KWALITEITSCENTRUM DIAGNOSTIEK VZW

Evaluatie van de CELF-5-NL

Met focus op de Vlaamse normering

Kwaliteitscentrum voor Diagnostiek vzw
Kortrijksesteenweg 405
9000 Gent

Website: www.kwaliteitscentrumdiagnostiek.be
E-mail: communicatie@kwaliteitscentrumdiagnostiek.be

Titel:	Evaluatie van de CELF-5-NL – met focus op de Vlaamse normering
Redactie:	Kwaliteitscentrum voor Diagnostiek vzw
Datum:	November 2020

Dit rapport kwam tot stand met de steun van de Vlaamse Overheid. In deze tekst komen onderzoeksresultaten van de auteur(s) naar voor en niet van de Vlaamse Overheid. Het Vlaams Gewest kan niet aansprakelijk gesteld worden voor het gebruik dat kan worden gemaakt van de meegedeelde gegevens.

Reviewprocedure

De evaluatie van de Clinical Evaluation of Language Fundamentals – versie 5 – Nederlandstalige versie (CELF-5-NL) kwam tot stand na raadpleging van twee onafhankelijke externe beoordelaars met expertise inzake testconstructie, psychometrie en/of het meten van taal- en communicatievaardigheden. Beide externe beoordelaars hebben het testmateriaal en de handleidingen van de CELF-5-NL grondig onderzocht en beoordeeld aan de hand van het beoordelingsmodel voor de beschrijving en evaluatie van psychologische en educatieve testen van de European Federation of Psychologists' Associations (EFPA; Evers et al., 2013), naar het Nederlands vertaald door het Kwaliteitscentrum voor Diagnostiek vzw[1]. De gemotiveerde beoordelingen werden onafhankelijk door beide externe experten bezorgd aan het Kwaliteitscentrum voor Diagnostiek vzw. Daarnaast heeft een wetenschappelijk onderzoeker van het Kwaliteitscentrum voor Diagnostiek vzw onafhankelijk een beoordeling gemaakt van de CELF-5-NL, eveneens aan de hand van het EFPA-beoordelingsmodel. Het Kwaliteitscentrum voor Diagnostiek vzw integreerde de drie onafhankelijke beoordelingen tot een eerste versie van dit adviesrapport dat vervolgens werd voorgelegd aan de externe beoordelaars en in onderling overleg tussen de beoordelaars werd besproken waarna enkele zaken zijn aangepast. Deze tweede versie van het adviesrapport werd voorgelegd aan de testauteurs. Via een gemotiveerd schrijven hebben de testauteurs gereageerd op de inhoud van het adviesrapport en extra toelichting gegeven omtrent enkele opmerkingen. Op basis van de inhoudelijke toelichting door de testauteurs werd het adviesrapport herwerkt door de wetenschappelijk medewerker van het Kwaliteitscentrum. Deze versie werd nogmaals nagelezen door de externe beoordelaars en opnieuw aangepast waar nodig. Dit resulteerde in de finale versie van het adviesrapport. Doorheen het geïntegreerd adviesrapport wordt de structuur van het EFPA-beoordelingsmodel zoveel mogelijk aangehouden.

Het adviesrapport heeft voornamelijk betrekking op het gebruik van de CELF-5-NL in Vlaanderen, en in mindere mate in Nederland. De beoordeling van het gebruik van de CELF-5-NL in Nederland gebeurde door de Commissie Testaangelegenheden Nederland (COTAN) en kan geraadpleegd worden op hun website (https://www.cotandocumentatie.nl/beoordelingen/).

[1] http://assessment.efpa.eu/documents-/ en https://portaal.kwaliteitscentrumdiagnostiek.be/wp-content/uploads/2020/09/EFPAtestbeoordeling2020.docx

Algemeen

Jaar van uitgave:	2019
Auteur(s):	E. Wiig, E. Semel, en W. E. Secord
Bewerker(s):	Pearson Benelux B.V. in samenwerking met dr. Jan de Jong
Uitgever:	Pearson Benelux B.V.
Referenties:	E. Wiig, E. Semel, & W. E. Secord (2019). *Clinical Evaluation of Language Fundamentals – Fifth edition – Nederlandstalige versie. Technische handleiding.* Amsterdam: Pearson Benelux B.V. E. Wiig, E. Semel, & W. E. Secord (2019). *Clinical Evaluation of Language Fundamentals – Fifth edition – Nederlandstalige versie. Afnamehandleiding.* Amsterdam: Pearson Benelux B.V.

Meetpretentie, doelgroep en gebruiksdoel

De Clinical Evaluation of Language Fundamentals, vijfde editie – Nederlandstalige versie (CELF-5-NL) is een instrument voor de identificatie, diagnose en opvolging van taal- en communicatiestoornissen bij kinderen en jongeren van 5 tot en met 18 jaar[2]. Specifiek richt deze test zich op problemen met taalproductie (expressieve taalvaardigheid), taalbegrip (receptieve taalvaardigheid) en problemen in alledaagse communicatie en sociale interactie. De test peilt niet naar fonologie of spraak.

De CELF-5-NL wordt door de auteurs omschreven als een test die bedoeld is voor *relatief minder belangrijke beslissingen op individueel niveau*[3]. Bovendien wijzen ze er terecht op dat de CELF-5-NL resultaten steeds in combinatie met andere (test)resultaten geïnterpreteerd moeten worden, vooraleer beslissingen op individueel niveau worden genomen. Mogelijk zal men in de Vlaamse praktijk de CELF-5-NL niet enkel inzetten voor minder belangrijke, maar ook voor *belangrijke beslissingen op individueel niveau*[4], zoals een verwijzing naar het buitengewoon onderwijs, onder meer omdat officiële instanties in Vlaanderen de CELF-5-NL aanwijzen als de beste in zijn categorie. Zo is de CELF-5-NL de enige test in de A-categorie van de limitatieve lijst van het Rijksinstituut voor ziekte- en invaliditeitsverzekering (RIZIV)[5]. Omwille van die reden wordt de CELF-5-NL in het huidige adviesrapport niet enkel beoordeeld voor het nemen van *relatief minder belangrijke beslissingen*, waarvoor de test volgens de auteurs is bedoeld, maar ook voor het nemen van *belangrijke beslissingen op individueel niveau*. De kwaliteitseisen betreffende normering en betrouwbaarheid gesteld aan testen voor belangrijke beslissingen zijn strenger dan deze gesteld aan testen voor minder belangrijke beslissingen.

[2] In het huidige adviesrapport wordt de term 'kinderen' gebruikt om te verwijzen naar kinderen en jongeren tussen 5 en 18 jaar.

[3] De COTAN geeft als voorbeelden voor relatief minder belangrijke beslissingen op individueel niveau: voortgangscontrole, beschrijvend gebruik van de testresultaten, therapie-indicatie en beroepskeuzebegeleiding (Evers, Boxtel, et al., 2010, pg. 22).

[4] Belangrijke beslissingen worden door de COTAN gedefinieerd als beslissingen die in principe op korte termijn, onomkeerbaar zijn, en die voor een belangrijk deel buiten de geteste persoon om worden genomen, zoals personeelsselectie, verwijzing naar speciaal onderwijs, opname/ontslag kliniek, certificering, … (Evers, Boxtel, et al., 2010, pg. 22).

[5] https://www.riziv.fgov.be/nl/professionals/individuelezorgverleners/logopedisten/Paginas/logopedisten-limitatieve-lijst-tests.aspx

Structuur

De Nederlandstalige CELF-5 omvat 11 subtests, namelijk Zinnen Begrijpen (ZB), Linguïstische Concepten (LC), Woordstructuur (WS), Woordcategorieën (WC), Aanwijzingen Volgen (AV), Zinnen Formuleren (ZF), Zinnen Herhalen (ZH), Tekstbegrip (TB), Definities van Woorden (DW), Zinnen Samenstellen (ZS) en Semantische Relaties (SR). Afhankelijk van de leeftijdscategorie van het kind of de jongere – 5-8, 9-12 of 13-18 jaar – kunnen niet alle subtests worden afgenomen. De subtests ZB, LC en WS zijn enkel bedoeld voor de jongste leeftijdsgroep (i.e., 5-8 jaar). DW, ZS en SR zijn daarentegen enkel bruikbaar voor de oudste leeftijdsgroep (i.e., 9-18 jaar). Door bepaalde subtests te combineren, afhankelijk van de leeftijd van het kind, kunnen verschillende indexscores worden berekend, namelijk de Receptieve Taal Index (RTI), Expressieve Taal Index (ETI), Taalinhoud Index (TII), Taalvorm Index (TVI) en Taalgeheugen Index (TGI). De Kernscore kan worden bepaald op basis van de afname van slechts vier subtests. De samenstelling hiervan is opnieuw afhankelijk van de leeftijd van het kind.

Naast de 11 subtests die bijdragen aan de indexscores en hoofdzakelijk taalvaardigheden in kaart brengen, kunnen ook drie observatieschalen of losstaande subtests worden ingezet voor het beoordelen van voornamelijk de functionele communicatievaardigheden van een kind, namelijk de Observatieschaal (OS), het Pragmatiekprofiel (PP) en de Checklist Pragmatiek in Activiteiten (PAC). De OS wordt ingevuld door het kind, een ouder en/of een docent en biedt informatie over luister-, spreek-, lees-, en schrijfvaardigheden van het kind thuis en op school. Het PP helpt testleiders om informatie te verzamelen over verbale en non-verbale pragmatische gedragingen van het kind die een negatieve invloed kunnen hebben op interactie en communicatie. De PAC is tot slot een observatie-instrument voor het in kaart brengen van functionele communicatievaardigheden tijdens een gedeelde activiteit, zoals het uitleggen en spelen van een spel.

Normen

De CELF-5-NL beschikt over afzonderlijke Nederlandse en Vlaamse normen voor alle subtests, indexscores en de Kernscore. Enkel voor de observatieschalen OS en PAC zijn geen normen voorhanden. Informatie verkregen aan de hand van de OS kan kwalitatief gebruikt worden bij de beoordeling van een kind. Voor de PAC is een criteriumscore beschikbaar gebaseerd op Amerikaans onderzoek.

