eventuele aanpassingen die nodig zijn om deze te optimaliseren.

Algemene vaststellingen over typetaken, taken, instructies en afleiders waren de volgende:

- de kleuters vinden de nieuwe tekeningen ‘mooi’, ‘leuk’ en duidelijk (m.a.w. aantrekkelijk maar zonder dat de details teveel afleiden);
- meerkeuze-opdrachten waarbij meerdere antwoordmogelijkheden kunnen worden gegeven zijn te hoog gegrepen voor de meeste kleuters en vormen nauwelijks een meerwaarde bij een taalvaardigheidsscreening.;

- toetsformats blijven best uniform: variatie, door bijvoorbeeld af te wisselen tussen single choice en multiple choice aanduidopties zorgen voor verwarring.

2.3 Conclusie

Na triangulatie van de gegevens uit (a) onderzoek en expertengesprekken, (b) feedback van resonansgroepleden en (c) ervaringen uit het vooronderzoek kunnen we besluiten dat een aantal taken en items uit SALTO een goede basis vormen voor een taalscreeningsinstrument voor kleuters, onder voorwaarden. Opvallend is dat deze drie bronnen van informatie in dezelfde richting wijzen: om echt tot een kwalitatieve meting te komen, zijn aanpassingen nodig aan toetsconstruct, toetsmatrijs en typetaken.

2.4 Aanpassingen toetsconstruct, toetsmatrijs en typetaken

Uit de informatie van experten en de literatuur, en uit feedback op de geselecteerde SALTO-taken kunnen we besluiten dat een toetsenbatterij met enkel ongewijzigde SALTO-taken niet voldoende zijn om de geselecteerde doelstellingen op een betrouwbare manier te testen bij kleuters.

Met name de typische toetstaken - een vraag gevolgd door enkele antwoordopties waaruit het correcte antwoord moet worden gekozen - dienen uitgebreid te worden: gestandaardiseerde observaties bijvoorbeeld vormen een belangrijke aanvulling. Waar SALTO een aantal doelstellingen op een indirecte manier toetste (door bijvoorbeeld te laten beoordelen of iemand een instructie correct uitvoert) bieden 'gestandaardiseerde observaties' als nieuwe typetaak de mogelijkheid om deze doelstellingen direct te testen.

Deze zelfde typetaak bleek ook van belang om doelstellingen te testen die nog niet voorkwamen in SALTO-toetsbatterij. Doelstellingen die te maken hadden met vragen naar persoonlijke voorkeuren en naar eigen ervaringen waren nog niet opgenomen in SALTO, maar wel geselecteerd als belangrijke doelstellingen uit het referentiekader NT2. De toetsmatrijs werd uitgebreid met deze doelstellingen.

Meerdere experten geven aan dat kleuters tijdens meerkeuze-opdrachten niet steeds ten volle hun taalvaardigheid kunnen tonen: het uitbreiden van de typetaken voor de taalscreening bij kleuters is ook daarom aangewezen. Meerkeuze-opdrachten liggen immers vrij ver af van de kleuterontwikkeling. Andere typetaken dan meerkeuzeopdrachten bieden kleuters - ook kleuters met minder schoolse ervaring en mogelijkheden - de kans om te laten zien wat ze kunnen.

Tot slot blijkt variatie in soorten opdrachten en contexten belangrijk om de fluctuerende aandacht van kleuters te kunnen vasthouden.

We beslisten verder om uitspraken over kleuters niet louter op één toetsmoment te baseren. Via een handleiding die de inbedding van KOALA kadert in een breder beeldvormingsproces vermijden we onbetrouwbare metingen. Het toetsgegeven werd op die manier uitgebreid met flankerende observaties om de informatie uit de toetsafname te bevestigen, verder te verkennen of te verdiepen.

2.5 Herwerking van SALTO-taken

De taken die we behielden uit de SALTO-batterij, kregen een 'make-over':

- alle tekeningen werden herwerkt: een meer hedendaagse vormgeving, in kleur, met jongere kinderen…;
- bij een aantal taken werden niet alle items behouden of werden de afleiders lichtjes aangepast;
- een aantal taken veranderde van type: een aantal meerkeuze-opdrachten die in SALTO doelstellingen indirect probeerden te meten zijn bijvoorbeeld aangepast naar gestandaardiseerde observaties om op een directe manier de doelstelling in kaart te brengen.

2.6 Nieuwe taken, geschikt voor kleuters

De taken uit SALTO bleken onvoldoende om de volledige toetsmatrijs met alle doelstellingen met overlap af te dekken. 17 nieuwe taken werden ontwikkeld om te komen tot een toetsbatterij die voldoende rekening houdt met de kleutercontext. De extra ontwikkeling gaf de mogelijkheid om:

- elke doelstelling uit de toetsmatrijs van meerdere taken te kunnen voorzien;
- de leefwereld van de derde kleuterklas voldoende aan bod te laten komen;
- de verschillende typetaken vorm te geven;
- de collectie bestaande taken aan te vullen met taken met lagere talige eisen.

2.7 Taken verwijderen

Na het doornemen van de feedback van de resonansgroep en de algemene en taakspecifieke beslissingen met betrekking tot aanpassingen behielden we 13 van de 17 geselecteerde taken uit SALTO in een nieuwe vorm. 4 taken werden niet weerhouden omdat het niet mogelijk was ze aan te passen aan de hierboven opgesomde feedback. In concreto zou de aanpassing overeenkomen met het ontwikkelen van een nieuwe taak.

2.8 Conclusie

Met nieuwe taakontwikkeling en met hergebruik van aangepaste SALTO-taken, bleek het mogelijk om - binnen het nieuwe toetsconstruct - een toetsbatterij van 30 taken te creëren die geschikt is om van 5-jarige kleuters af te nemen.

Deze 30 toetstaken – de nieuw ontwikkelde taken en de geselecteerde SALTO-taken – moesten echter nog getest worden in kleinere en grotere afnames. In elke van deze test- en ontwikkelrondes speelde input van afnemers, kleuters en experten een rol om de uiteindelijke toetsenbatterij vorm te geven. Na triangulatie van deze data gebeurden in elke ontwikkelronde aanpassingen aan de screening.

3 ▪ Tweede test-en ontwikkelronde: pilootonderzoek

Het pilootonderzoek was grootschaliger dan het vooronderzoek. Twaalf scholen en meer dan 300 kleuters namen deel aan het pilootonderzoek. De afnamecondities waren gelijkaardig aan de condities van het kalibratie-onderzoek. De kleuters mochten bijvoorbeeld niet meer reageren tijdens de afname. De data werden systematisch verzameld en werden nadien geanalyseerd. Daarnaast kon via participatieve observatie ook kwalitatieve informatie over de screening als geheel en de verschillende taken en items verkregen worden.

3.1 Vaststellingen uit het pilootonderzoek

De 30 taken die na het vooronderzoek werden herwerkt, werden in het kader van het pilootonderzoek (zieHoofdstuk 3: Vooronderzoek, voor een volledige beschrijving) opnieuw voorgelegd aan de resonansgroep. Naast deze feedback werd ook de feedback van afnemers verzameld via feedback – en observatieformulieren en informatie uit de kwantitatieve analyses meegenomen.

Op vlak van afnamecondities en taken en taaktypes werd weinig negatieve feedback ontvangen na de grondige aanpassingen op basis van de eerdere feedback. Enkel voor de instructies voor de nieuw-ontwikkelde gestandaardiseerde observaties bleken in sommige gevallen nog ruimte voor verbetering, om op die manier beter tot een gestandaardiseerde uitvoering en consistente verbetering te komen.

Het pilootonderzoek leverde voornamelijk informatie op over de werking van items en afleiders. Via de kwantitatieve analyses identificeerden we enkele problematische items: zeer gemakkelijke items (door bijna alle kinderen correct beantwoord), zeer moeilijke items of inconsistente items (bijvoorbeeld items die sterk taalvaardige kinderen toch fout maken).

De feedback van afnemers en van resonansgroepleden wezen op een paar problemen bij enkele taken, items of afleiders: soms was het nodig om meer uitdaging te creëren, in andere gevallen werd een grote moeilijkheidsgraad gesignaleerd.

Signalen van te eenvoudige items:

- volledig te gemakkelijk item;
- inputtekst is te eenvoudig;
- (te) weinig items in een taak;
- (te) weinig afleiders bij overzichtsprenten;
- te veel aandacht voor een element in de input;

- te expliciet opgenomen in de afbeelding van de antwoordoptie.

Signalen van te moeilijke items:

- enkele te lange (delen van) verhalen;
- te weinig aandacht voor een element van een antwoord in de input;
- meervoudige interpretaties mogelijk;
- antwoordoptie te weinig expliciet opgenomen in de afbeelding;
- afbeeldingen van emoties te klein en onduidelijk.

Uit de triangulatie van gegevens bleken de inzichten uit pilootonderzoek en feedback van afnemers elkaar te bevestigen.

3.2 Herwerking na het pilootonderzoek

Na het pilootonderzoek werden de dertig taken van de screening herwerkt op basis van een tweede feedbackronde door de resonansgroep, feedback van de toetsafnamers en de kwantitatieve resultaten uit het pilootonderzoek.

Te eenvoudige items werden aangepast door:

- het item te schrappen;
- meer informatie/uitdaging op te nemen in de inputtekst;
- items toe te voegen aan een taak;
- extra afleiders of elementen in een afleider toe te voegen;
- minder expliciet opnemen van antwoordelementen in de input;
- minder expliciet opnemen van antwoordelementen in de tekening (output).

Te moeilijke items of taken werden aangepast door

- lange stukken tekst te splitsen; soms werd een tekening toegevoegd als tussenblad, zodat kleuters niet afgeleid werden door de antwoordmogelijkheden van een vraag, maar hun aandacht op een verhaal konden richten;
- tekst van de input eenduidiger te maken om verschillende interpretaties te vermijden;
- antwoordelementen explicieter op te nemen in de input;
- antwoordelementen explicieter op te nemen in de tekening (output);
- in te zoemen op gezichten bij items waar emoties centraal stonden.

Items die als problematisch werden geïdentificeerd door de kwantitatieve analyses, werden in detail onderzocht en aangepast op basis van de veronderstelde problemen. De algemene lijnen van feedback en de concrete feedback van de resonansgroepleden of afnemers waren hierbij richtinggevend.

- Voorbeeld van een gemakkelijk item:
  Item 4.1 van taak JELLE: In deze doe-en zoekopdracht worden kleuters gevraagd om een persoon aan te duiden op een familiefoto. Voor het (te) gemakkelijke item blijkt echter geen waarschijnlijke afleider beschikbaar: kleuters moeten immers een man aanduiden, terwijl er maar één man zichtbaar is.

  Voorstel voor herwerking: De prent aanpassen door een extra man toe te voegen.