Afname en scoring

De CELF-5-NL kan worden afgenomen door een logopedist, taal-/spraakpatholoog, schoolpsycholoog, (ortho)pedagoog of een andere persoon die op grond van opleiding en ervaring in staat is gestandaardiseerde taaltests af te nemen en te interpreteren en die een gedegen kennis heeft van de regels van het Nederlandse taalsysteem. Bovendien is het van belang dat de testleider goed vertrouwd is met de taalvariëteit(en) (standaardtaal, tussentaal of dialect) die worden gesproken in de omgeving van het kind.

Een afname kan via een papier-en-potlood versie of een digitale versie (Q-interactive) verlopen. Let op, de normen van de CELF-5-NL zijn uitsluitend gebaseerd op onderzoek met de digitale versie. Bij de digitale versie beschikken zowel het kind als de testleider over een tablet (uitsluitend iPads). De iPads communiceren met elkaar via een bluetoothverbinding. Het kind krijgt de items met visuele stimuli te zien op de cliënt-iPad en voor bepaalde subtests dienen de items te worden beantwoord door een afbeelding op het scherm aan te raken. De testleider gebruikt zijn iPad om aanwijzingen voor de testleider te raadplegen, antwoorden te registreren en te scoren, en aantekeningen te maken. Al het testmateriaal van de CELF-5-NL en de scoringsmodule zijn beschikbaar via het scherm.

De afnameduur van de CELF-5-NL is afhankelijk van de afnamewijze (i.e., digitaal versus papier-en-potlood), leeftijd van het kind, vaardigheden van het kind en het aantal subtests dat wordt afgenomen. De CELF-5-NL batterij wordt meestal niet volledig afgenomen, aangezien op basis van een 4- tot 6-tal subtests de algemene taalvaardigheid van een kind kan worden beoordeeld. Het digitaal afnemen van de subtests die nodig zijn voor het bepalen van de Kernscore vergt 30 tot 35 minuten tijd. Het digitaal afnemen van de subtests nodig

voor het bepalen van de Kernscore, de Receptieve Taal Index en de Expressieve Taal Index duurt ongeveer 45 minuten bij 5-8 jarigen en ongeveer 60 minuten voor de leeftijden 9-18 jaar. Om de volledige testbatterij af te nemen moet men ongeveer anderhalf uur uittrekken. De papier-en-potlood versie vergt meestal meer tijd dan een digitale afname.

Wanneer de CELF-5-NL digitaal wordt afgenomen, gebeurt de scoring gedeeltelijk automatisch, via Q-interactive. Bij subtests met een objectieve scoring (bv. Linguïstische Concepten en Woordcategorieën) antwoordt het kind via de iPad door een antwoord aan te klikken of registreert de testleider een verbaal antwoord van het kind door de scoreknop aan te klikken op de testleider-iPad. Deze antwoorden worden automatisch opgeslagen en gescoord door het systeem. Bij subtests waar ruimte is voor interpretatie moeten de testleiders zich een aantal scoringscriteria eigen maken (bv. Zinnen Formuleren). Bij deze subtests registreert de testleider het antwoord van het kind letterlijk op de testleider-iPad; standaard wordt ook een geluidsopname gemaakt. De scoring gebeurt in dit geval bij voorkeur achteraf door de testleider op de testleider-iPad. Een afname met papier-en-potlood kan ofwel manueel gescoord worden met de daartoe voorziene scoringsformulieren, ofwel via het (betalende) Q-global platform. In het laatste geval volstaat het om de ruwe scores in te geven in Q-global. Het programma genereert, rekening houdend met de socio-demografische gegevens van het kind (i.e., leeftijd en Nederlands/Vlaams), automatisch geschaalde subtestscores, indexscores, percentielscores, leeftijdsequivalenten en groeiscore-equivalenten. Ook worden significanties van verschilscores en kritieke waarden bij vergelijkingen tussen indexen automatisch berekend. Al deze resultaten worden gebundeld in een digitaal testrapport of voortgangsrapport. Deze methode vereenvoudigt het administratieve werk aanzienlijk.

Er worden verschillende referentieschalen gehanteerd doorheen de verwerking van de gegevens van de CELF-5-NL:

- Genormaliseerde standaardscores (geschaalde scores) met een gemiddelde van 10 en standaarddeviatie 3 voor de subtests;
- Genormaliseerde standaardscores (geschaalde scores) met een gemiddelde van 100 en standaarddeviatie 15 voor de indexen en Kernscore;
- Percentielscores voor de subtests, indexen en Kernscore;
- Leeftijdsequivalenten corresponderend met ruwe subtestscores;
- Een criteriumscore voor de Checklist Pragmatiek in Activiteiten;
- Gegevens voor scorevergelijking: kritieke waarden en prevalenties van indexscore-verschillen;
- Groeiscore-equivalenten corresponderend met ruwe subtestscores.

Het vergt van de testleider de nodige opleiding en een gedegen inzicht in de specifieke eigenschappen van deze schalen om er op een gepaste manier mee te kunnen omgaan bij de interpretatie.

Afname

Afname:	Individueel
Wijze:	Papier-en-potlood, digitaal (Q-interactive)
Door:	Logopedist, taal-/spraakpatholoog, schoolpsycholoog, (ortho)pedagoog of een andere persoon die op grond van opleiding en ervaring in staat is gestandaardiseerde taaltests af te nemen en die een gedegen kennis heeft van de regels van het Nederlandse taalsysteem en bovendien vertrouwd is met het taalgebruik in de omgeving van het kind
Tijdsduur:	Variabel: alle subtests ongeveer 1,5 uur; Kernscore ongeveer 30 tot 35 minuten (digitale afname)

Scoring

Papier-en-potlood:	Handmatig
Q-global & Q-interactive:	Invoer van de ruwe itemscores gebeurt (deels) handmatig Het scoren van subtests en indexen gebeurt geheel automatisch

Interpretatie

Door:	Logopedist, taal-/spraakpatholoog, schoolpsycholoog, (ortho)pedagoog en anderen die op grond van opleiding en ervaring in staat zijn gestandaardiseerde taaltests te interpreteren en die een gedegen kennis hebben van de regels van het Nederlandse taalsysteem

! **Het is belangrijk op te merken dat de Vlaamse en Nederlandse normeringsdata enkel werden verzameld met de digitale versie van de CELF-5-NL, aan de hand van Q-interactive. Dit rapport heeft dan ook enkel betrekking op de kwaliteit van de Vlaamse normen aan de hand van de digitale versie en betreft dus niet de papier-en-potlood afname.**

Beoordeling

Kwaliteit uitgangspunten, presentatie en beschikbare informatie

Uitgangspunten van de test

Zoals bij andere gestandaardiseerde taaltests gaat men bij de CELF-5-NL uit van één overkoepelende taalfactor die bestaat uit verschillende modaliteiten en dimensies, namelijk receptieve en expressieve taal enerzijds, en taalinhoud en taalvorm/taalgeheugen anderzijds. Het bestaan van één overkoepelende taalfactor wordt theoretisch onderbouwd door te verwijzen naar bestaande literatuur (bv. Tomblin & Zhang, 2006). Een verdere uiteenzetting van relevante literatuur die de structuur van de CELF-5-NL onderbouwt, voornamelijk het bestaan van de verschillende indexen, is echter beperkt. Bovendien, zoals wordt aangegeven in de Technische handleiding (pg. 20), toont onderzoek aan dat opsplitsing tussen een receptieve en expressieve taalindex eerder arbitrair en kunstmatig is (Leonard, 2009). Het onderscheid zou voornamelijk van praktisch belang zijn omdat het vaak nodig is in de communicatie met zorgverzekeraars en omdat de meeste classificatiesystemen het onderscheid maken. Het feit dat de indexscores onderling sterk correleren en factoranalyse een enkele factor oplevert, bevestigt dat deze indices niet opgevat dienen te worden als onafhankelijk, van elkaar te onderscheiden factoren, maar eerder als nuttige aanknopingspunten voor kwalitatieve interpretatie of therapie-indicatie. De inhoud van de subtests (i.e., wat meten de subtests) wordt eveneens behoorlijk theoretisch gestaafd aan de hand van relevante literatuur, wat bijdraagt tot de inhoudsvaliditeit van de CELF-5-NL.

De procedure die werd doorlopen om de Nederlandstalige CELF-5 te ontwikkelen uitgaande van de Amerikaanse versie wordt duidelijk toegelicht in de Technische handleiding. Hierbij werd onder andere aandacht besteed aan de behoeften vanuit het veld, de vertaling van het Engels (Amerikaanse CELF-5) naar het Nederlands, verschillen (in taalgebruik) tussen Vlaanderen en Nederland, en bias en fairness. Er werd een piloot-, try-out- en scoringsonderzoek uitgevoerd. Waarom precies werd gekozen om een Amerikaans instrument te vertalen om een goed Nederlandstalig *taal*instrument te ontwikkelen wordt echter weinig onderbouwd. Bij navraag verduidelijken de testauteurs dat deze keuze werd gemaakt omwille van de voordelen die internationale wetenschappelijke en praktische vergelijkbaarheid biedt.

Een aantal zaken dienen te worden opgemerkt bij de ontwikkelingsprocedure:

- Er wordt in de test te weinig rekening gehouden met de huidige realiteit wat betreft het gebruik van tussentaal in Vlaanderen. Bepaalde items bevatten zinnen in standaardtaal die door Vlaamse kinderen niet of nauwelijks worden gebruikt. In de CELF-5-NL had men meer rekening kunnen houden met deze gevoeligheden rond het normatief gebruik van Vlaamse tussentaal door items te selecteren die zowel in de standaardtaal als tussentaal gangbaar zijn.
- Bij de bespreking van de behoeften uit het veld wordt (nog) verwezen naar het Vlaamse M-decreet (Technische handleiding, pg. 45-46). Echter, dit zal met ingang van september 2021 (ten vroegste) vervangen worden door een begeleidingsdecreet.

Beschikbare documentatie

Globaal genomen verschaffen de Technische handleiding en Afnamehandleiding van de CELF-5-NL voldoende en duidelijke informatie over de ontwikkeling, normen, betrouwbaarheid en validiteit van de CELF-5-NL. Ze zijn bovendien toegankelijk geschreven en te begrijpen zonder dat hiervoor doorgedreven psychometrische kennis noodzakelijk is. Zoals ook verder wordt besproken doorheen dit rapport worden een aantal zaken echter onvoldoende toegelicht of ontbreekt nuttige informatie.

Bij de beschrijving van de normeringssteekproef wordt geen informatie verstrekt over de verdeling van de stratificatievariabelen per leeftijdsgroep, met uitzondering van geslacht. Bovendien wordt niet duidelijk toegelicht waarom enkel in Nederland een klinische steekproef werd gerekruteerd en niet in Vlaanderen, en

hoe de taalontwikkelingsstoornis bij deze kinderen werd gediagnosticeerd. Bij de toelichting van de validiteitsstudies, meer specifiek de beschrijving van de responsprocessen, is het bovendien jammer dat de moeilijkheidsgraden van de items niet te raadplegen zijn.

De informatie verstrekt in de handleidingen over de digitale testrapporten die verkregen worden na een digitale scoring is beperkt. In de Afnamehandleiding wordt kort vermeld welke referentieschalen worden berekend door Q-Interactive of Q-Global (bv. geschaalde subtestscores en percentielscores) en dat deze kunnen worden gedownload in de vorm van een rapport (Afnamehandleiding pg. 16/17 en 106). Er worden ook twee voorbeeldpagina's uit een dergelijk rapport weergegeven met grafieken en tabellen, maar geen volledig voorbeeldrapport. In navolging van deze opmerking plaatsten de testauteurs een dergelijk voorbeeld-rapport op hun website (te raadplegen op https://www.pearsonclinical.nl/pub/media/productfile/c/e/celf-5-nl_voorbeeldrapport_watermerk.pdf). Tot slot wordt nergens een controle van het systeem gerapporteerd, waardoor er, zonder zelf een handmatige controle uit te voeren, niet zonder meer kan worden vanuit gegaan dat deze rapporten valide en betrouwbaar zijn en dat de computerscoring correct verloopt (i.e., of deze equivalent is aan een manuele scoring). De testauteurs lieten weten dat dit uitvoerig gecontroleerd werd in een geautomatiseerd kwaliteitscontroleproces.

Procedurele instructies voor de testleider

Na het doornemen van de Technische handleiding en Afnamehandleiding, uitvoerig geïllustreerd met item-voorbeelden en casestudies, beschikken testleiders over uitgebreide informatie over hoe ze een CELF-5-NL afname kunnen voorbereiden, uitvoeren en scoren. Een aantal zaken zijn echter niet volledig duidelijk, voornamelijk met betrekking tot de doelgroep van de CELF-5-NL.

Er wordt in de handleidingen niet expliciet gemaakt wat de beperkingen zijn voor de inzetbaarheid van de CELF-5-NL. Uit de in- en exclusiecriteria die werden gehanteerd bij het rekruteren van de normeringssteekproeven en het feit dat geen verdere validiteitsstudies uitgevoerd werden kan wel indirect worden afgeleid dat de normen niet zonder meer bruikbaar zijn voor bepaalde doelgroepen, bijvoorbeeld voor kinderen met ernstige verstandelijke, motorische, auditieve of visuele beperking. Bij deze doelgroepen moeten de normen dus met voorzichtigheid gehanteerd worden. Bij rapportage en interpretatie van de resultaten dient hier extra aandacht aan besteed te worden. Daarnaast is de test niet geschikt voor kinderen die het Nederlands onvoldoende machtig zijn. In de normgroep werden (meertalige) kinderen die minder dan 7 jaar in het Nederlandstalige taalgebied wonen geëxcludeerd. In welke mate een bepaald niveau van leesvaardigheid vereist is, is bovendien ook onduidelijk. Ondanks dat alle items mondeling worden aangeboden, worden de items van sommige subtests ondersteund met een schriftelijke aanbieding. Het is mogelijk dat een beperkte leesvaardigheid een impact heeft op de score op deze subtests, dit werd niet onderzocht.

Daarnaast ontbreken instructies over hoe een testleider na afname feedback kan verlenen aan het kind en aan bevoegde derden. Tot slot wordt in de handleidingen geen informatie gegeven over hard- en softwarevereisten en technische ondersteuning bij een digitale afname. Er wordt hiervoor wel verwezen naar de website van Pearson. De testontwikkelaars melden hierbij dat dit een bewuste keuze is, aangezien technische vereisten betreffende hard- en software en supportmogelijkheden veranderlijk zijn en het van belang is dat gebruikers de meest recente informatie ter beschikking hebben.

Aandachtspunten

- De indexscores van de CELF-5-NL kunnen gebruikt worden voor een kwalitatieve interpretatie van de aard van de taalstoornis en zijn nuttige aanknopingspunten voor therapie, maar mogen niet opgevat worden als onafhankelijke van elkaar te onderscheiden factoren en kunnen bijgevolg niet zonder meer gebruikt worden voor het maken van een vaststaande differentiaaldiagnose betreffende een subtype van een taalstoornis (bv. expressieve taalstoornis).
- De normgegevens zijn niet zonder meer bruikbaar voor een CELF-5-NL afname bij (meertalige) kinderen die minder dan 7 jaar[6] in het Nederlandstalige taalgebied wonen, of kinderen met een ernstige verstandelijke, motorische, auditieve of visuele beperking.

Algemeen genomen kan de **kwaliteit van de uitgangspunten, presentatie en beschikbare informatie** als **goed** worden beoordeeld voor zowel het nemen van *belangrijke* als *minder belangrijke beslissingen op individueel niveau*. Bovenstaande aandachtspunten dienen evenwel in acht genomen te worden.

[6] Een termijn van 7 jaar werd gehanteerd in de exclusiecriteria van de normeringssteekproef. Deze termijn is richtinggevend maar niet absoluut. De CELF-5-NL is niet bruikbaar bij kinderen die het Nederlands onvoldoende beheersen. De kans hierop is groter, maar niet absoluut, wanneer het Nederlands niet de moedertaal is van het kind en in het geval het kind minder dan 7 jaar in het Nederlandse taalgebied woont.

Kwaliteit van het testmateriaal

Papier-en-potlood versie

Een CELF-5-NL testkoffer bevat twee opgaveboeken, twee leeftijdsgroep-specifieke bundels van elk 25 antwoordformulieren (i.e., 5-8 jaar en 9-18 jaar), een blok met 50 Observatieschaal-vellen en twee handleidingen (i.e., de Afnamehandleiding en de Technische handleiding). De antwoordformulieren zijn zeer goed opgesteld en degelijk uitgevoerd. Mede doordat de formulieren de afname-instructies bevatten, worden ze als bijzonder gebruiksvriendelijk beoordeeld. De opgaveboeken zijn eveneens handig in gebruik. De plastic ringen en de geperforeerde bladen zijn echter vrij kwetsbaar. De gehele testkoffer is bijzonder zwaar en log, onaangenaam om te dragen, moeilijk te sluiten en ook niet duurzaam, wat een negatieve invloed heeft op de gebruiksvriendelijkheid en duurzaamheid van de papier-en-potlood versie.

Daarnaast dient opgemerkt dat geen Nederlands-Vlaams normeringsonderzoek werd verricht met de papier-en-potlood versie, waardoor het gebruik van die versie niet kan worden aangeraden. De beschikbare normen zijn uitsluitend gebaseerd op de digitale afname via Q-interactive.

Digitale versie (Q-interactive)

Voor een digitale afname zijn twee iPads vereist, één voor de testleider en één voor de cliënt, die via een beveiligde bluetoothverbinding met elkaar communiceren. Het kind kan de visuele stimuli bekijken en de items beantwoorden via zijn iPad. De testleider-iPad wordt gebruikt om instructies te geven, antwoorden te registreren, te scoren en aantekeningen te maken. Via Q-interactive, een digitaal platform voor afname, kan de CELF-5-NL worden afgenomen en gescoord, en kan een rapport worden gegeneerd van de resultaten. Een testafname en een cliëntprofiel kunnen worden aangemaakt, klaargezet, en beheerd op Central, een beveiligde online omgeving.

Alle subtests kunnen via de iPads worden afgenomen. Er is dus geen extra materiaal nodig. In tegenstelling tot de papier-en-potlood afname, waar geen opnameapparatuur wordt voorzien, worden met de iPads automatische geluidsopnames gemaakt van de subtests die een verbale respons vereisen van het kind, zoals Zinnen Formuleren. Op die manier kunnen deze responsen achteraf gescoord worden, wat de efficiëntie en nauwkeurigheid van het scoringsproces aanzienlijk verhoogt. Bij gebruik van de digitale versie zal de testleider de online handleidingen moeten raadplegen die ter beschikking zijn op Central, papieren versies zijn niet inbegrepen in een Q-Interactive jaarlicentie.

Wat betreft gegevensbeveiliging en versleuteling geeft testuitgeverij Pearson aan dat Q-Interactive en Central voldoen aan de hoogste standaarden en aan de General Data Protection Regulation (GDPR). De details kunnen geraadpleegd worden op de website van Pearson: https://www.pearsonclinical.nl/gdpr en https://www.pearsonclinical.nl/platform_security_privacy. Echter, uit deze informatie is onvoldoende op te maken in welke mate Pearson voldoet aan alle vereisten van de GDPR. De beveiliging van de digitale CELF-5-NL lijkt op het eerste gezicht goed te zijn, maar het vraagt specialistische kennis die buiten de scope van deze beoordeling valt om hier een definitieve uitspraak over te doen. Hierbij wordt gevraagd aan Pearson om de informatie op hun website te verhelderen en precies aan te geven hoe ze de Europese regelgeving toepassen.

Over het algemeen is een digitale afname van de CELF-5-NL gebruiksvriendelijk. De software is van goede kwaliteit en de gebruikersinterface is duidelijk. Bovendien worden op Central gebruikershandleidingen voorzien waarin aan de hand van screenshots stap voor stap wordt toegelicht hoe je aan de slag kan gaan. Drie details kunnen de gebruiksvriendelijkheid inperken. Ten eerste moet het kind bij bepaalde subtests (bv. Aanwijzingen Volgen) alle figuren één voor één aanklikken wanneer gevraagd wordt een rij aan te duiden. Kinderen hebben de neiging om te 'swipen' of vegen over de ganse rij, in welk geval het systeem de respons als foutief registreert. Deze foute registratie kan eenvoudig gecorrigeerd worden door de testleider. Ten tweede kan de testleider bij de subtest Aanwijzingen Volgen telkens enkel het laatst door de deelnemer aangeklikte item (van een reeks) verwijderen/corrigeren, niet de items daarvoor. Wanneer het kind een foute reeks aanduidt en zich nadien corrigeert, worden zowel de foute als de juiste responsen opgeslagen en moet de testleider alle responsen verwijderen (te beginnen bij de laatste) en dan de correcte opnieuw ingeven. Tot

slot kan de testleider bij het herbeluisteren van een respons het item niet meer zien op de iPad (wat wel kan tijdens de afname). Een dergelijke functie was handig geweest aangezien het bij een aantal subtests belangrijk is dat de respons van de deelnemer past bij het item (bv. de afbeelding, het verhaal). De items kunnen wel worden geraadpleegd op Central en specifiek voor de subtest Zinnen Formuleren eveneens in Bijlage III van de Afnamehandleiding.