- Voorbeeld van een moeilijk item:
  Item 51.1 van taak Konijntjes: *"Waar is het huis van Paultje? Waar woont Paultje?"*. Konijntjes is een meerkeuze-taak waarbij kleuters na het horen van een verhaal gevraagd wordt om de juiste tekening te kiezen uit vier opties. Omdat bij dit item kleuters eerder getailleerde informatie moesten identificeren die vrij vroeg in de tekst aan bod kwam, is het niet vreemd dat slechts 5 van de 31 kinderen hier het correcte antwoord hebben gegeven.

  Voorstel voor herwerking: de tekst opsplitsen, en onmiddellijk de vraag stellen na het stuk verhaal waarin het over de woonplaats van Paultje gaat. In de vraagstelling bovendien het woord 'huis' door 'holletje' vervangen, en ook de holletjes op de tekeningen groter en duidelijker maken.

3.3 Concrete voorbeelden van herwerking

Voorbeeld aanpassing na pilootonderzoek; taak Konijntjes

> tussenblad voor kleuters wanneer ze moeten luisteren naar het verhaal (aanpassen instructie toetsafnemer)

1. Geen opmerkingen van toetsafnemers en leraren
2. Opmerkingen van resonansgroep: is prent 7 wel een goede afleider?
3. Analyses > alle afleiders worden gekozen, moeilijker item (measure 1.18), prestaties in lijn der verwachtingen (taalvaardige kinderen scoren beter)

Conclusie: geen aanpassing

Voorbeeld aanpassing na pilootonderzoek; taak Zandtafel

1. Observaties: kleuters vinden dit zeer leuk
2. Resonansgroep: kleutercontext +
3. Analyses: makkelijke taak > discrimineert onvoldoende

Conclusie: meer afleiders (aanpassen afbeelding)

Voorbeeld aanpassing na pilootonderzoek; taak In bad

Rosanne heeft de hele middag in de zandbak van de speeltuin gespeeld. Nu hangt ze vol zand: het zit tussen haar tenen, aan haar handen, in haar haren. [aanwijzen] "In bad,[i]", zegt mama. Rosanne doet eerst haar kleren uit en springt dan in bad. Rosanne vindt het leuk om in bad te gaan, omdat ze dan kan spelen met haar badeendjes. "

Wat heeft Rosanne de hele middag gedaan? Waar speelde Rosanne? Trek een kring rond de juiste tekening.

1, Geen bedenkingen
2. Geen bedenkingen
3, Analyses: problematisch item (outlier) met hoge outfit waarde (>2) slechts één kind heeft fout geantwoord, waarbij het gaat om een kind met een hogere vaardigheidsscore

Conclusie
> Item behouden omwille van succeservaring bij begin van de taak (~literatuur)
> moeilijker maken door ook andere afleiders in verhaal te vernoemen (aanpassen vraag)

4 ▪ Instrument voor Kalibratie-Onderzoek

4.1 Definitie van taken en items

Een 'taak' vormt een geheel van meerdere vragen ('items') dat inhoudelijk coherent is. Dat wil zeggen dat er aan het begin van een taak een specifieke context wordt gecreëerd (bv. taak Lievelingsboeken: *In de klas staan veel boeken. De juf zoekt het lievelingsboek van elk kind. Jullie moeten kijken welk van de vier boeken het lievelingsboek is van elk kind.*; taak Kabouters: *Ik ga een verhaal over twee kabouters vertellen. Eerst vertel ik een stukje van het verhaal. Bij het verhaal zijn veel tekeningen gemaakt. Jij moet straks kiezen welke tekeningen juist zijn.*). Alle vragen hebben betrekking op deze context maar worden onafhankelijk van elkaar beoordeeld.

4.2 Typetaken in het kalibratieonderzoek

Er zijn drie typetaken (zie 3 ▪ TYPETAKEN IN KOALA). Elk van deze typetaken werden opgenomen in het kalibratieonderzoek en gespreid voor afname over de verschillende clusters.

Het gaat met name over:

- Doe-en zoekopdrachten op papier/tablet (6 taken)
- Meerkeuze-opdrachten op papier/tablet (15 taken)
- Gestandaardiseerde observaties (9 taken)

4.3 Aantal taken en items

Het instrument dat we voor het kalibratieonderzoek hebben gebruikt bevat 30 taken, bestaande uit 148 items in totaal. Alle taken hebben een taaknummer (van 1 tot 30) toegewezen gekregen op basis van de alfabetische volgorde van de taaktitels.

De meeste taken (27) bevatten 4, 5 of 6 items. Drie taken, allen gestandaardiseerde observaties, wijken hiervan af: taak Eten heeft 3 items, taak Hoepel heeft 7 items, en taak Bewegen heeft 12 items.

4.4 Papieren en digitale versie

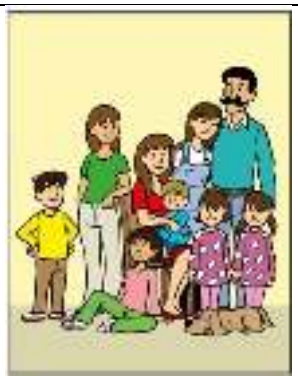
Van een aantal taken bestond er een papieren versie én een digitale versie die via tablets kon worden afgenomen. Het gaat hierbij om alle 15 meerkeuze-opdrachten en om 3 doe-en zoekopdrachten (zie hierboven).

Inhoudelijk waren beide versies van de taken identiek en was in principe de visuele presentatie ook hetzelfde: Wat kleuters bij de papieren afname op een DIN A4 blad zagen, kregen kleuters tijdens digitale afnames op het scherm van de tablet te zien.

Naar gelang een taak op papier of digitaal werd afgenomen, verschilde wel de gevraagde respons door de kleuters:

- Papier: Bij meerkeuze-opdrachten trekt de kleuter een kring rond de juiste afbeelding. Bij doe-en zoekopdrachten (taak Jelle, taak Waar is? en taak Zandtafel) zet de kleuter een kruisje. Instructie: *Trek een kring rond … / Zet een kruisje op ….*
- Digitaal: Bij meerkeuze-opdrachten tikt de kleuter op de juiste afbeelding op het beeldscherm. Bij doe-en zoekopdrachten (taak Jelle, taak Waar is? en taak Zandtafel) tikt de kleuter op een detail van de tekening.
  Instructie: *Tik op …*

Technisch gezien waren de digitale afnames hetzelfde bij het pilootonderzoek en het kalibratieonderzoek (geprogrammeerd in Qualtrics).

Voorbeeld van papieren versie en digitale versie (taak Jelle)	
PAPIER	*DIGITAAL*
	

5 ▪ Definitief instrument

Net als in de vorige test- en ontwikkelrondes wordt de informatie uit het kalibratieonderzoek - de informatie uit de statistische analyses, de feedback van toetsafnemers en scholen en van toetsassistenten - getrianguleerd om te komen tot een definitief instrument.

Voor de selectie van de definitieve items werd rekening gehouden met psychometrische, inhoudelijke en vormelijke criteria. Op basis van deze criteria werden 7 taken bestaande uit 33 items in totaal geselecteerd (zie Hoofdstuk 9: Samenstelling Definitieve instrument). Met dit geheel blijkt het mogelijk om een valide en betrouwbaar screeningsinstrument op te leveren dat ingezet kan worden om op een efficiënte en inzetbare wijze kleuters te detecteren die nood hebben aan talige ondersteuning.

De handleiding en de instructiebundel helpen leraren om bij het gebruik rekening te houden met de vaststellingen uit dit onderzoek.

HOOFDSTUK 9: SAMENSTELLING DEFINITIEVE INSTRUMENT

1 ▪ Screening Items kalibratie-onderzoek

Voor het kalibratieonderzoek werden 30 taken en 148 items uitgetest bij bijna 2000 kinderen, op papier en met tablets. Al deze items komen theoretisch gezien in aanmerking om opgenomen te worden in het definitieve instrument.

Om te komen tot een screening die voldoet aan de eerder opgestelde criteria van 'validiteit', 'betrouwbaarheid', 'efficiëntie' en 'inzetbaarheid', onderzoeken we de verschillende items uitgebreid om te komen een set van valide en betrouwbare items binnen een efficiënte en inzetbare toets.

Daartoe werden de 148 items inhoudelijk en psychometrisch gescreend. Dit stelt ons in staat om de beste items te selecteren voor het finale instrument. Hieronder worden de stappen van het selectieproces beschreven.

1.1 Psychometrische screening

1.1.1 Misfit

Van de 148 items die getest werden in het kalibratie-onderzoek werden vijf items geïdentificeerd als 'outfit items' (1.2 DATASET VAN BETROUWBARE TOETSITEMS) Deze items hebben een outfit MNSQ-waarde > 2, wat aangeeft dat er ruis zit op die items, en dat andere aspecten dan de luistervaardigheid meespelen in de score van een kleuter. Deze vijf items met misfit werden niet meegenomen in de verdere analyses en uitgesloten van selectie voor het definitieve instrument.

Tabel 60 geeft een overzicht van de niet betrouwbare items en hun misfit waarden:

Item	Taak	Infit		Outfit	
		MNSQ	ZSTD	MNSQ	ZSTD
1.4	Bewegen	.83 |	-.64	3.59	3.34
1.7	Bewegen	1.16	.72	2.77	2.59
4.2	Eenzaam	1.21	3.09	3.02	8.73
12.2	Konijntjes	1.21	1.74	2.63	4.76
18.4	Naar bad	1.19	3.45	2.13	7.64

Tabel 60: Overzicht van niet-betrouwbare items

Na uitsluiting van de vijf misfittende items blijven er 143 bruikbare items over. Het moeilijkste item heeft een measure van 2,42. Het gemakkelijkste items heeft een measure van -3,19.

1.1.2 Differential Item Functioning (DIF)

Voor het definitieve instrument willen we items selecteren waarbij de moeilijkheid niet verschilt naargelang afnamecondities (papier of tablet) of achtergrondkenmerken van de kleuter (gezinstaal en opleiding moeder). Op die manier zijn de resultaten van de screening betrouwbaar en gelijk, ongeacht achtergrond of afnamemodaliteit.

Aan de hand van DIF-analyses (Differential Item Functioning) gingen we voor elk van de 143 items na of er sprake is van significante DIF-contrasten. We gingen na of de item measures significant verschillend zijn tussen:

- Afnameconditie: papier versus digitaal
- Gezinstaal: enkel Nederlands versus meertalig
- Opleiding van de moeder: hoogopgeleide versus laagopgeleide moeder

We identificeren **25 items** (17,4%) met één of meerdere significante DIF-contrasten. Het gaat over de volgende items:

		Significante DIF-contrasten		
		afnamemodaliteit	gezinstaal	opleiding moeder
1.2	Bewegen		X	
1.5	Bewegen		X	
8.2	juf is jarig	X		X
9.2	Kabouters		X	
10.2	Klasafspraken	X		
11.2	Klastaakjes		X	
12.6	Konijntjes		X	
13.3	Lievelingsboeken	X		
15.2	Mona's hoeken	X		
15.4	Mona's hoeken	X		
16.4	Mug en olifant		X	
17.3	Myriam		X	
17.4	Myriam	X	X	
17.6	Myriam		X	
19.4	Park	X		
20.2	Rommel in de eetzaal		X	
20.3	Rommel in de eetzaal		X	
22.4	Speeltijd		X	
26.1	Varken en rups	X		
26.2	Varken en rups	X		
26.3	Varken en rups	X	X	

27.1	Verjaardagsfeest		X	X
27.5	Verjaardagsfeest		X	
29.1	Waar is		X	
30.3	Zandtafel	X		

Items met DIF houden we uit de selectie voor het definitieve instrument. Op deze manier blijven er 118 items over om een instrument mee samen te stellen.

○ Discriminerende waarde

We gingen op zoek naar items met een hoge discriminerende waarde. Zulke items zijn goed in staat om kleuters met een verschillend luistervaardigheidsniveau van elkaar te onderscheiden.

Bij de selectie van items voor het definitieve instrument geven we voorrang aan items met een hoge(re) discriminerende waarde. We onderscheiden drie categorieën:

Discriminerende waarde	Geschikt voor instrument?	Aantal items
>= 1	Zeer geschikt	78 items
>= 0,70 en < 1	Voldoende geschikt	52 items
< 0,70	Minder geschikt	13 items

Bovenstaande tabel laat zien dat 120 van de 143 items geschikt zijn om opgenomen te worden in het definitieve instrument. Slechts **13 items** (9%) lijken op basis van de discriminerende waarde minder geschikt, omdat ze weinig informatie geven over de luistervaardigheid van kleuters. Vijf van deze items werden ook geïdentificeerd als minder interessant om op te nemen omwille van significante DIF-contrasten.

Item	Taak	Discriminatiewaarde	DIF
2.2	Boekenhoek	0,47	
4.4	Eenzaam	0,39	
9.2	Kabouters	0,67	X
11.2	Klastaakjes	0,59	X
12.1	Konijntjes	0,47	
12.6	Konijntjes	0,38	X
15.1	Mona's hoeken	0,27	
16.1	Mug en olifant	0,51	
17.4	Myriam	0,69	X
18.2	Naar bad	0,67	

19.3	Park	0,56	
26.3	Varken en rups	0,60	x
26.4	Varken en rups	0,66	

○ Verdeling over vaardigheidsniveaus

Er zijn dus **110 items** zonder significante DIF-contrasten én met een voldoende hoge discriminerende waarde (>= 0,70). Uit deze items willen we een selectie maken voor het definitieve instrument. We streven daarbij naar een doordachte verdeling over de verschillende vaardigheidsniveaus (kleurenzones) die worden onderscheiden door cesuren (zie Hoofdstuk 7: Cesuren bij koala).

- Items in de groene zone:
  Sluiten aan bij het vaardigheidsniveau van kleuters boven cesuur A.
- Items in de oranje zone:
  Sluiten aan bij het vaardigheidsniveau van kleuters onder cesuur A.
- Items in de rode zone:
  Sluiten aan bij het vaardigheidsniveau van kleuters onder cesuur B.

Kleurenzone	Situering	Measure	Aantal items
Groen	Boven cesuur A	>= 0,52	29
Oranje	Tussen cesuur A en B	>= -0,63 en < 0,52	43
Rood	Onder cesuur B	< -0,63	38

We stellen vast dat er voldoende items zitten in elke kleurenzone. Dit geeft ons de ruimte om een instrument op te stellen dat enerzijds kleuters kan situeren in de verschillende kleurenzones, en anderzijds ook binnen deze kleurenzones kleuters met een verschillend luistervaardigheidsniveau kan identificeren.

Daarnaast stellen we vast dat het grootste aantal items zich bevindt in de oranje en de rode zone. Dit is niet vreemd, gezien de doelstelling van de toets: namelijk kleuters identificeren die (extra) taalstimulering nodig hebben. Dit werd ook eerder bevestigd in de psychometrische analyses: de taalscreening slaagt er inderdaad in om kleuters met lagere taalvaardigheidsniveaus te onderscheiden.

Voor het definitieve instrument willen we die taken selecteren waarbij er voldoende items uit elke kleurenzone vertegenwoordigd zijn, en die voldoende spreiding over de measures binnen elke kleurenzone verzekeren.

1.2 Inhoudelijke screening

De 110 psychometrisch geschikte items om kleuters te identificeren die nood hebben aan (extra) taalstimulering, werden vervolgens aan een inhoudelijke screening onderworpen. Omwille van efficiëntie en inzetbaarheid in de kleuterklas, keken we naar een verdeling over de verschillende doelstellingen van de toetsmatrijs (inzetbaarheid), een verdeling over de typetaken (betrouwbaarheid en validiteit) en naar het aantal items per taak (efficiëntie).

1.2.1 Verdeling over de doelstellingen

Voor het definitieve instrument laten we elk van de vier doelstellingen van luistervaardigheid aan bod komen. Op die manier wordt luistervaardigheid voldoende breed in kaart gebracht, en krijgt een leraar ook basale informatie over de mate waarin een kleuter elke doelstelling beheerst. Om die reden willen we voor elke doelstelling we minstens één taak selecteren.

De geschikte items zijn als volgt verdeeld over de verschillende doelstellingen:

- Instructies begrijpen (30 geschikte items)
- Informatieve mededelingen begrijpen (31 geschikte items)
- Vragen begrijpen (38 geschikte items)
- Verhalen begrijpen (11 geschikte items)

We kunnen concluderen dat de geschikte items mooi verdeeld zijn over de vier doelstellingen, en dat deze verdeling ons hoogstwaarschijnlijk in staat stelt om de alle doelstellingen aan bod te laten komen in het instrument.

Enkel voor de doelstelling ‘verhalen begrijpen’ blijft de keuze beperkt tot 11 geschikte items van de oorspronkelijk 22 items: 6 van deze 22 items vertoonde immers DIF, 4 items bleken te weinig te discrimineren, en een laatste item was een misfit. Mogelijk wordt het voor deze doelstellingen moeilijker om geschikte taken en items te selecteren.

1.2.2 Verdeling over de typetaken

Er zijn drie typetaken: taken met gestandaardiseerde observaties, doe- en zoekopdrachten en meerkeuze-opdrachten . Uit het kalibratieonderzoek bleek dat de variatie tussen deze drie typetaken erg gewaardeerd werd door zowel de kleuters als de toetsafnemers. Het zorgt voor afwisseling en de kleuters kunnen langer gefocust blijven op de taken. De mix van typetaken draagt op die manier bij aan de betrouwbaarheid van de toets. Kleuters die zich minder op hun gemak voelen bij een bepaalde typetaak kunnen via andere typetaken hun luistervaardigheid aantonen.

De geschikte items zijn als volgt verdeeld over de verschillende typetaken:

- Gestandaardiseerde observaties (44 geschikte items)
- Doe- en zoekopdrachten (25 geschikte items)
- Meurkeuze-opdrachten (41 geschikte items)

De verdeling over de verschillende typetaken doet vermoeden dat een selectie met de verschillende typetaken mogelijk is voor het definitieve instrument en dat dus het streven naar variatie over de typetaken lukt. Het lijkt wel waarschijnlijk dat een definitief instrument meer items zal bevatten van het type gestandaardiseerde observaties en meerkeuze-opdrachten dan van de doe- en zoekopdrachten.

Het aanbod laten komen van alle typetaken in de taalscreening, heeft automatisch tot gevolg dat een deel van de taalscreening individueel per kleuter gebeurt en een ander deel in kleine groep. De precieze werkwijze van beide afnamecondities wordt toegelicht in de handleiding en de instructiebundel voor de leraar.

1.3 Vormelijke screening

1.3.1 Aantal taken en items

In het kalibratie-onderzoek legde elke kleuter zes taken met gemiddeld 28 items af. Deze hoeveelheid bleek haalbaar voor de meeste kleuters (zie Hoofdstuk 6: Resultaten kalibratieonderzoek).

We moeten er echter rekening mee houden dat de lengte van een toets en de betrouwbaarheid van een meting aan elkaar gerelateerd zijn. Naarmate het aantal items in een instrument afneemt, daalt ook de betrouwbaarheid. Uit het kalibratie-onderzoek bleek de betrouwbaarheid op het niveau van de kleuter (person reliability) met .78 voldoende. Om betrouwbaarheidsredenen is een kleine uitbreiding ten opzichte van het kalibratie-onderzoek te verantwoorden.

Tegelijkertijd laat het concentratievermogen van 5-jarige kleuters het niet toe een (zeer) uitgebreid instrument voor te leggen: Het moet haalbaar blijven voor jonge kinderen om de toets te maken. Met een reeks van ongeveer zes à zeven taken bestaande uit in totaal een 30-tal items kunnen we een goed evenwicht vinden tussen haalbaarheid en betrouwbaarheid.

1.3.2 Aantal items per taak

Als een taak in aanmerking komt om opgenomen te worden in het definitieve instrument, nemen we alle geschikte items binnen die taak op. Niet-geschikte items - items met een significant DIF-contrast, met lage discriminerende waarde of met misfit - vermijden we. Het meenemen van de moeilijkheidsgraad (measure) maakt een evenwichtige spreiding mogelijk, algemeen en ten opzichte van de cesuren. Met oog op de efficiëntie van het instrument blijven er idealiter minstens drie geschikte items over binnen een taak.

De volgende taken bevatten minder dan drie geschikte items:

- Taak 4: Eenzaam
- Taak 15: Mona's hoeken
- Taak 16: Mug en olifant

- Taak 18: Naar bad

Als we deze taken buiten beschouwing laten, blijven er 26 taken en 104 items over om een geschikt instrument samen te stellen.

2 ▪ Selectie Voor KOALA

Na alle bovenstaande criteria in overweging te hebben genomen, maakten we een selectie van 7 taken (33 items). Om de verschillende typetaken een plaats te geven in de taalscreening, werd KOALA opgedeeld in twee delen: een deel met taken van het type gestandaardiseerde observaties (die individueel worden afgenomen) en een deel met doe- en zoekopdrachten en meerkeuze-opdrachten (die in groep worden afgenomen). We noemen de delen respectievelijk Deel A en Deel B.

2.1 Deel A

Een eerste deel bestaat uit drie gestandaardiseerde observatietaken die individueel worden afgenomen en al meteen een goed beeld geven van de taalvaardigheid van de kleuter.

In dit deel werden 17 items opgenomen, met een gemiddelde moeilijkheidsgraad van -.0,03. Deze gemiddelde moeilijkheidsgraad ligt net tussen cesuur A en cesuur B. De discriminerende waarde van de items in Deel A ligt met een gemiddelde van 1,11 en een kleine standaarddeviatie erg hoog.