Algemene kwaliteit

Het testmateriaal (bv. afbeeldingen, items en instructies), beschikbaar via zowel papier-en-potlood als digitale afname, is over het algemeen verzorgd, duidelijk, hedendaags en voldoende aangepast aan de leefwereld van kinderen en jongeren. Instructies werden onderzocht naar verstaanbaarheid voor zowel Vlaamse als Nederlandse kinderen en werden aangepast naargelang de afnamemethode (i.e., papier-en-potlood versus digitaal). Om te onderzoeken of er sprake was van taalkundige bias op itemniveau tussen Vlaamse en Nederlandse kinderen werd gebruik gemaakt van Differential Item Functioning (DIF). Items die bias vertoonden werden aangepast of verwijderd. De Technische handleiding vermeldt dat eveneens DIF werd toegepast om verschillen tussen beide geslachten te onderzoeken (pg. 145). Hierover worden echter geen resultaten gegeven of conclusies beschreven. Wat betreft de afbeeldingen valt het op dat bij de subtest Woordstructuur, bij items over verkleinwoorden, dezelfde afbeeldingen worden uitvergroot of verkleind om een grote of kleine versie van een object/dier/persoon weer te geven, terwijl in de realiteit grote en kleine objecten, dieren en personen niet dezelfde morfologie hebben en in meerdere opzichten van elkaar verschillen dan enkel in afmetingen. Men kan zich afvragen of dit een invloed heeft op het antwoordgedrag van kinderen.

Er zijn bovendien inspanningen geleverd om de culturele fairness van de CELF-5-NL te waarborgen. Items tonen personen met diverse achtergronden, tijdens diverse activiteiten en in een variëteit van sociale rollen. Mogelijke bias met betrekking tot etnisch-culturele achtergrond, opleiding of cultuur werd tijdens het pilootonderzoek beoordeeld door experten en items werden aangepast waar nodig. Items die bijvoorbeeld als te 'Amerikaans' werden ervaren, werden verwijderd. In tegenstelling tot het onderzoek naar taalkundige en geslachtsbias werden jammer genoeg geen DIF-analyses uitgevoerd aan de hand van de normdata om interculturele fairness te onderzoeken. Dergelijk of ander toekomstig onderzoek zou sterkere conclusies toelaten omtrent fairness en culturele bias. Hoewel de ontwikkelaars van de CELF-5-NL aandacht hebben besteed aan de culturele fairness van de test, wat als positief wordt beoordeeld, kan niet gesteld worden dat de CELF-5-NL niet cultuurgeladen is. Bijgevolg is het, zoals bij elke (taal)test, van belang dat testleiders zich bewust zijn van een mogelijke culturele invloed op de resultaten.

Aandachtspunten

- Ondanks de aandacht die werd besteed aan de culturele fairness van de test kan, zoals bij andere (taal)tests, niet gesteld worden dat de CELF-5-NL niet cultuurgeladen is. Het is van belang dat testleiders zich altijd bewust zijn van een mogelijke culturele invloed.

Mits het in acht nemen van het bovenstaande aandachtspunt kan **de kwaliteit van het testmateriaal** als **goed** beoordeeld worden voor zowel het nemen van *belangrijke* als *minder belangrijke beslissingen op individueel niveau*.

Vlaamse normen

Het normeringsonderzoek is gebaseerd op de testgegevens van 1234 kinderen en jongeren uit Nederland en 608 uit Vlaanderen. Aangezien MANOVA[7] analyses wijzen op systematische, significante verschillen in de ruwe subtest- en indexscores tussen kinderen uit Nederland versus kinderen uit Vlaanderen (i.e., kinderen uit Nederland scoorden systematisch hoger), werd besloten om aparte normen te creëren. Het huidige rapport beoordeelt in hoofdzaak het Vlaamse normeringsonderzoek en de daaruit voortkomende Vlaamse normen.

Het valt op dat in Vlaanderen geen kinderen met een taalontwikkelingsstoornis (TOS; klinische groep) werden gerekruteerd. De testauteurs melden dat deze keuze bewust werd gemaakt omwille van praktische overwegingen. Om toch een breed scorebereik te verkrijgen, inclusief kinderen met scores in het lagere bereik die in de meeste gevallen geen regulier onderwijs volgen, werden acht kinderen met een TOS uit Nederland (1.3%) toegevoegd aan de Vlaamse normeringssteekproef (i.e., clinical seeding techniek). Het is jammer dat hiervoor geen Vlaamse kinderen met een TOS werden getest. Bijgevolg zijn de Vlaamse normen niet uitsluitend op Vlaamse testgegevens gebaseerd.

Algemeen

De Vlaamse normeringsteekproef bestond uit 608 kinderen en jongeren en werd gestratificeerd naar leeftijd, geslacht, regio, opleidingsniveau moeder, schooltype kind en etnisch-culturele achtergrond, en gecontroleerd voor thuistaal, urbanisatiegraad en onderwijsnet. De streefcijfers zijn gebaseerd op diverse bronnen, waaronder Statistics Belgium (Statbel) (2017) en het statistisch jaarboek van het Vlaams onderwijs 2016-2017 (Vlaams Departement Onderwijs en Vorming, 2015). Tijdens de dataverzameling werd gepoogd deze populatiecijfers zo goed mogelijk te benaderen. Tevens werd een wegingsprocedure uitgevoerd tijdens de data-analyse voor een aantal variabelen (i.e., geslacht, regio, opleidingsniveau moeder en etnisch-culturele achtergrond) om een zo groot mogelijke overeenkomst te bekomen tussen de steekproef en de populatie. Hierbij werd maximaal met een factor 2 gewogen zoals opgelegd door COTAN (Evers, Boxtel, et al., 2010) en werd de finale steekproefgrootte teruggebracht naar een eerder beperkte 568, wat overeenkomt met het vooropgestelde aantal voor de Vlaamse steekproef.

Dataverzameling

De dataverzameling vond plaats tussen augustus 2017 en januari 2019. Testleiders werden geselecteerd op basis van hun ervaring in het afnemen en scoren van testen en ervaring in het vakgebied van logopedie, (toegepaste) psychologie, (ortho)pedagogiek of taalwetenschap. Het is positief dat de meerderheid van de testleiders hun opleiding in één van de bovenstaande domeinen reeds had afgerond en slechts een beperkt aantal studenten deel uitmaakten van de groep testleiders.

De data werden systematisch verzameld met de digitale versie van de CELF-5-NL. Er werd tijdens de dataverzameling veel aandacht besteed aan het gestandaardiseerd afnemen van de test en aan het minimaliseren van de kans op fouten, wat als positief wordt beoordeeld. Zo kregen alle testleiders een intensieve training in de digitale afname, werden ze begeleid door ervaren coördinatoren, voerden ze oefenafnames uit waarop feedback werd voorzien, hadden ze de mogelijkheid om doorheen het hele proces vragen te stellen en werd de scoring van subtests waarbij interpretatie een rol speelt steekproefsgewijs gecontroleerd door externen.

De versie van de CELF-5-NL die gebruikt werd tijdens de dataverzameling is identiek aan de finale versie, met uitzondering van de afbreekregels. Tijdens de dataverzameling werden langere afbreekregels aangehouden teneinde over meer informatie te beschikken voor de analyses. De testauteurs deelden ons echter mee dat de normen werden bepaald door gebruik te maken van de definitieve afbreekregels (i.e., wanneer een kind een punt behaalde na het definitieve afbreekpunt werd dit niet meegerekend), wat als positief wordt beoordeeld. Vermoeidheid, aandacht en motivatie kunnen een grotere rol spelen bij subtests die langer

[7] MANOVA = Multivariate ANalyses Of VAriance

duren. Aangezien het slechts drie subtests betreft waarbij de afbreekregel telkens slechts één item langer was in de normeringsversie dan in de finale versie, is het aannemelijk dat dit geen invloed heeft gehad op de normen.

Representativiteit

De representativiteit van de Vlaamse steekproef wordt over het algemeen als goed beoordeeld. Per variabele worden de (gewogen) steekproefpercentages naast de vooropgestelde streefpercentages (i.e., verdeling in de populatie) geplaatst (Tabellen 3.13b, 3.14b, 3.15b, 3.16b, 3.17b, 3.18b, 3.19b). Afwijkingen van 5% of minder worden als (zeer) acceptabel beschouwd door de testauteurs. Er wordt echter terecht op gewezen dat de aanvaardbaarheid van een afwijking afhankelijk is van de veronderstelde invloed van een variabele (Technische handleiding, pg. 59). Na weging zijn er slechts twee kleine afwijkingen op het 5% criterium waar te nemen bij de stratificatievariabelen, namelijk bij regio (i.e., West: steekproefpercentage 18%, streefpercentage 12.7%) en bij schooltype kind (i.e., a-stroom: steekproefpercentage 20.1%, streefpercentage 26.1%). Er worden iets grotere afwijkingen geobserveerd bij de controlevariabelen onderwijsnet voor de leeftijdsgroep 5;00-11;11 (i.e., respectievelijk VGO en GO: steekproefpercentages 56.9% en 24.8%%, streefpercentages 62.2% en 14.9%) en voor de leeftijdsgroep 12;00-18;11 (i.e., VGO: steekproefpercentage 69.3%, streefpercentage 74.2%), en urbanisatiegraad (i.e., respectievelijk stad en niet-stad: steekproefpercentages 48.6% en 51.4%, streefpercentages 56.3% en 43.7%). Verder kunnen nog een aantal opmerkingen gegeven worden over de representativiteit van de Vlaamse steekproef.