Taak	Doelstelling	Typetaak	Itemnummer	Measure	Discriminerende waarde
Hoepel (7 items)	Instructies begrijpen	Doe-opdracht	1.1	-1,78	1,02
			1.2	-1,83	1,07
			1.3	-1,11	1,09
			1.4	0,76	1,02
			1.5	0,37	0,82
			1.6	1,48	1,20
			1.7	0,44	1,17
Turnles (6 items)	Informatieve mededelingen begrijpen	Doe-opdracht	2.1	0,08	1,21
			2.2	-1,62	0,99
			2.3	-0,86	1,06
			2.4	-0,02	1,25
			2.5	-0,58	1,17
			2.6	0,66	1,22
Vingerpop (4 items)	Instructies begrijpen	Doe-opdracht	3.1	1,65	1,06
			3.2	1,11	1,18
			3.3	1,34	1,17
			3.4	-0,58	1,12

			gemiddelde	-0.03 (SD 1.15)	1,11 (SD 0,11)

Tabel 61: Overzicht taken en items Deel A

Toetsdeel A begint met een taak met aantal gemakkelijke items, wat interessant is om een kleuter succes te laten ervaren. Binnen dezelfde taak zijn echter ook items opgenomen van een gemiddelde en hogere moeilijkheidsgraad. De tweede taak in Deel A, Turnles, bevat opnieuw enkele gemakkelijke items die de kleuter vertrouwen kunnen geven, en ook enkele items van een hogere moeilijkheidsgraad. De derde taak van Deel A, Vingerpop, is duidelijk de moeilijkste taak van Deel A.

Deel A bevat items van verschillende moeilijkheidsgraden. 5 items sluiten aan bij het luistervaardigheidsniveau van kleuters onder cesuur B, 6 bij het luistervaardigheidsniveau van kleuters tussen cesuur A en B, en 6 bij het luistervaardigheidsniveau van kleuters boven cesuur A. Enkele van deze items liggen dicht bij de cesuren A of B. De discriminerende waarde van bijna alle items ligt hoger dan 1.

We concluderen dat de items in Deel A mooi gespreid zijn over de luistervaardigheidsniveaus en de groepen die door de cesuren worden onderscheiden en in het algemeen goed discrimineren. Op die manier geeft Deel A al een vrij goed beeld van de kleuters vaardigheid en van de vaardigheid ten opzichte van de cesuren.

2.2 Deel B

Het tweede deel van de taalscreening bestaat uit 4 taken op papier, die afgenomen worden in kleine groep. In Deel B werden in totaal 16 items opgenomen, met een gemiddelde moeilijkheidsgraad van 0.47. Deze gemiddelde moeilijkheidsgraad ligt tussen cesuur A en cesuur B, en erg dicht bij cesuur A. De gemiddelde discriminatiegraad ligt net niet op 1, wat wijst op een groot discriminerend vermogen.

Taak	Doelstelling	Typetaak	Itemnummer	Measure	Discriminerende waarde
Deel Rommel eetzaal (5 items)	Instructies begrijpen	Zoek-opdracht	4.1	-0,95	0,92
			4.2	0,07	0,71
			4.3	0,82	1,54
			4.4	0,79	1,18
			4.5	1,42	1,18
Konijntjes (4 items)	Verhalen begrijpen	Kies-opdracht	5.1	2,42	0,47
			5.2	0,82	0,79
			5.3	-0,58	1,09
			5.4	0,54	1,11
Lievelingsboeken (3 items)	Vragen begrijpen	Kies-opdracht	6.1	-0,57	1,05
			6.2	0,63	1,10

			6.3	1,33	0,92
Juf is jarig (4 items)	Instructies begrijpen	Kies-opdracht	7.1	-0,34	0,97
			7.2	0,22	0,90
			7.3	0,72	0,74
			7.4	0,23	1,12
			gemiddelde	0,47 (SD 0,86)	0,99 (SD. 0,25)

Tabel 62: Overzicht taken en items Deel B

Toetsdeel B begint met een taak met twee gemakkelijke items, wat interessant is om een kleuter succes te laten ervaren. De andere taken in Deel B bevatten vooral items van een gemiddeld en hogere moeilijkheidsgraad.

Deel B bevat items van verschillende moeilijkheidsgraden. 9 items sluiten aan bij het luistervaardigheidsniveau van kleuters boven cesuur A, 6 bij het luistervaardigheidsniveau van kleuters tussen cesuur A en B, en 1 bij het luistervaardigheidsniveau van kleuters onder cesuur B. Twee van deze items liggen dicht bij A, twee erg dicht bij cesuur B. De discriminerende waarde van bijna alle items ligt hoger dan 0.9.

We kunnen concluderen dat de items in Deel B mooi gespreid zijn over de gemiddelde en hogere luistervaardigheidsniveaus. Op die manier kan Deel B goed nagaan of het luistervaardigheidsniveau van de kleuters boven cesuur A of tussen cesuur A en B ligt en of de kleuters dus tot de oranje of de groene groep behoren.

3 ▪ Controle van de selectie

De verschillende psychometrische, inhoudelijke en vormelijke criteria voor de geselecteerde taken en items werden opgenomen in Tabel 63: het taaknummer, de moeilijkheidsgraad (measure), categorisering ten opzichte van de cesuren (kleurcodes groen, oranje, rood), de outfitmnsqr als indicator van misfit, de discriminerende waarde, de aanwezigheid van DIF, de in kaart gebrachte doelen en het taaktype.

Nr.	Naam	Meas	Misfit	Discr	DIF	Doel	Typetaak
6.1	Hoepel	-1,78	1,04	1,02		instructies begrijpen	doe-opdracht
6.2	Hoepel	-1,83	0,64	1,07		instructies begrijpen	doe-opdracht
6.3	Hoepel	-1,11	0,67	1,09		instructies begrijpen	doe-opdracht
6.4	Hoepel	0,76	0,94	1,02		instructies begrijpen	doe-opdracht
6.5	Hoepel	0,37	1,13	0,82		instructies begrijpen	doe-opdracht
6.6	Hoepel	1,48	0,93	1,20		instructies begrijpen	doe-opdracht
6.7	Hoepel	0,44	0,81	1,17		instructies begrijpen	doe-opdracht
8.1	Juf is jarig	-0,34	1,09	0,97		instructies begrijpen	Kies-opdracht
~~8.2~~	~~juf is jarig~~	~~0,78~~	~~1,05~~	~~0,91~~	~~X~~	~~instructies begrijpen~~	~~kies-opdracht~~
8.3	Juf is jarig	0,22	1,01	0,90		instructies begrijpen	kies-opdracht
8.4	Juf is jarig	0,72	1,14	0,74		instructies begrijpen	kies-opdracht
8.5	Juf is jarig	0,23	0,85	1,12		instructies begrijpen	kies-opdracht
12.1	Konijntjes	2,42	1,48	0,47		verhaal begrijpen	kies-opdracht
12.3	Konijntjes	0,82	1,13	0,79		verhaal begrijpen	kies-opdracht
12.4	Konijntjes	-0,58	0,84	1,09		verhaal begrijpen	kies-opdracht
12.5	Konijntjes	0,54	0,93	1,11		verhaal begrijpen	kies-opdracht
~~12.6~~	~~Konijntjes~~	~~1,23~~	~~1,41~~	~~0,38~~	~~X~~	~~verhaal begrijpen~~	~~kies-opdracht~~
13.1	Lievelingsboeken	-0,57	0,92	1,05		vragen begrijpen	kies-opdracht
13.2	Lievelingsboeken	0,63	0,92	1,10		vragen begrijpen	kies-opdracht
~~13.3~~	~~Lievelingsboeken~~	~~0,51~~	~~0,88~~	~~1,21~~	~~X~~	~~vragen begrijpen~~	~~kies-opdracht~~
13.4	Lievelingsboeken	1,33	1,06	0,92		vragen begrijpen	kies-opdracht
20.1	Rommel in de eetzaal	-0,95	1,19	0,92		instructies begrijpen	zoekopdracht
20.2	Rommel in de eetzaal	0,07	1,36	0,71	X	instructies begrijpen	zoekopdracht
20.3	Rommel in de eetzaal	0,82	0,67	1,54	X	instructies begrijpen	zoekopdracht
20.4	Rommel in de eetzaal	0,79	0,88	1,18		instructies begrijpen	zoekopdracht
20.5	Rommel in de eetzaal	1,42	0,89	1,18		instructies begrijpen	zoekopdracht

24.1	Turnles	0,08	0,79	1,21		Info. mededeling begrijpen	doe-opdracht
24.2	Turnles	-1,62	1,29	0,99		Info. mededeling begrijpen	doe-opdracht
24.3	Turnles	-0,86	1,00	1,06		Info. mededeling begrijpen	doe-opdracht
24.4	Turnles	-0,02	0,71	1,25		Info. mededeling begrijpen	doe-opdracht
24.5	Turnles	-0,58	0,68	1,17		Info. mededeling begrijpen	doe-opdracht
24.6	Turnles	0,66	0,84	1,22		Info. mededeling begrijpen	doe-opdracht
28.1	Vingerpop	1,65	1,00	1,06		instructies begrijpen	doe-opdracht
28.2	Vingerpop	1,11	0,89	1,18		instructies begrijpen	doe-opdracht
28.3	Vingerpop	1,34	0,91	1,17		instructies begrijpen	doe-opdracht
28.4	Vingerpop	-0,58	0,83	1,12		instructies begrijpen	doe-opdracht

Tabel 63: Overzicht van psychometrische, inhoudelijke en vormelijke criteria

In de volgende paragrafen worden de verschillende criteria voor de geselecteerde taken en items besproken.

3.1 Psychometrische criteria

3.1.1 Verdeling over de vaardigheidsniveaus

Voor het volledige instrument ligt het zwaartepunt op items in de groene en oranje zone omdat we met dit instrument zo betrouwbaar mogelijk willen meten of een kleuter al dan niet in de groene zone zit (boven cesuur A) of in de oranje zone (tussen cesuur A en cesuur B). Kleuters moeten dus maximaal de kans krijgen om hun luistervaardigheidsniveau te tonen aan de hand van items met een vaardigheidsniveau uit de groene en/of oranje zone.

aantal items groene zone	15
aantal items oranje zone	12
aantal items rode zone	6
totaal aantal items	33
laagste measure	-1,83
hoogste measure	2,42
gemiddelde measure	0,21
Standaarddeviatie	1,03

3.1.2 Discriminerende waarde

De geselecteerde items en taken hebben zoals, vooropgesteld, een discriminerende waarde van .70. Een uitzondering daarop is een item van de taak Konijntjes. Aangezien we slechts uit een beperkt aantal geschikte verhaaltaken en –items konden kiezen, maakten we de afweging tussen discriminatiewaarde (betrouwbaarheid) en aantal items voor een doelstelling en taak (inzetbaarheid). Uiteindelijk hebben we beslist om dit item te behouden ondanks de lagere discriminatiegraad om zo toch de doelstelling 'verhalen begrijpen' uitgebreider in kaart te kunnen brengen.

3.1.3 DIF

Items die DIF bevatten werden uit de geselecteerde items gefilterd. Op die manier bevat de KOALA geen enkel item dat anders functioneert op papier of tablet, of voor kleuters met bepaalde achtergrondkenmerken. Twee items uit 'Rommel in de eetzaal' werden behouden ondanks hun significante DIF-waarden voor gezinstaal omdat het effect van deze DIF voor het ene item in het voordeel was van de kleuters die thuis Nederlands praatten, de andere keer in het voordeel van de kleuters uit een meertalige gezinsomgeving, en op die manier elkaar als het ware opheffen. Door beide items te behouden, konden we vijf items voor dit taaktype selecteren. In deze overweging namen we ook mee dat deze taak als enige behoorde tot de typetaak 'doe- en zoekopdracht'.

3.1.4 Testinformatie

Voor de geselecteerde items gaan we na a.d.h.v. de testinformatie functie na waar de piek in de informatiewaarde zich situeert. Idealiter situeert de piek in de informatiewaarde zich op of vlakbij de cesuren: we willen immers zoveel mogelijk informatie hebben rond die moeilijkheidsgraad, opdat we met grote zekerheid kunnen aangeven dat een kleuter zich boven of onder een van de cesuren bevindt.

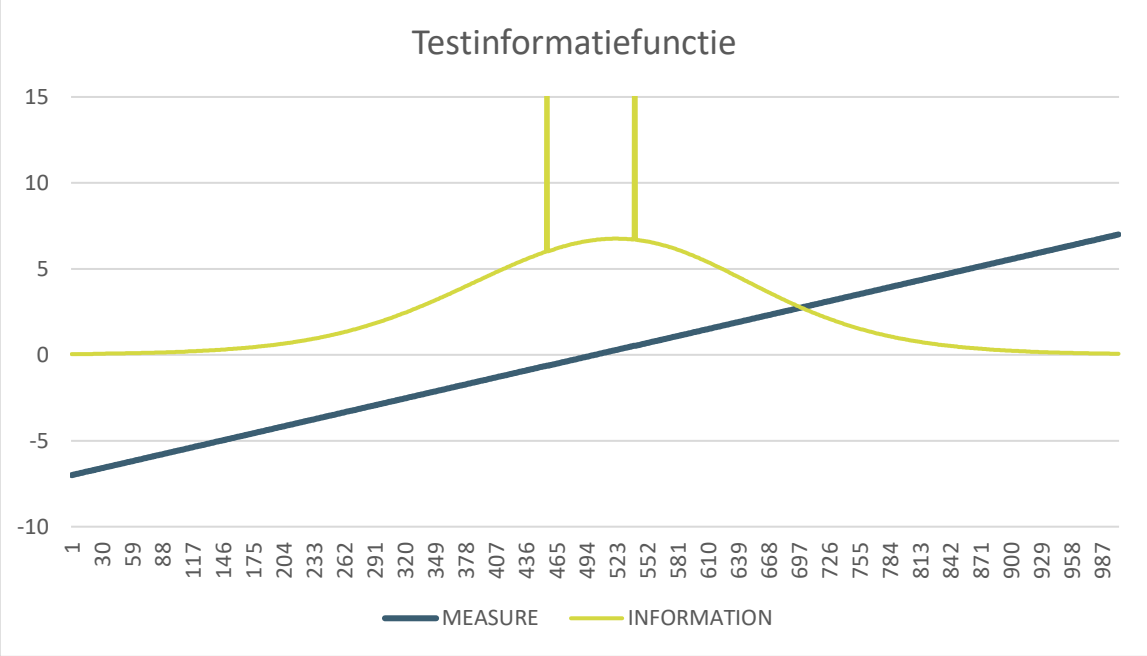
Tabel 64 vat de testinformatiewaarden samen die relevant zijn voor ons instrument, namelijk de piek de en verhouding tot de beide cesuren.

	MEASURE	SCORE	INFORMATION	S.E.M.
KOALA				
Piek informatiewaarde	0,31	17,16	**6,76**	0,38
Cesuur A	0,52	18,58	6,7	0,39
Cesuur B	-0,63	11,06	6,02	0,41

Tabel 64: Testinformatiewaarden KOALA

Als we kijken naar de informatiewaarde van het instrument stellen we vast dat de piek (zeer) kort achter cesuur A valt. Dit betekent dat het instrument vooral helpt om kleuters te identificeren die wel/niet boven cesuur A liggen, m.a.w. of een kleuter in de groene of oranje zone terecht komt. Ook de informatiewaarde in de buurt van cesuur B ligt nog steeds erg hoog. Het standaard instrument geeft dus ook betrouwbare informatie om te bepalen of een kleuter in de oranje of rode zone terechtkomt. Dit komt overeen met wat we wilden bereiken met deze aangepaste versie, met name een goed beeld krijgen van de kleuters die ondersteuning nodig hebben en de mate van ondersteuning die zij nodig hebben. .

Deze verhouding tussen de cesuren en de informatiewaarden wordt ook nog eens visueel voorgesteld in FIGUUR 34. De blauwe lijn stelt de lineair stijgende moeilijkheidsgraad van de items voor. De oranje curve geeft aan hoe de informatiewaarde van de test zich verhoudt tot de moeilijkheidsgraad van de items. Elke oranje verticale lijn situeert een cesuur ten opzichte van de informatiewaarde van de test. De linkse groene lijn geeft aan welke informatiewaarde de test heeft ter hoogte van de moeilijkheidsgraad van cesuur B. De rechtse verticale lijn geeft de informatiewaarde van de test aan ter hoogte van de moeilijkheidsgraad van cesuur A.



Figuur 34: Testinformatiefunctie KOALA ten opzichte van Cesuur A en B

De visuele voorstelling geeft eveneens aan dat de beide cesuren zich bevinden in het 'topje' van de informatiewaarde van de test. Met andere woorden: de test geeft de meeste informatiewaarde rond de measure die overeenkomt met cesuur A en cesuur B. We stellen vast dat de piek in de informatiewaarde tussen cesuur A en B valt. De informatiewaarde rond cesuur A is lichtjes hoger dan deze rond cesuur B.

3.2 Inhoudelijke criteria

3.2.1 Verdeling over de doelstellingen

Alle vooropgestelde doelstellingen komen aan bod in de voor de KOALA geselecteerde items. De doelstelling ‘instructies begrijpen’ komt het meest aan bod (20 items). Vervolgens komen ‘informatieve mededelingen begrijpen’ (6 items), ‘verhalen begrijpen’ (4 items) en ‘vragen begrijpen’ (3 items) .

3.2.2 Verdeling over de taaktypes

De verschillende taaktypes komen aan bod: er zijn drie taken met gestandaardiseerde observaties (Hoepel, Turnles en Vingerpop, in totaal 17 items), 1 doe-en zoekopdracht (Rommel in de eetzaal, 5 items), en 3 meerkeuze-opdrachten (Lievelingsboeken, Juf is jarig en Konijntjes, in totaal 11 items).

3.3 Vormelijke criteria

3.3.1 Aantal taken en items

We selecteerden 7 taken van in totaal 33 items, wat iets meer is als het aantal taken in het kalibratie-onderzoek. Op die manier kunnen we voldoende betrouwbaar uitspraken doen over kleuters, en kunnen we een leraar ook informatie geven over de verschillende doelstellingen die opgenomen werden in de toetsmatrijs.

3.3.2 Aantal items per taak

De vooropgestelde benedengrens is gehaald: elke taak bestaat uit minstens drie items. Zes van de zeven taken bestaan zelfs uit minstens vier items. Op die manier blijft KOALA voldoende efficiënt.

4 ▪ Aangepaste versie voor kleuters met grote ondersteuningsnoden

Uit het kalibratieonderzoek bleek een zeer beperkt aantal kleuters (1,2%, 1.8.2 KLEUTERS MET EEN (ZEER) LAGE TAALVAARDIGHEID) minder dan 20% van de toetsitems correct te kunnen oplossen. Om te vermijden dat voor deze kleuters de taalscreening een frustrerende ervaring wordt, stellen we een aangepaste versie voor kleuters met een zeer lage taalvaardigheid voor. Meer bepaald bieden we kleuterleraren met kleuters die heel weinig correct kunnen oplossen uit het eenvoudigste deel van de toets (eerste deel van Deel A) de mogelijkheid om te stoppen (afbreken). Om de leraar toch nog zinvolle informatie over deze kleuter te geven, mogen deze kleuters een aangepaste versie (de sterretjesversie) maken.

Dit afbreken blijft een uitzondering en is erop gericht om demotivatie, frustraties en afhaken te vermijden bij de kleine groep kleuters die erg zwak scoren. Tegelijkertijd willen we ook van deze kleuters het beeld verder verfijnen om zicht te krijgen op hun precieze ondersteuningsnoden. We concretiseren dit in een afbreekpunt én een selectie van extra toetsitems.

4.1 Plaatsen van het afbreekpunt

De eerste twee taken in Deel A hebben een gemakkelijke of gemiddelde moeilijkheidsgraad, om kleuters op deze manier aan te moedigen en gemotiveerd te krijgen voor de taalscreening. Deze opbouw biedt bovendien ook de mogelijkheid om kleuters die het al vanaf het begin erg moeilijk hebben met deze gemakkelijk items, snel te detecteren.

Als we de volledige toets opsplitsen in een eerste deel (taak 1-2) en tweede deel (taak 3-7), krijgen we het volgende beeld:

Taak 1-2	
aantal items groene zone	3
aantal items oranje zone	5
aantal items rode zone	5
totaal aantal items	13
laagste measure	-1,83
hoogste measure	1,48
gemiddelde measure	-0,31
Standaarddeviatie	1,07

Taak 3-7	
aantal items groene zone	12
aantal items oranje zone	7
aantal items rode zone	1
totaal aantal items	20
laagste measure	-0,95
hoogste measure	2,42
gemiddelde measure	0,55
standaarddeviatie	0,87

Bovenstaande gegevens bevestigen dat een afbreekpunt na de eerste twee taken zinvol is. Deze twee taken ('Hoepel' en 'Turnles') geven ons voldoende informatie om in te kunnen beslissen of een kleuter de screening zonder grote moeilijkheden verder kan zetten of doorverwezen kan worden naar een aangepaste versie voor kleuters met grote ondersteuningsnoden.

4.2 Extra toetsitems voor Deel B*

We komen tot een selectie van 4 taken (16 items) die voldoen aan de criteria die hierboven werden beschreven. We kiezen er bewust voor om deze groep één taak minder aan te bieden, omdat het voor hen moeilijker is om lang geconcentreerd te blijven.