Met uitzondering van de variabele geslacht worden geen tabellen weergegeven over de verdeling van de stratificatievariabelen per leeftijdsgroep. Het is bijgevolg niet mogelijk om na te gaan of de afzonderlijke leeftijdsgroepen binnen de steekproef een representatieve afspiegeling zijn van de populatie. Aangezien bij continue normering voornamelijk de uiterste leeftijdsgroepen van belang worden geacht (Evers, Sijtsma, et al., 2010), is vooral in die steekproefgroepen een goede weerspiegeling van de populatie belangrijk. Het wordt positief beoordeeld dat de jongste leeftijdsgroepen (i.e., 5;0-5;5, 5;6-5;11, 6;0-6;5 en 6;0-6;11) en oudste leeftijdsgroep (i.e., 17;0-18;11) in verhouding, dus per levensjaar, iets groter zijn dan de middelste leeftijdsgroepen. Of deze groepen representatief zijn voor de populatie kan echter niet worden nagegaan aan de hand van de beschikbare informatie.

Normeringsmethode

Continue normering en de mean-shift benadering

Voor het ontwikkelen van de Nederlandse normen werd een inferentieel normeringsmodel gehanteerd, een variatie op continu normeren. Daarbij werd gezocht naar de best passende functie voor het gemiddelde, de standaardafwijking (*SD*) en de scheefheid van de geobserveerde scores van elke subtest per leeftijdsgroep. Aangezien de Vlaamse steekproef te klein was om dezelfde methode te gebruiken voor de ontwikkeling van de Vlaamse normen, werd gebruikgemaakt van een mean-shift benadering met als referentie de Nederlandse steekproef. Meer specifiek werd voor de Vlaamse normering enkel een best passende functie bepaald voor de gemiddelde scores en werden de geselecteerde functies voor *SD* en scheefheid overgenomen uit de Nederlandse steekproef.

Aangezien de *SD* en scheefheid voor Vlaanderen gelijkgesteld werden aan die voor Nederland lijkt het ons van belang om naast de Vlaamse eveneens de Nederlandse normeringsmethode te beoordelen. Hierbij kan worden opgemerkt dat de methode van continue normering over het algemeen vrij summier wordt gerapporteerd in de Technische handleiding. Er worden bijvoorbeeld nergens goodness-of-fit maten weergegeven voor de geselecteerde functies (Lenhard et al., 2019). Er wordt enkel gesteld dat ‘de best passende’ functie werd geselecteerd voor elk moment (Technische handleiding, pg.76 en 84). Bijkomende informatie, die werd aangeleverd door de testauteurs, over het verloop van de geschaalde scoremomenten in functie van de leeftijd (zowel voor Vlaanderen als Nederland) wijst evenwel in de richting van een goede fit.

Steekproefgrootte

Nederlandse steekproef

Om aan te tonen dat het aantal deelnemers per leeftijdsgroep in de Nederlandse steekproef voldoende groot is, wordt in de Technische handleiding een vergelijking gemaakt tussen klassieke en continue normering volgens de methode van Bechger et al., (2009). De standaardfout van het gemiddelde werd berekend voor twaalf subtests – per leeftijdsgroep – voor twee klassieke normeringsmodellen (i.e., *N*=200 en *N*=300) en vergeleken met de standaardfout van het gemiddelde zoals verkregen onder het continue normeringsmodel. Op basis van deze vergelijking concludeert de testontwikkelaar dat de grootte van de Nederlandse steekproef goed te noemen is. Hierbij wordt verwezen naar de COTAN richtlijnen (Evers, Boxtel, et al., 2010).

Een aantal zaken dienen te worden opgemerkt. Ten eerste moet bij het hanteren van de methode van Bechger et al., (2009) rekening gehouden worden met een aantal statistische veronderstellingen (bv. normaliteit, homoscedasticiteit). In de handleiding van de CELF-5-NL wordt niet aangetoond dat aan deze veronderstellingen werd voldaan. De testauteurs wijzen erop dat deze informatie niet werd meegedeeld, evenmin als in de handleiding van vele andere testen, omdat er tot op heden nog geen objectieve methodieken en criteria voor zijn opgesteld.

Ten tweede werd enkel de standaardfout van het gemiddelde gebruikt om klassieke en continue normering met elkaar te vergelijken. Verder onderzoek waarbij ook andere parameters zoals de standaarddeviatie van de verdeling worden vergeleken tussen beide normeringsmethoden is wenselijk (Evers, Boxtel, et al., 2010).

Ten derde werd hier een vergelijking gemaakt met klassieke normeringsmodellen die volgens het beoordelingsmodel van EFPA als (net) voldoende (*N*=200-299) en goed (*N*=300-399) worden beoordeeld voor testen voor *relatief minder belangrijke beslissingen* (Evers et al., 2013). In het geval van *belangrijke beslissingen* worden de klassieke normeringsmodellen, waarmee hier een vergelijking werd gemaakt, door EFPA beoordeeld als onvoldoende (*N*=200-299) en (net) voldoende (*N*=300-399) (Evers et al., 2013). Op basis van de verstrekte informatie kon in het huidige adviesrapport geen vergelijking gemaakt worden met een klassiek normeringsmodel dat volgens EFPA als goed wordt beoordeeld voor testen voor belangrijke beslissingen (*N*=400-999). Wanneer belangrijke beslissingen, zoals een doorverwijzing naar het buitengewoon onderwijs, worden genomen op basis van de resultaten van de CELF-5-NL, kan op basis van de huidige argumentatie in de Technische handleiding dus niet zonder meer gesteld worden dat de Nederlandse normen adequaat zijn. Hieronder wordt verder toegelicht waarom dit voornamelijk voor een aantal specifieke subtests het geval is.

Voor de subtests Zinnen Begrijpen, Linguïstische Concepten en Woord Structuur, dewelke een beperkt leeftijdsbereik hebben (i.e., 5-8 jaar), kan voor de uiterste twee leeftijdsgroepen vastgesteld worden dat de standaardfout, zoals bekomen met continue normering, groter is dan wanneer een klassiek normeringsmodel met *N*=200 zou worden gehanteerd. Bij het nemen van *relatief minder belangrijke beslissingen* raden wij dan ook aan voor deze drie subtests bij interpretatie van de resultaten van kinderen uit de uiterste twee leeftijdsgroepen (i.e., 5;0-5;5 jaar en 8;0-8;11 jaar) rekening te houden met de grotere standaardfout.

Voor dezelfde drie subtests (i.e., Zinnen Begrijpen, Linguïstische Concepten en Woord Structuur) kan bovendien worden waargenomen dat voor de uiterste vier leeftijdsgroepen de standaardfout, zoals bekomen met continue normering, groter is dan wanneer een klassiek normeringsmodel met *N*=300 zou worden toegepast (zie Technische handleiding pg. 79-80). Enkel voor de middelste twee leeftijdsgroepen wordt een kleine winst geboekt ten opzichte van klassieke normering met *N*=300. Zoals eerder aangehaald, wordt een steekproefgrootte van *N*=300 door EFPA beoordeeld als (net) voldoende voor het nemen van belangrijke beslissingen (Evers et al., 2013). Voor deze drie subtests wordt de steekproefgrootte als onvoldoende beoordeeld voor het nemen van *belangrijke beslissingen*. Voor het nemen van dergelijke beslissingen zijn de standaardfouten voor deze subtests immers te groot, wat mogelijk kan leiden tot onnauwkeurige normen. Bij de meeste andere subtests kan voor de jongste leeftijdsgroepen eveneens waargenomen worden dat de standaardfouten onder continue normering groter zijn dan wanneer klassieke normering met *N*=300 zou worden toegepast. Aangezien dit enkel de uiterste leeftijdsgroepen betreft en de winst tegenover klassieke normering met *N*=300 in de andere leeftijdsgroepen voldoende groot is, is dit in eerste instantie

aanvaardbaar. Hier dient echter opnieuw te worden opgemerkt dat geen vergelijking kon worden gemaakt met klassieke normering met *N*=400 (i.e., (net) goed volgens EFPA voor het nemen van belangrijke beslissingen).

Vlaamse steekproef

Zoals hoger vermeld, werd bij de kleinere Vlaamse steekproef gebruik gemaakt van een mean-shift benadering om de normen te ontwikkelen, waarbij de theoretische populatieverdelingen gebaseerd werden op de Nederlandse normen. Dit houdt in dat de curves van opeenvolgende (volgens leeftijden) steekproevengemiddelden voor de subtests voor Vlaanderen middels een regressiemodel werden gerelateerd aan de vergelijkbare curves voor de Nederlandse normeringsdata. Op die manier werden mogelijke onregelmatigheden in het verloop van steekproefgemiddelden over opeenvolgende leeftijdsgroepen, ten gevolge van steekproeftoevalligheden, gecorrigeerd. Dit is een zinvolle werkwijze en de gebruikte data-analyse is degelijk. Echter, deze methode houdt een impliciete aanname in over equivalentie in taalontwikkeling over de taalgebieden heen. Bovendien is het niet helemaal duidelijk waarom voor de Vlaamse normen de functies voor SD en scheefheid werden overgenomen uit Nederland en niet werden geschat op een gecombineerde Nederlands-Vlaamse steekproef. Dat laatste was immers ook mogelijk geweest met de beschikbare data.

Virtueel karakter extreem hoge en lage scores

Bij het ontwikkelen van de normen werd gebruik gemaakt van een inferentieel normeringsmodel. Deze methode zorgt ervoor dat het mogelijk is om de frequenties van extreem hoge of lage scores te schatten, ook als die niet worden geobserveerd in een steekproef (Technische handleiding pg. 76). De normtabellen bestrijken de (theoretische) range voor de Kernscore gaande van 40 tot 160 en voor de indexscores van 45 tot 155. Aangezien de normen voor extreem hoge en extreem lage scores voornamelijk gebaseerd zijn op extrapolaties op basis van een mathematisch model raden we aan om voorzichtig om te springen met interpretaties van scores in de extreme zones.