○ Overzicht taken en items

Taak	Doelstelling	Typetaak	Itemnummer	Measure	Discriminerende waarde
Speeltijd (5 items)	Informatieve mededelingen begrijpen	Zoek-opdracht	3.1*	-0,98	1,00
			3.2	-0,03	1,18
			3.3	-1,09	1,17
			3.4	-0,72	1,18
			3.5	-0,70	1,20
Zandtafel (3 items)	Vragen begrijpen	Zoek-opdracht	4.1	-0,64	0,86
			4.2	-1,45	0,97
			4.3	0,50	0,74
Kabouters (4 items)	Verhalen begrijpen	Kies-opdracht	5.1	-0,74	0,97
			5.2	-0,83	0,98
			5.3	0,68	0,74
			5.4	-0,01	0,95

Boekenhoek (4 items)	Vragen begrijpen	Kies-opdracht	6.1	-0,80	1,00
			6.2	0,00	1,18
			6.3	-0,13	1,15
			6.4	0,30	0,99

5 ▪ Controle van de selectie voor de aangepaste versie

5.1 Psychometrische criteria

5.1.1 Verdeling over de vaardigheidsniveaus

Als we deze extra toetsitems samenvoegen met het gemeenschappelijk gedeelte (taak 1-2) verkrijgen we een instrument waarbij het zwaartepunt ligt op items in de oranje en rode zone. Deze versie van het instrument maakt het mogelijk om op een betrouwbare manier te meten of een kleuter die grote moeilijkheden ervaart tijdens taak 1-2 daadwerkelijk in de rode zone zit (onder cesuur B). Let wel: als het taalvaardigheidsniveau van de kleuter toch hoger blijkt te zijn dan gesuggereerd door de afname van taak 1 en 2, dan kan de kleuter met deze versie nog in de groene of oranje zone terechtkomen en kan op die manier het foute beeld worden bijgesteld.

Het zwaartepunt in de extra selectie taken (taak 3*-6*) voor kleuters met grote ondersteuningsnoden ligt dan ook op items in de rode en oranje zone. De extra selectie laat toe om het beeld van deze groep kleuters verder te verfijnen.

Taak 1-2	
aantal items groene zone	3
aantal items oranje zone	5
aantal items rode zone	5
totaal aantal items	13
laagste measure	-1,83
hoogste measure	1,48
gemiddelde measure	-0,31
Standaarddeviatie	1,07
standaarddeviatie	0,83

Taak 3*-6*	
aantal items groene zone	1
aantal items oranje zone	6
aantal items rode zone	9
totaal aantal items	16
laagste measure	-1,45
hoogste measure	0,68
gemiddelde measure	-0,42
standaarddeviatie	0,61

5.1.2 Discriminerende waarde

Alle items in deze extra selectie hebben een voldoende discriminerende waarde. Geen enkel item heeft een discriminerende waarde die lager ligt dan .70.

Discriminerende waarde	Geschikt voor instrument?	Aantal items

>= 1	Zeer geschikt	8 items
>= 0,70 en < 1	Voldoende geschikt	8 items
< 0,70	Minder geschikt	0 items

5.1.3 DIF

Geen enkel item in deze aangepaste selectie heeft een significant DIF-contrast.

5.1.4 Testinformatie

Voor de geselecteerde items gaan we na a.d.h.v. de testinformatie functie na waar de piek in de informatiewaarde zich situeert. Idealiter situeert de piek in de informatiewaarde zich op of vlakbij de cesuren: we willen immers zoveel mogelijk informatie hebben rond die moeilijkheidsgraad opdat we met grote zekerheid kunnen aangeven dat een kleuter zich boven of onder een van de cesuren bevindt.

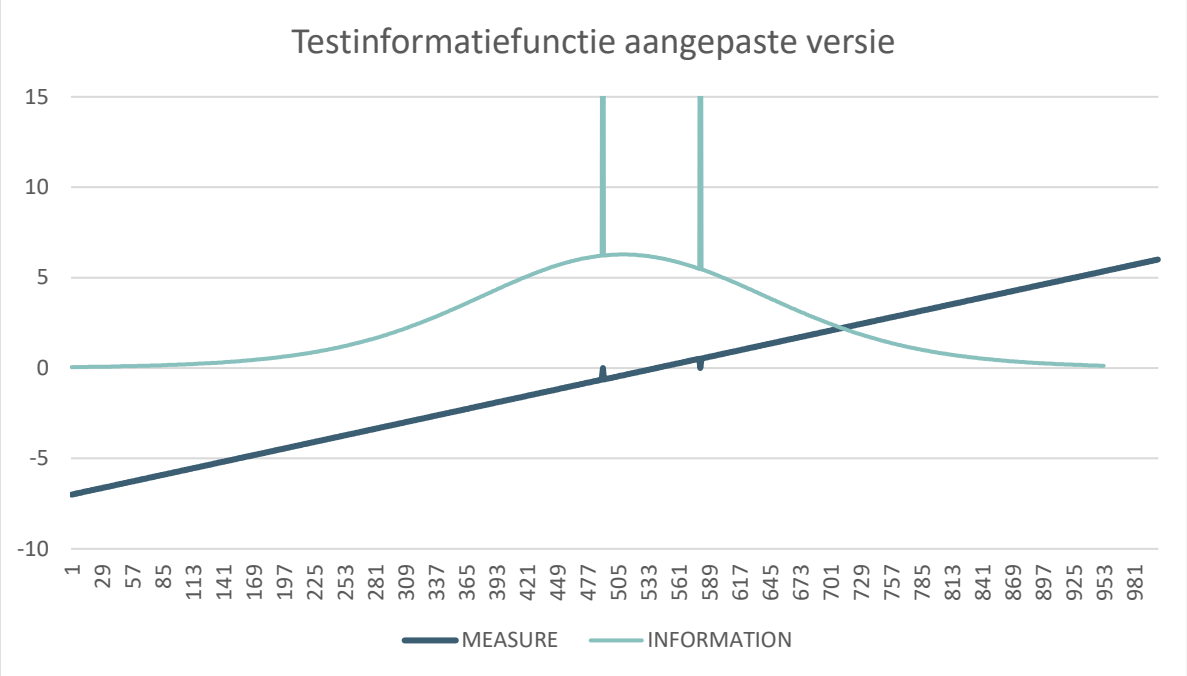
Tabel 65 vat de testinformatiewaarden samen die relevant zijn voor het aangepaste instrument, namelijk de piek en de verhouding tot de beide cesuren.

	MEASURE	SCORE	INFORMATION	S.E.M.
KOALA (AANGEPASTE VERSIE)				
Piek informatiewaarde	-0,34	14,79	6,28	0,4
Cesuur A	0,52	19,93	5,48	0,43
Cesuur B	-0,63	13	6,22	0,4

Tabel 65: Testinformatiefunctie aangepaste versie

Voor de aangepaste versie valt de piek in de informatiewaarde zo goed als samen met cesuur B. Dit betekent dat dit gedeelte van de screening zich uitstekend leent om kleuters te identificeren die boven/onder cesuur B liggen. Zoals vooropgesteld kunnen we met dit gedeelte een verfijnder beeld krijgen van kleuters met een erg lage score op het gemeenschappelijk gedeelte.

De verhouding tussen de cesuren en de informatiewaarden worden ook nog eens visueel voorgesteld in Figuur 35. De blauwe lijn stelt de lineair stijgende moeilijkheidsgraad van de items voor. De grijze curve geeft aan hoe de informatiewaarde van de test zich verhoudt tot de moeilijkheidsgraad van de items. Elke grijze verticale lijn situeert een cesuur ten opzichte van de informatiewaarde van de test. De linkse lichtblauwe verticale lijn geeft aan welke informatiewaarde de test heeft ter hoogte van de moeilijkheidsgraad van cesuur B. De rechtse verticale lijn geeft de informatiewaarde van de test aan ter hoogte van de moeilijkheidsgraad van cesuur A.



Figuur 35: Testinformatiefunctie aangepaste versie

De visuele voorstelling geeft aan dat de beide cesuren zich bevinden in het 'topje' van de informatiewaarde van de test. Met andere woorden: de test geeft de meeste informatiewaarde rond de measure die overeenkomt met cesuur A en cesuur B. De informatiewaarde rond cesuur B bevindt zich zelfs mooi op het topje van de curve: rond deze waarde geeft de aangepaste versie van de toets de meeste informatie.

Dit betekent dat de aangepaste versie van de screening zich er goed toe leent om kleuters te identificeren die boven/onder cesuur B liggen. Zoals vooropgesteld kunnen we met dit gedeelte een verfijnder beeld krijgen van kleuters met een erg lage score op het gemeenschappelijk gedeelte.

5.2 Inhoudelijke criteria

5.2.1 Verdeling over de doelstellingen

In deze aangepaste versie met afbreekpunt komen alle vier de vooropgestelde doelstellingen aan bod.

De doelstelling ' 'informatieve mededelingen begrijpen' (11 items) komt het meest aan bod. Ook de doelstellingen 'vragen begrijpen' (7 items) en 'instructies begrijpen' (7 items) komen veel voor. Tot slot zijn er ook voor deze aangepaste versie enkele items geselecteerd voor de doelstelling 'verhalen begrijpen' (4 items).

5.2.2 Verdeling over de typetaken

Het gemeenschappelijk gedeelte (taak 1-2) bestaat uit twee doe-opdrachten. Het extra gedeelte uit twee doe- en zoek-opdrachten en twee meerkeuzeopdrachten. Ook in deze aangepaste versie kunnen we spreken van een evenwichtige verdeling over de verschillende typetaken.

5.3 Vormelijke criteria

5.3.1 Aantal taken en items

We selecteerden 6 taken met in totaal 29 items, wat iets korter is als het aantal taken in het kalibratie-onderzoek. Op die manier kunnen we voldoende betrouwbare uitspraken doen over kleuters, en kunnen we een leraar ook informatie geven over de verschillende doelstellingen die opgenomen werden in de toetsmatrijs.

5.3.2 Aantal items per taak

Deze aangepaste versie telt minstens 3 items (Zandtafel) en maximum 7 items (Hoepel) per taak. Daarmee is er een goede balans tussen efficiëntie en haalbaarheid voor de kleuters.

6 ▪ Besluit

Met de vooropgestelde brede criteria in het achterhoofd (validiteit, betrouwbaarheid, efficiëntie en inzetbaarheid) werden de verschillende items die uitgeprobeerd werden tijdens het kalibratie-onderzoek gescreend voor hun bruikbaarheid als taalscreeningsinstrument voor de kleuterklas.

Op basis van deze screening werd een selectie gemaakt van taken en items voor KOALA, die aan de vooropgestelde criteria blijkt te voldoen. Daarnaast werd ook een voorstel voor een aangepaste versie gedaan, die ingezet kan worden in het kader van een zinvolle screening van de allerzwakste kleuters.