Equivalentie papier-en-potlood versus digitale afname

Zoals eerder vermeld werden de data voor het normeringsonderzoek systematisch verzameld aan de hand van de digitale versie van de CELF-5-NL. In de Technische handleiding (pg. 87-90) wordt vermeld dat de normen gebruikt kunnen worden voor zowel digitale als papier-en-potlood afnames, aangezien verschillende studies de equivalentie tussen beide afnamewijzen in voldoende mate hebben aangetoond. Deze onderzoeken tonen echter enkel aan dat er op groepsniveau wellicht geen significante verschillen optreden wanneer de ene of de andere testmodaliteit wordt gebruikt. Een individu kan echter nog steeds verschillende resultaten behalen afhankelijk van de gehanteerde modaliteit, dit kan niet worden uitgesloten op basis van de beschreven onderzoeken. Bovendien vereist COTAN dat wanneer voor capaciteits- en vaardigheidstests normen zijn verzameld met behulp van een papier-en-potlood versie, terwijl de te beoordelen versie een computerversie betreft (of vice versa), er nieuwe normen worden verzameld (Evers, Boxtel, et al., 2010, pg. 21).

Met betrekking tot de equivalentie van de testversies is de verstrekte informatie eerder summier, met vage verwijzingen naar studies zonder vermelding te maken van de kwaliteitscriteria van deze studies zoals steekproefgroottes. Er wordt voornamelijk verwezen naar één Amerikaanse studie (Daniel et al., 2014). Wanneer we hier meer informatie over opzoeken wordt duidelijk dat hier slechts vier subtests werden vergeleken binnen een erg kleine steekproef (*N*=20) van kinderen en jongeren tussen 5;00 en 13;11 jaar. Wij willen er dan ook op wijzen dat bij het gebruik van de papier-en-potlood versie van de CELF-5-NL de huidige normen niet zonder meer gebruikt kunnen worden. De huidige beoordeling heeft dan ook enkel betrekking op de digitale CELF-5-NL.

Criteriumscore Checklist Pragmatiek in Activiteiten

Voor de Checklist Pragmatiek in Activiteiten (PAC) werden geen Nederlandse en Vlaamse data verzameld en geen normen ontwikkeld. Er wordt wel een criteriumscore voorzien. Wanneer een score van 9 of lager wordt behaald, stellen de testauteurs dat dit wijst op een adequate ontwikkeling van pragmatische vaardigheden. De criteriumscore is echter overgenomen van de Amerikaanse CELF-5 en dus gebaseerd op Amerikaans onderzoek. De motivering hiervoor is onvoldoende. De meeste items beschrijven inderdaad wel gedragingen die op het eerste gezicht gelden over culturen heen. Echter, er bestaan mogelijks kleine cultuurverschillen in wat sociaal aanvaardbaar gedrag is. Daarnaast zijn er zeker culturele verschillen terug te vinden in het gebruik van letterlijke/figuurlijke betekenissen, wat een invloed kan hebben op de scoring van items 16 en 17 'Begreep letterlijke/figuurlijke betekenissen niet.' Tot slot werden naar alle waarschijnlijkheid niet alle items letterlijk vertaald (bv. waar cultuurverschillen waarschijnlijk zijn), wat ook een invloed kan hebben op de scoring. Wij zijn dan ook van oordeel dat deze criteriumscore niet zonder meer kan worden gebruikt in Vlaanderen.

Bovendien kunnen vragen gesteld worden bij de kwaliteit van het Amerikaans onderzoek. Er werden data verzameld bij 61 kinderen van 5;0 tot en met 14;11 jaar met een pragmatische taalstoornis, en deze werden vergeleken met de scores van normaal functionerende leeftijdsgenoten. Gegevens van jongeren tussen 15;0 en 18;11 jaar ontbreken. Desondanks wordt in de handleidingen aangegeven dat de PAC en bijhorende criteriumscore kan gebruikt worden bij kinderen en jongeren tussen 5;0 en 18;11 jaar. Bovendien lijkt het tegen de intuïtie in te gaan dat voor alle leeftijden zondermeer dezelfde criteriumscore kan gehanteerd worden. Bepaalde (sociale) verwachtingen kunnen immers verschillen naargelang de leeftijd, bijvoorbeeld het item 'praatte te veel' zal waarschijnlijk sneller worden aangekruist bij een adolescent dan een jong kind wanneer hij/zij veel praat. Tot slot is het onduidelijk hoe de pragmatische taalstoornis werd vastgesteld in de klinische groep en of dit op een betrouwbare en valide manier werd gedaan.

Aandachtspunten

- Er werd veel aandacht besteed aan de stratificatie van de Vlaamse steekproef. Hierbij dient echter opgemerkt dat de kinderen met een taalontwikkelingsstoornis die in de Vlaamse steekproef werden opgenomen uit Nederland afkomstig zijn. De Vlaamse normen zijn bijgevolg niet op uitsluitend Vlaamse testgegevens gebaseerd.
- De Vlaamse normen werden ontwikkeld aan de hand van een mean-shift benadering, toegepast op de Nederlandse normen. Bij de ontwikkeling van de Nederlandse normen moet opgemerkt worden dat voor de subtests Zinnen Begrijpen, Linguïstische Concepten en Woord Structuur de gehanteerde steekproeven van de vier uiterste leeftijdsgroepen (5;0-5;5, 5;6-5;11, 7;0-7;11 en 8;0-8;11) te klein waren om aan de hand van continue normering voldoende nauwkeurige normen te bekomen voor het nemen van *belangrijke beslissingen*.
- De equivalentie tussen de papier-en-potlood versie en de digitale versie van de CELF-5-NL is onvoldoende aangetoond (op individueel niveau). Het gebruik van de huidige normen bij een afname aan de hand van de papier-en-potlood versie wordt daarom afgeraden.
- Het gebruik van de Amerikaanse criteriumscore bij de Checklist Pragmatiek in Activiteiten is onvoldoende onderbouwd en wordt afgeraden.

Voor het nemen van *minder belangrijke beslissingen op individueel niveau* kan de **Vlaamse normering** als **goed** worden beoordeeld.

Voor het nemen van *belangrijke beslissingen op individueel niveau* wordt de **Vlaamse normering** als **voldoende** beoordeeld.

De **criteriumgerichte normering van de PAC** wordt als **onvoldoende** beoordeeld aangezien het criterium gebaseerd is op Amerikaans onderzoek in een beperkte leeftijdsrange.

Bovenstaande aandachtspunten zijn van toepassing. De Nederlandse normeringssteekproef werd uitgebreider beoordeeld door COTAN. De COTAN beoordeling kan geraadpleegd worden op hun website: https://www.cotandocumentatie.nl/beoordelingen/ .

Betrouwbaarheid

Betrouwbaarheid verwijst naar de mate waarin scores vrij zijn van meetfoutvariantie. Er dient hierbij een onderscheid gemaakt te worden tussen verschillende types betrouwbaarheid (i.e., interne consistentie, test-hertestbetrouwbaarheid en interbeoordelaarsbetrouwbaarheid). Daarnaast kan de betrouwbaarheid en stabiliteit van een test ook worden onderbouwd door het beschrijven van standaardmeetfouten, betrouwbaarheidsintervallen en significantie van verschillen tussen scores. Bij de beoordeling dient verder rekening gehouden te worden met het doel van de test. Voor de Kernscore en indexscores worden in het huidige adviesrapport zowel de criteria voor minder belangrijke beslissingen gehanteerd, aangezien de CELF-5-NL voor zulke beslissingen wordt voorgesteld door de testontwikkelaars, als de criteria voor belangrijke beslissingen, aangezien de CELF-5-NL in de praktijk in Vlaanderen vaak wordt gebruikt bij het nemen van dergelijke beslissingen (zie supra). Aangezien het niet gerechtvaardigd is om op basis van één subtest een uitspraak te doen over de taalvaardigheid van een kind, gelden de criteria voor minder belangrijke beslissingen voor de subtests.

De **interne consistentie** van de CELF-5-NL werd onderzocht aan de hand van Guttmans lambda2 in zowel de Vlaamse als de Nederlandse normeringssteekproef. Deze statistiek verschilt in de praktijk weinig van Cronbach's alfa, hoewel Cronbach's alfa een iets strengere maat is en Guttmans lambda2 geacht wordt de reële interne consistentie beter te schatten. In het Vlaamse onderzoek zijn de betrouwbaarheidscoëfficiënten van de Kernscore voor alle leeftijdsgroepen te beoordelen als goed (≥.80) tot uitstekend (≥.90) volgens de criteria voor relatief minder belangrijke beslissingen. Ook voor de indexscores varieerden de coëfficiënten tussen .80 (goed) en ≥.90 (uitstekend), met uitzondering van de *Receptieve Taalindex* voor de leeftijdsgroep 8;0 jaar die een coëfficiënt van .78 (voldoende) had. Wanneer de criteria voor belangrijke beslissingen worden gehanteerd wordt de interne consistentie van de Kernscore en indexscores in de meeste leeftijdsgroepen beoordeeld als voldoende (≥.80) tot goed (≥.90), met uitzondering van de interne consistentie van de *Receptieve Taalindex* voor de leeftijdsgroep 8;0-8;11 jaar die in dit geval als onvoldoende wordt beoordeeld (.78).

De interne consistentie van de subtests toonde een meer variabel beeld. De betrouwbaarheidscoëfficiënten van de subtests varieerden van <.70 (onvoldoende) tot ≥.90 (uitstekend), waarbij de meerderheid een goede betrouwbaarheid had (.80-.90). De betrouwbaarheidscoëfficiënten tussen de .60 en .70 (onvoldoende) worden door de testauteurs wellicht terecht toegeschreven aan steekproeffluctuaties aangezien deze slechts in bepaalde (niet aangrenzende) leeftijdsgroepen voorkomen en deze meestal niet worden bevestigd voor de andere taalregio. Voor de coëfficiënten onder de .60 wordt in twee gevallen verwezen naar mogelijke plafondeffecten als oorzaak van de lage betrouwbaarheid, namelijk voor *Zinnen Begrijpen* in de twee oudste leeftijdsgroepen (7;0 en 8;0). Dit wijst erop dat deze subtest mogelijks een verminderd onderscheidingsvermogen heeft in deze leeftijdsgroepen voor de hogere scores. Ook over leeftijdsgroepen heen is de interne consistentie van *Zinnen Begrijpen* onvoldoende (.66). In een klinische groep van kinderen met een taalontwikkelingsstoornis (TOS) had deze subtest wel een betrouwbaarheidscoëfficiënt die voldoende was (.72) over leeftijden heen. Toch is voorzichtigheid geboden bij het interpreteren van resultaten van deze subtest, voornamelijk in de hoge range. Opmerkelijk is dat voor de lage interne consistentie van *Linguïstische Concepten* in de leeftijdsgroep 8;0 jaar (.61) niet wordt verwezen naar een dergelijk plafondeffect in de Vlaamse steekproef, terwijl dit wel wordt gedaan voor de Nederlandse steekproef. De betrouwbaarheidscoëfficiënt is iets lager in de Nederlandse (.49) dan in de Vlaamse steekproef, maar beide worden beoordeeld als onvoldoende en lijken veroorzaakt door een mogelijk plafondeffect (hoge gemiddelde ruwe scores in deze leeftijdsgroep). De testauteurs bevestigen dit, en zullen dit aanvullen in een volgende druk. Ook voor deze subtest is voorzichtigheid geboden bij interpretatie van resultaten in de hoge range bij kinderen zonder een taalstoornis in deze leeftijdsgroep.