HOOFDSTUK 10: BELEIDSAANBEVELINGEN

Op basis van het gevoerde onderzoek formuleren we enkele aanbevelingen voor de Vlaamse overheid en voor verder onderwijsonderzoek. Deze aanbevelingen hebben enerzijds betrekking op de taalscreening zelf en de evolutie ervan doorheen de jaren, en anderzijds op de implementatie van het instrument.

(1) Volgend onderzoek rond de taalscreening zelf strekt tot aanbeveling:

- Het is belangrijk om een taalscreening steeds up-to-date te houden met recent referentiemateriaal. De prestaties van individuen, klassen en scholen worden in deze screening afgezet ten opzichte van de prestaties van individuen, klassen en scholen zoals deze waren in de periode oktober-december 2021. Enkel als op regelmatige basis nieuw referentie-materiaal wordt verzameld voor deze toets, zullen scholen en leraren de validiteit van de taalscreening en de bijbehorende cesuren voldoende naar waarde blijven schatten. Een nieuwe referentie bepalen voor de taalscreening in de nabije toekomst is extra van belang omdat de dataverzameling plaatsvond op het moment dat kleuters, leraren en scholen al meer dan een half jaar functioneerden binnen de restricties van corona (met afstandsonderwijs, verlengde vakanties, en verarmde mogelijkheden tot interactie tot gevolg); de precieze impact van corona blijft – in het bijzonder voor deze leeftijdsgroep - erg onduidelijk.
- Breed geïmplementeerde toetsen zoals deze taalscreening lopen een risico op veroudering: items lijken minder relevant, de maatschappij verandert, het onderwijs verandert. Een efficiënte, betrouwbare en goedkope manier om veroudering tegen te gaan is het 'zaaien' van items - waarbij extra taken of items worden toegevoegd en gekalibreerd door deze als proefitem op te nemen in het bestaande toetsinstrument. Het is dan ook aangewezen om stappen te zetten in functie van het 'zaaien' van items voor deze taalscreening.
- De cesuren bij KOALA werden bepaald door experten op basis van hun verwachtingen en ervaringen rond de groei van vijfjarige kleuters doorheen de derde kleuterklas. Wij hebben in Vlaanderen momenteel geen gegevens over de gemiddelde talige vooruitgang van kleuters op vijf jaar. Onderzoek dat de leerwinst van kleuters doorheen de derde kleuterklas meet, kan de gelegde cesuren (en/of de interpretatie ervan) bij KOALA bevestigen of, indien nodig, geïnformeerd bijsturen.
- Binnen het onderzoeksopzet bij deze taalscreening was het niet mogelijk om de predictieve validiteit van KOALA te onderzoeken; nagaan in welke mate een individuele leerlingenuitkomst op de KOALA voorspellend is voor de prestaties Nederlands en/of het onderwijssucces van een leerling, viel met andere woorden buiten de opdracht. Het nagaan van de predictieve validiteit van het instrument zou de waarde van taalscreenings in het algemeen en van dit instrument in het bijzonder kunnen versterken.

(2) Op vlak van de implementatie van het instrument legde ons onderzoek de volgende noden bloot:

- Kleuterleraren zijn blij met de bevestigingen en aanvullingen die de taalscreening hen biedt. Tegelijkertijd leven er grote noden en vragen rond de acties en plannen die volgen op de afname en de resultaten. Eenmaal de afname achter de rug is, hebben kleuterleraren met andere woorden behoefte aan concrete en praktijkgerichte handvatten over de manier waarop zij de taalstimulering concretiseren voor de kleuters die door de taalscreening geïdentificeerd worden als risicokleuters.
- Met KOALA wordt voor het eerst een ondersteuningsbeleid aangestuurd door een verplichte gecentraliseerde screening. Dit is uniek in Vlaanderen (en wereldwijd ook vrij uniek) voor leraren bij zulke jonge leerders. Het is daarom in het belang van het onderwijsbeleid om de implementatie van de KOALA en de manier waarop het eraan gekoppelde ondersteuningsbeleid vorm krijgt, nauw op te volgen. Goede inzichten in de implementatiepraktijk van KOALA kunnen niet alleen ingezet worden om het gebruik ervan in de toekomst bij te sturen en te optimaliseren, maar ook worden gebruikt ter inspiratie bij de implementatie van centrale toetsen in het lager onderwijs.

LITERATUURLIJST

Alderson, J. C., Clapham, C., and Wall, D. (1995). *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press.

Bailey, E. (2017). Theoretical and developmental issues to consider in the assessment of young learners English language proficiency. In K.M. Wolf & Y.B. Butler (Eds.), *English proficiency assessment for young leaners.* New York and London: Routledge

Bagnato, S.J. (2007). *Authentic assessment for early childhood intervention: Best practices*. New York: Guilford Press.

Bijlsma, B. (2004). *Het kiezen van een kindvolgsysteem*. Leeuwarden: Partoer.

Blessing, E. (2019). *Appropriate Uses of Early Assessments. National Association for the Education of Young Children.*

Brassard, M., & Boehm, A. *Preschool Assessment: Principles and Practices*. Guilford Publications, 2007.

Burke Hadley, E., & Dickinson, D. K. (2020). Measuring young children's word knowledge: A conceptual review. *Journal of Early Childhood Literacy*. Vol. 20(2) 223–251. DOI: 10.1177/1468798417753713

Byers-Heinlein, K., & Lew-Williams, C. (2013). Bilingualism in the Early Years: What the Science Says. *LEARNing landscapes*, 7(1), 95–112.

Cameron, L. (2001). *Teaching languages to young learners*. Ernst Klett Sprachen.

Conboy, B., & Montanari, S. (2016). Early Lexical Development in Bilingual Infants and Toddlers. In E. Nicoladis & S. Montanari (Eds.), *Lifespan perspectives on bilingualism: Factors moderating language proficiency*. Washington D.C.

Colpin, M., Gysen, S., Jaspaert, K., Heymans, R., Van den Branden, K., & Verhelst, M. (2006). *Studie naar de wenselijkheid en haalbaarheid van de invoering van centrale taaltoetsen in Vlaanderen in functie van gelijke onderwijskansen.* Leuven: Centrum voor Taal en Onderwijs.

Conti-Ramsden, G., & Durkin, K. (2012). Language Development and Assessment in the Preschool Period. *Neuropsychol Rev* (2012) 22:384–401. DOI 10.1007/s11065-012-9208.

Dale, B. , Mcintosh, D., & Rothlisberg, B. (2011) . Profile analysis of the kaufman assessment battery for children, second edition, with african american and caucasian preschool children. *Psychology in the Schools, Vol. 48(5),* 476-487.

Dean, J. B. (1997). The washback effect of language tests. University of Havai'i *Working Papers in ESZ*, Vol. 16, No. l, p.2745.

De Fraine, B. (2003) *The Effect of Schools and Classes on Language Achievement*. British Educational Research Journal, Vol. 29(6):841-859.

De Mayer, S., & Rymenans, R. (2004) *Onderzoek naar kenmerken van effectieve scholen*. Academia Press.

Deygers, B. (2019). Fairness and social justice in English language assessment. *Second handbook of English language teaching*, 541-569.

Djalal, FM., Ameel, E., & Storms, G. (2016). The Typicality Ranking Task: A New Method to Derive Typicality Judgments from Children. *PLoS ONE* 11(6): e0157936.

Dockrell, J.E., & Marschall, C. (2015). Measurement Issues: Assessing language skills in young children. *Child and Adolescent Mental Health 20*, No. 2, pp. 116–125 doi:10.1111/camh.12072.

Fellowes, J., & Oakley, G. (2010). *Language, Literacy and Early Childhood Education*. Victoria: Oxford University Press; 604 pp.

Fiore, Lisa B. (2012). *Assessment of Young Children. A Collaborative Approach*. Taylor and Francis Group.

Frijns, C., & Jaspaert, K. (2016). *Kleuters aan het woord. Een taakgericht neologismenverhaal voor taalonderzoek bij niet-Nederlandstalige kleuters*. Leuven: Centrum voor Taal en Onderwijs.

Fijns, C. (2017). A*ls we 't de kinderen vragen. Het potentieel van productieve interactie voor tweedetaalverwerving vanaf het prille begin*. KU Leuven.

Fulcher, G., & Davidson, F. (2012). *The Routledge Handbook of language testing*. Routledge. London and New York.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. Developmental Psychology, 40(2), 177–190. https://doi.org/10.1037/0012-1649.40.2.177

Goorhuis-Brouwer, S., & de Bo, K. (2005). Heeft vroeg vreemde-talenonderwijs een negatief effect op de Nederlandse taalontwikkeling van kinderen? *Levende Talen,* Vol.5, 3-7.

Gysen, S., K., Rossenbacker & M. Verhelst (1999). *KOBI-TV. Kleuterobservatie-instrument Taalvaardigheid*. Leuven: Centrum voor Taal en Migratie, Steunpunt NT2.

Harsch, C., & Hartig, C. (2016). *Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills*. Language Testing Vol. 33(4) 555–575.

Hasselgreen, A., & Caudwell, G. (2000). *Assessing the language of young learners,* British Council Monographs, Vivien Berry & Barry O'Sullivan (series ed.).

Heugh, K., (2009a). Contesting the monolingual practices of a bilingual to multilingual policy. *English Teaching: Practice and Critique,* 8, 96–113.

Hill, K., & McNamara, T. (2011). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing* Vol. 29(3) 395–420.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology* 45: 337–374.

Hudziak, J., & Archangeli, C. (2017). The Future of Preschool Prevention, Assessment, and Intervention. Child Adolesc Psychiatric Clin N Am 26 (2017) 611–624. http://dx.doi.org/10.1016/j.chc.2017.02.010

Ioannou-Georgiou, S. (2003). *Assessing Young Learners (Resource Books for Teachers*). Oxford University Press, USA.

Kane, M. (2010). Validity and fairness. *Language testing*, 27(2), 177-182.

Law, J., Boyle, J. Harris, F., Harkness, A., & Nye, C. (2000). The feasibility of universal screening for primary speech and language delay: findings from a systematic review of the literature. *Developmental Medicine & Child Neurology,* 42: 190–200.

Leysen, H., & Schraeyen, K. (2020). Testafname bij meertalige kinderen. In J. de Waal-Bogers (red.) *Taaldiagnostiek bij kinderen* (pp. 83-100). Pearson.

Linacre, M. (2012). *Winsteps Rasch Tutorial 3: Standard error and reliability*. https://winsteps.com/a/winsteps-tutorial-3.pdf

Łockiewicz, M., Sarzała-Przybylska, Z., & Lipowska, M. (2018). Early predictors of learning a foreign language in pre-school - Polish as a first Language, English as a foreign language. *Frontiers in Psychology*, 9, 1-11. https://doi.org/10.3389/fpsyg.2018.01813

Lonigan, J., Allan, N., & Lerner, M. (2011). Assessment of Preschool Early Literacy Skills: Linking Children's Educational Needs with Empirically Supported Instructional Activities, *Psychol Sch. 2011 May 1;* 48(5): 488–501. doi:10.1002/pits.20569.