In twee klinische groepen van kinderen met een diagnose TOS of dyslexie varieerden de betrouwbaarheidscoëfficiënten van de subtests tussen ≥.70 (voldoende) en ≥.90 (uitstekend). Dit wijst erop dat de CELF-5-NL subtests ook in deze groepen betrouwbaar kunnen worden ingezet. Dit betrouwbaarheidsonderzoek werd echter enkel uitgevoerd bij Nederlandse kinderen (met een TOS of

dyslexie) en er worden geen betrouwbaarheidscoëfficiënten weergegeven voor de Kernscore en de indexscores wat opmerkelijk is aangezien net deze scores het meest gehanteerd worden in de praktijk en dus de focus van het betrouwbaarheidsonderzoek zouden moeten vormen, conform met het testbeoordelingsmodel van EFPA (Evers et al., 2013).

Samengevat blijkt dat de Kernscore en indexscores een betere betrouwbaarheid en interne consistentie vertonen dan de afzonderlijke subtestscores. Zoals eerder aangehaald is het dan ook niet gerechtvaardigd om op basis van afzonderlijke subtests uitspraken te doen over de aan- of afwezigheid of de aard van een (sub)taalstoornis.

De **test-hertestbetrouwbaarheid** werd onderzocht in een samengestelde steekproef van 171 kinderen (123 uit Nederland en 48 uit Vlaanderen) door de CELF-5-NL tweemaal af te nemen met een gemiddeld tijdsinterval van 35 dagen. De (gecorrigeerde) stabiliteitscoëfficiënten van de Kernscore en indexscores wijzen op een uitstekende test-hertestbetrouwbaarheid (≥.80) in alle drie de onderzochte leeftijdsgroepen (5;0-8;11, 9;0-12;11 en 13;0-18;11) en dit voor het nemen van relatief minder belangrijke beslissingen. Enkel de *Taalinhoud Index* (9;0-12;11 jaar) haalde een iets lagere stabiliteit van .77 (goed). Voor het nemen van belangrijke beslissingen worden dezelfde coëfficiënten respectievelijk beoordeeld als goed (≥.80) en voldoende (.70-.80). De test-hertest betrouwbaarheidscoëfficiënten van de subtests varieerden tussen voldoende (≥.60) en uitstekend (≥.80). Over het algemeen kan bijgevolg gesproken worden over een goede test-hertestbetrouwbaarheid van de CELF-5-NL.

Volgende opmerkingen zijn evenwel van toepassing. Ten eerste is de volledige test-herteststeekproef voldoende groot volgens de EFPA-criteria (Evers et al., 2013), maar zijn de steekproeven van de drie afzonderlijke leeftijdsgroepen die gebruikt werden om de stabiliteitscoëfficiënten te berekenen te klein. Dergelijke kleine steekproeven (<100) worden door EFPA als onvoldoende beoordeeld. Bovendien is het onduidelijk of alle leeftijden vertegenwoordigd zijn in deze steekproef. Er wordt enkel gesteld dat er in het *vooraf* opgestelde steekproefplan spreiding van respondenten was over alle leeftijden (Technische handleiding, pg. 97). Ten tweede valt op dat, hoewel de steekproef redelijk gestratificeerd is volgens de streefpercentages, er een aantal duidelijke verschillen naar voor treden (bv. Vlaams streefpercentage 'hoog opleidingsniveau moeder' 40% is beduidend lager dan het gerealiseerde Vlaams steekproefpercentage 58%). Er is echter geen reden om er van uit te gaan dat dit een probleem vormt voor het onderzoeken van de test-hertestbetrouwbaarheid.

In het test-hertestonderzoek worden (kleine tot middelgrote) leereffecten teruggevonden voor de CELF-5-NL. Op basis hiervan wordt het afgeraden om, kort na een afname, een kind opnieuw te testen met de CELF-5-NL. Toch wordt er in de Technische handleiding (pg. 101) geen richtlijn gegeven voor een minimale tussentijd. De rationale die hiervoor wordt gegeven, is dat de gewenste tussentijd nog niet bepaald is door onderzoek. In de Afnamehandleiding wordt echter wel aangeraden om in de praktijk minstens een half jaar te laten tussen twee CELF-5-NL afnames. Het is onduidelijk van waar deze aanbeveling komt, en of een half jaar voldoende lang is om kunstmatig hogere scores op een hertesting te vermijden.

De **interbeoordelaarsbetrouwbaarheid** werd onderzocht voor de drie subtests die een beoordelend proces inhouden en dus vatbaar zijn voor de interpretatie van de testleider, namelijk Zinnen Formuleren, Zinnen Herhalen en Definities van Woorden. Hiervoor werden 103 (68 Nederlandse en 35 Vlaamse) willekeurige afnames gescoord door zowel de oorspronkelijke testleider, als door een tweede onafhankelijke beoordelaar op basis van een geluidsopname. De interbeoordelaarsbetrouwbaarheid was voor alle drie de subtests uitstekend. De Intraclass Correlatie Coëfficiënten (ICC) lagen tussen .86 en .99. Er wordt echter niet gerapporteerd welk type ICC werd berekend (voor een overzicht van verschillende types ICC's, zie Hallgren, 2012). De onderzochte interbeoordelaarsbetrouwbaarheid betreft bovendien enkel *de scoring* van drie subtests, niet de afname (van de ganse test). Daarom blijft het belangrijk om voldoende training te voorzien voor testleiders, zodat de afname zo goed mogelijk gestandaardiseerd verloopt en de invloed van de beoordelaar minimaal is.

Verder worden ook **standaardmeetfouten (SEM) en betrouwbaarheidsintervallen** beschreven om de betrouwbaarheid van de CELF-5-NL verder te onderbouwen. Een SEM is omgekeerd evenredig aan de

betrouwbaarheid, hoe kleiner de SEM hoe betrouwbaarder de subtest- of indexscore. Het feit dat de SEMs worden weergegeven, wordt als positief beoordeeld (Evers et al., 2013). De SEMs in de (kleine) Vlaamse steekproef zijn bovendien gelijkwaardig aan deze uit de (grotere) Nederlandse steekproef, wat de betrouwbaarheid en veralgemeenbaarheid van de SEMs onderbouwt.

Tot slot worden de **verschillen tussen indexscores** beschreven en besproken in de Technische handleiding. Deze verschilscores worden vaak gebruikt bij de interpretatie van de prestaties van een kind op de CELF-5-NL. Er wordt in de Technische handleiding terecht op gewezen dat hierbij zowel de statistische significantie als de klinische waarde (frequentie van voorkomen) van een verschil een rol spelen, en dat verschilscores met de nodige voorzichtigheid moeten worden geïnterpreteerd. Met betrekking tot de statistische significantie van de verschillen worden kritieke waarden weergegeven voor zowel het .15- als .05-significantieniveau (cf. bijlage E.1 van de Afnamehandleiding). Afhankelijk van hoe snel een testleider getriggerd wil worden voor een mogelijk verschil kan hij gebruik maken van ofwel het .15- (snellere detectie, maar meer risico op loos alarm) ofwel het .05-significantieniveau. Verder is het in de tabellen van bijlage II. E.2 onduidelijk dat de percentages weergeven hoeveel procent kinderen uit de steekproef een dergelijk verschil *of kleiner verschil* behaalden. De tabeltitel 'Prevalentie van indexscoreverschillen' doet vermoeden dat het percentage dat wordt vermeld naast een scoreverschil aangeeft welk percentage kinderen uit de normeringssteekproef dat specifieke scoreverschil behaalde. Dit is misleidend en zou kunnen worden opgelost door een '≥' voor elk scoreverschil te plaatsen. Bovendien zijn de prevalenties van de indexscoreverschillen van kinderen met een TOS (Technische handleiding, Tabel 4.6, p. 106) enkel gebaseerd op een Nederlandse steekproef, en kunnen deze bijgevolg niet zonder meer worden gebruikt bij het interpreteren van scoreverschillen bij kinderen met een TOS in Vlaanderen.

In de Technische handleiding, bij de bespreking van de indexscoreverschillen, wordt aangegeven dat ongeveer de helft van de kinderen in de normeringssteekproef hogere scores behaalde op de ene index (bv. ETI) en de helft op de andere index (bv. RTI). Dit doet vermoeden dat er bij (bijna) alle kinderen sprake is van een daadwerkelijk verschil tussen beide indexen. Bij deze beschrijving is geen rekening gehouden met de significantie van deze verschillen. Ook verschillen van 1 punt worden hier meegeteld, en deze zijn nooit significant. Om verwarring te vermijden zou het beter zijn om te beschrijven dat gemiddeld genomen een verschil van X punten een statistisch significant verschil is, en X% van de kinderen een dergelijk indexscoreverschil behaalt, waarbij X% hoger scoort op de ene en X% op de andere index. Tot slot wordt bij verschillen tussen ETI en RTI over grote verschillen gesproken bij 20 punten of meer, en bij verschillen tussen TII en TGI bij 10 punten of meer. Het zou eenduidiger zijn om te spreken over klinische relevante verschillen in plaats van over 'grote' verschillen (bv. indien <15% van de populatie dit verschil vertoont, Sattler, 2008).

Aandachtspunten

- Het doen van uitspraken over de aan- of afwezigheid of de aard van een (sub)taalstoornis op basis van individuele subtests wordt sterk afgeraden.
- Voorzichtigheid is vooral geboden bij de interpretatie van de scores in de hoge range op de subtests *Zinnen Begrijpen* en *Linguïstische Concepten*, vanwege hun lage interne consistentie ten gevolgen van plafondeffecten in bepaalde leeftijdsgroepen, waardoor de scores onvoldoende discrimineren.
- Een korte tussentijd bij hertesting wordt afgeraden. Het is echter onduidelijk hoeveel tijd minimaal moet worden gerespecteerd tussen twee opeenvolgende afnames.