Losardo, A., & Notari-Syverson, A. (2001). *Alternative approaches to assessing young children*. Baltimore, MD: Brookes.

McKay, P. (2005). *Research into the assessment of school-age language learners*. Annual Review of Applied Linguistics, 25, 243-263.

Manual, A. L. T. E. (2011). *Manual for Language Test Development and Examining. For use with the CEFR*. Produced by ALTE [Association of Language Testers in Europe] on behalf of the Language Policy Division, Council of Europe.

Puglisi, M.L., Hulme, C., Hamilton, L.G., & Snowling, M.J. (2017). The Home Literacy Environment Is a Correlate, but Perhaps Not a Cause, of Variations in Children's Language and Literacy Development*, Scientific Studies of Reading*, 21:6, 498-514, DOI:10.1080/10888438.2017.1346660

McAfee, O. (2004). *Basics of Assessment. A Primer for Early Childhood Educators*. National Association for the Education of Young Children.

Meisels, S.J., Xue, Y., & Shamblott, M. (2008). Assessing language, literacy, and mathematics skills with work sampling for Head Start. *Early Education and Development*, 19(6), 963-981.

Messick, S. (1990). *Validity of Test Interpretation and Use. Educational Testing Service*. Princeton. New Jersey.

Miller, D., Bayram, F., Rothman, J., & Serratrice, L. (2018). *Bilingual Cognition and Language : The State of the Science Across Its Subfields.* John Benjamins Publishing Company.

OESO (2004). *Learning for Tomorrow's World. First Results from PISA 2003*. Geraadpleegd op 21 september 2020.

Popham, W. J. (1997). *Consequential validity: Right concern-wrong concept. Educational measurement: Issues and practice,* 16(2), 9-13.

Oller, K., & Eilers, E. (2002). *Language and Literacy in Bilingual Children.* Channel View Publications Ltd.

Ortiz, M., Folsom, J.S., Al Otaiba, S., Greulich, L., Thomas-Tate, S., & Connor, C.M. (2012). The Componential Model of Reading: Predicting First Grade Reading Performance of Culturally Diverse Students From Ecological, Psychological, and Cognitive Factors Assessed at Kindergarten Entry. *Journal of Learning Disabilities*. Vol.45 (5), p.406-417.

Peleman, B., Vandenbroeck, M., & Van Avermaet, P. (2019). *De overgang naar de kleuterschool voor kinderen uit gezinnen in armoede.* UGent.

Peng, P., Lin, X., Ünal, Z.E., Lee, K., Namkung, J., Chow, J., & Sales, A. (2020). Examining the Mutual Relations Between Language and Mathematics: A Meta-Analysis. *Psychological Bulletin*. Advance online publication. http://dx.doi.org/10.1037/bul0000231

Penninckx, M., Vanhoof, J., Quintelier, A., De Maeyer, S., & Van Petegem, P. (2017). *Zicht op leerwinst. Scenario's voor gestandaardiseerde toetsen*. Leuven: Acco.

Pinter, A. (2014). Child participant roles in applied linguistics research. *Applied Linguistics , Vol.35(2), 168–183.*

Ramaut, G., Roppe, S., Verhelst, M., & Heymans, R. (2007). *SALTO. Screeningsinstrument Aanvang Lager Onderwijs Taalvaardigheid.* Achtergronden. Leuven: KU Leuven, Centrum voor Taal en Onderwijs.

Robinson, P. (Ed.). (2011). *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (Vol. 2). John Benjamins Publishing.

Rowe, M. (2008). Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. J. *Child Lang. Vol. 35*, 185–205. doi:10.1017/S0305000907008343

Rubio-Codina, M., Araujo, M.C., Attanasio, O., Muñoz, P., & Grantham-McGregor, S. (2016). Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies*. PLoS ONE 11(8):* e0160962. doi:10.1371/journal.pone.0160962

Rueda, R., & Yaden, D. (2006). The literacy education of linguistically and culturally diverse young children: An overview of outcomes, assessment, and large-scale interventions. In B. Spodek and O.N. Saracho (Eds.), *Handbook of research on the education of young children* (2nd ed., pp. 167-186). Mahwah, NJ: Lawrence Erlbaum Associates.

Schachter, R.E., Strang, T. M., & Piasta, S.B. (2019). Teachers' experiences with a state-mandated kindergarten readiness assessment, *Early Years, 39:1,* 80-96. DOI: 10.1080/09575146.2017.1297777

Schouwstra, S., & Vloedgraven, J. (2020). *Wetenschappelijke verantwoording Kleuter in beeld* – Taal. Cito Arnhem.

Shohamy, E., Ben-Rafael, E., & Barni, M. (2010). *Linguistic landscape in the city*. Bristol by Multilingual Matters.

Sijtstra, J. et al., (2002). *Balans van het taalonderwijs aan het einde van de basisschool* 3. Uitkomsten van de derde peiling in 1998. Arnhem: Cito-groep.

Skehan, P. (Ed.). (2014). *Processing perspectives on task performance* (Vol. 5). John Benjamins Publishing Company.

Siu, A. L. (2015). *Screening for Speech and Language Delay and Disorders in Children Aged 5 Years or Younger: US Preventive Services Task Force Recommendation Statement*. www.pediatrics.org/cgi/doi/10.1542/peds.2015-1711

Skarakis-Doyle, E., & Dempsey, L. (2008). Assessing Story Comprehension in Preschool Children. *Topics in language disorders*. vol. 28(2), p. 131.

Smeyers, M., & Vandommele, G., (2003). Taalvaardigheid en taalvaardigheidstoetsen in Vlaanderen. In: T. Koole, B. Tahitu & J. Nortier (eds.), *Artikelen van de vierde sociolinguïstische conferentie*, Delft: Eburon.

Sneyers, E., Vanhoof, J., & Mahieu, P. (2020). Bias in primary school teachers' expectations of students? A study of general and specific bias towards ses, ethnicity and gender. *Studia paedagogica,* 25(2), 71-96.

Stein, Z. (2016). *Social Justice and Educational Measurement*. Routledge.

Sun, H., Steinkrauss, R., Tendeiro, J., & De Bot, K. (2015). Individual differences in very young children's English acquisition in China: Internal and external factors. *Bilingualism: Language and Cognition*,19(3), 550-577. https://doi:10.1017/S1366728915000243

Tarnanen, M., & Huhta, A. (2008). Interaction of Language Policy and Assessment in Finland, *Current Issues in Language Planning*, 9:3, 262-281, DOI: 10.1080/14664200802139547.

Teddlie, C., & Reynolds, D. (2000) The International Handbook of School Effectiveness Research. *School Leadership and Management,* Vol. 20(3):394 – 395.

Vanbuel, M., Boderé, A., & Van den Branden, K. (2017). *Helpen talenbeleid en taalscreening taalgrenzen verleggen? Een reviewstudie naar effectieve taalstimuleringsmaatregelen.* Gent, Steunpunt Onderwijsonderzoek.

Vanbuel, M., Vandommele, G., & Van den Branden, K. (2020a). *Taalvaardigheid screenen aan de start. De implementatie en implicaties van een verplichte low-stakes taalscreening in lagere en secundaire scholen.* Gent: Steunpunt Onderwijsonderzoek.

Van den Branden, K., Van den Nulft, D., Verhallen, M., & Verhelst, M. (2001). *Referentiekader vroege tweede taalverwerving. Een referentiekader voor doelstellingen rond vroege NT2-verwerving in Nederland en Vlaanderen.* Den Haag: Nederlandse Taalunie. Het Referentiekader is te downloaden op www.taalunieversum.org.

Van der silk, F. (2010). Acquisition of Dutch as a second language: The explanative power of cognate and genetic linguistic distance measures for 11 West European first languages. *Studies in Second Language Acquisition,* 401-432.

Van Petegem, P., & Vanhoof, J. (2002). *Evaluatie op de testbank: Een handboek voor het ontwikkelen van alternatieve evaluatievormen.* Mechelen: Wolters Plantyn.

Verhelst, M. (2002). De *relatie tussen mondeling taalaanbod en woordenschatverwerving van het Nederlands als tweede taal door 2,5-jarige allochtone kleuters in Brussel*. Leuven: KU Leuven, Faculteit Letteren (proefschrift).

Verhoeven, L., & Vermeer, A. (2005). Het ongelijk van Netelenbos. Toetsing van kleuters en hun prestaties op de Cito Eindtoets Basisonderwijs. In: *Toegepaste Taalwetenschap in Artikelen*, 74. 123-134, 2005.

Vermeir, K. (2019*). Implementatie van onderwijsinnovatie: artefacten, ondersteuners, agenda's en onderhandeling.* Proefschrift aangeboden tot het verkrijgen van de graad van doctor in de Pedagogische Wetenschappen.

Vermeir, K. (2020). Gewoon een kamishibai?! Implementatiepraktijken als het resultaat van interpretatieve onderhandelingsprocessen. *Pedagogische Studiën,* Vol. 97(1):24-41.

Vlaamse Overheid (2019). *Peiling Nederlands basisonderwijs: lezen – luisteren – schrijven*. Brussel.

Vlaamse Onderwijsraad (2021). *Taalscreening bij kleuters als hefboom voor een betere beheersing van het Nederlands. Kritische voorwaarden om de taalscreening in scholen te implementeren.* Brussel.

Ward, K.E., & Rothlisberg, B. (2011). Special issue: preschool assessment and intervention. *Psychology in the schools, vol. 48(5).* View this article online at wileyonlinelibrary.com/journal/pits DOI: 10.1002/pits.2056

Weir, C. (2005). Language Testing and Validation. An Evidence-Based Approach. In: *Research and Practice in Applied Linguistics*, Palgrave Macmillan.

Werkgroep Taaltoetsen (1995). *TAL. Taalvaardigheidstoets Aanvang Lager onderwijs*. Leuven: Steunpunt NT2.

Wheadon, C., & Stockford, I. (2010). *Classification accuracy and consistency in GCSE and A level examinations offered by the Assessment and Qualifications Alliance (AQA*). Ofqual/11/4823. Available: https://pdfs.semanticscholar.org/d725/366c7c074ddff4faa9f8bf115d825b900968.pdf

Willinger, U., Schmoeger, M., Deckert, M., Eisenwort, B., Loader, B., Hofmair, A., & Auff, E. (2017). Screening for Specific Language Impairment in Preschool Children: Evaluating a Screening Procedure Including the Token Test. J *Psycholinguist Res (2017)* 46:1237–1247. DOI 10.1007/s10936-017-9493-z

Yeung, S S., & Chan, C.K K. (2013). Phonological awareness and oral language proficiency in learning to read English among Chinese kindergarten children in Hong Kong. *British Journal of Educational Psychology*, 83, 550–568. https://doi.org/10.1111/j.2044-8279.2012.02082.x