De **betrouwbaarheid** van de CELF-5-NL wordt als **goed** beoordeeld voor zowel het nemen van *belangrijke* als *minder belangrijke beslissingen op individueel niveau*. Bovenstaande aandachtspunten zijn van toepassing.

Validiteit

De validiteit van een test verwijst naar de mate waarin de test aan zijn doel beantwoordt. Er bestaan verschillende types of vormen van validiteit, die elk een bijdrage kunnen leveren tot het ondersteunen van de algemene validiteit van de test. In deze beoordeling volgen we de klassieke driedeling; inhoudsvaliditeit, begripsvaliditeit en criteriumvaliditeit.

Inhoudsvaliditeit en responsprocessen

Inhoudsvaliditeit betreft de mate waarin de inhoud van de test een goede weergave is van het te meten construct, in dit geval taalvaardigheid. De inhoudsvaliditeit van de CELF-5-NL wordt positief beoordeeld. Ondersteuning hiervoor wordt voornamelijk geboden aan de hand van feedback van experts en gebruikers en verwijzingen naar de literatuur. Zoals eerder beschreven, wordt de inhoud van iedere subtest van de CELF-5-NL behoorlijk theoretisch gestaafd (cf. Uitgangspunten van de test).

Verder is er veel aandacht besteed aan responsprocessen (i.e., worden de verwachte cognitieve processen gebruikt door het kind bij beantwoorden van items) in zowel het Vlaams-Nederlands als het Amerikaans onderzoek, wat extra validiteitsbewijs oplevert. Antwoorden op pilootitems werden geanalyseerd, moeilijkheidsgraden van items en responsfrequenties werden onderzocht, scoringsonderzoek werd uitgevoerd, en de feedback van testleiders tijdens de piloot- en try-out fases werd meegenomen in de finale constructie van de item-set. Idealiter zouden ook gegevens over de moeilijkheidsgraden van de items beschikbaar zijn. Deze informatie is immers belangrijk voor (het beoordelen van) de correcte itemvolgorde en gerelateerd het correct functioneren van de test. Daarnaast kan deze informatie ook klinisch relevant zijn. Op basis van de gegevens over de moeilijkheidsgraden van de items kan bijvoorbeeld mogelijks duidelijk worden dat de afstand tussen items qua moeilijkheid niet steeds dezelfde is, en er mogelijks soms sprake is van een sprong in moeilijkheid. Deze gegevens kunnen informatief zijn bij het afnemen van de test en het interpreteren van de testresultaten.

Begripsvaliditeit

Meet de CELF-5-NL daadwerkelijk wat hij bedoelt te meten? Deze vraag peilt naar begripsvaliditeit. De Technische handleiding levert hiervoor evidentie aan door het beschrijven van een confirmatorische factoranalyse, de interne structuur met analyse van intercorrelaties, verschillen tussen relevante groepen, en relaties met andere testen.

Uit de ***confirmatorische factoranalyse (CFA)*** blijkt dat structuren met één tweede-orde factor (Kernscores) en twee eerste-orde factoren (ofwel taalinhoud en taalvorm/taalgeheugen, ofwel receptieve taal en expressieve taal) de data algemeen goed lijken te fitten. Hierbij moeten echter een aantal zaken worden opgemerkt.

Ten eerste is het jammer dat geen factorladingen worden weergegeven in de Technische handleiding ter ondersteuning van de bevindingen.

Ten tweede wordt in de Technische handleiding slechts een beperkte rationale gegeven voor de gekozen structuren. Het is onduidelijk waarom enkel deze structuren werden getest, en waarom de fit hiervan niet vergeleken werd met deze van andere mogelijke structuren, bijvoorbeeld een structuur met één tweede-orde factor en *vier* eerste-orde factoren (taalinhoud, taalvorm, receptieve taal en expressieve taal). De testauteurs lieten ons hierover weten dat de geteste structuur diegene is die werd gevonden in het Amerikaans onderzoek en het voornamelijk de bedoeling was om na te gaan of deze structuur kon worden teruggevonden in de CELF-5-NL.

Ten derde tonen de fit-indices dat modellen met receptieve en expressieve taal als tweede-orde factoren (model 1) de data beter fitten dan modellen met taalinhoud en taalvorm/taalgeheugen als tweede-orde factoren (model 2). Dit is onder andere te zien aan de AIC en BIC fit-indices die in alle leeftijdsgroepen lager zijn voor model 1 dan voor model 2. Deze verschillen tussen beide typen modellen worden nergens vermeld.

Ten vierde vertoont model 2 in de leeftijdsgroep van 9-12 jarigen een zwakke fit (i.e., RMSEA >.08, AGFI <.90 en hoge AIC en BIC). Aangezien dit slechts in één leeftijdsgroep het geval is, kan dit mogelijks te wijten zijn aan steekproeftoevalligheden waardoor het geen probleem vormt. Een vermelding hiervan had echter wenselijk geweest.

Ten vijfde kan worden opgemerkt dat de AGFI fit-index niet wordt toegelicht in de Technische handleiding, in tegenstelling tot de andere fit-indices. Het is bijgevolg onduidelijk voor de lezer welke waarde van AGFI wijst op een aanvaardbare fit. Recente literatuur geeft aan dat een AGFI van .90 of hoger wijst op een acceptabele fit (Wang et al., 2019). De testauteurs laten weten dat ze een toelichting van de AGFI-index zullen toevoegen in een volgende druk van de Technische handleiding.

Ten zesde wordt ook de χ^2/df index gebruikt om de fit van de modellen te onderzoeken, aangezien deze minder gevoelig zou zijn voor grote steekproeven dan χ^2. De literatuur wijst er echter op dat χ^2 enkel gevoelig is aan de steekproefgrootte bij incorrecte modellen, dat *df* niet afhangt van de steekproefgrootte (en χ^2 delen door *df* dus niet echt zin heeft) en dat er geen duidelijke richtlijnen zijn die aangeven welke waarde van χ^2/df wijst op een acceptabele fit (Kline, 2011). Aangezien er nauwelijks statistische grond is voor de χ^2/df fit-index raden we dan ook af om deze te gebruikten.

Tot slot vermelden de testauteurs dat de Nederlandse en Vlaamse steekproeven werden samengenomen omdat wordt aangenomen dat de structuur dezelfde is in beide landen. Een afzonderlijke CFA in de Nederlandse en Vlaamse steekproef had echter kunnen bijdragen tot het aantonen van deze aangenomen 'structuurinvariantie'. Daarvoor wordt momenteel geen bewijs aangeleverd.

De **interne structuur** wordt in de Technische handleiding onderbouwd aan de hand van een beschrijving van **intercorrelaties tussen subtests en indexscores** voor de Vlaamse en Nederlandse normeringssteekproef afzonderlijk. Er wordt geen informatie verstrekt over item-subtest of item-index verbanden. Verder is het niet mogelijk om de verbanden te beoordelen per leeftijdsgroep, aangezien enkel correlaties worden weergegeven over alle leeftijdsgroepen heen. Aangezien de samenstelling van de indexen verschilt over leeftijden heen had een opsplitsing in drie brede groepen (i.e., 5-8, 9-12 en 13-18 jaar) een logischere keuze geweest.

A priori werden verwachtingen geformuleerd over de sterkte van de verbanden. Deze verwachtingen werden vrij vaag geformuleerd zonder daarbij te refereren naar theoretische taalmodellen. Met betrekking tot de subtests werd verwacht dat deze allen 'in meer of mindere mate' samenhangen. Deze eerder vrijblijvende verwachting wordt bevestigd (r=.35-.66). Vervolgens werd ook een bepaalde mate van divergentie verwacht, aangezien bepaalde indexen en subtests ontwikkeld zijn om deelaspecten van taalvaardigheid te meten. Er is echter geen duidelijk patroon terug te vinden in de verbanden (bv. een sterkere samenhang tussen subtests die tot dezelfde index behoren dan tussen subtests van te onderscheiden indexen). Het is jammer dat hierover geen melding wordt gemaakt. Verder zijn de gecorrigeerde correlaties tussen subtests en indexen waartoe de subtests behoren niet sterker dan de verbanden tussen subtests en indexen waartoe ze niet behoren, wat we wel zouden verwachten. Met betrekking tot de indexen onderling werden de laagste correlaties verwacht tussen indexen die ontwikkeld zijn om te onderscheiden deelaspecten van taal te meten, namelijk ETI versus RTI, en TGI/TVI versus TII. Deze correlaties zijn over het algemeen inderdaad lager (r=.68-.76) dan de overige index-correlaties (r=.81-.96). Volgens de testauteurs ondersteunen deze resultaten het idee dat deze indexen verschillende aspecten van taalvaardigheid meten. Echter, de 'zwakkere' correlaties waren nog steeds ± .70 (i.e., hoog), en de verbanden tussen TII en ETI (r=.76), en TVI en RTI (r=.74) vertonen eveneens een dergelijke 'zwakkere' relatie. Deze indexen meten echter geen te onderscheiden deelaspecten van taal, maar delen net zoals de te onderscheiden aspecten eveneens geen of nauwelijks subtests met elkaar. Dit wijst er op dat de 'zwakkere' correlaties eerder te wijten zijn aan het feit dat bepaalde indexen geen of nauwelijks subtests delen, dan aan het feit dat ze te onderscheiden taalgebieden meten. Fundamenteel moet worden gesteld dat de CELF-5-NL vooral één grote factor "taalvaardigheid" meet en niet echt in staat is om deelaspecten van die taalvaardigheid te onderscheiden. Een laatste bedenking betreft de verwaarloosbare verbanden tussen het Pragmatiek Profiel (PP) en de overige subtests en indexscores (r<.30). De testauteurs geven aan dat deze bevinding overeenkomt met de verwachtingen, aangezien deze subtest

iets anders meet dan de andere onderdelen van de CELF-5-NL en omgekeerd. Hierbij kan men zich afvragen of het PP niet eerder moet worden beschouwd als een afzonderlijk (aanvullend) instrument dan een subtest van de CELF-5-NL.

De verwachtingen die werden geformuleerd met betrekking tot de hierboven beschreven intercorrelaties zijn allen gebaseerd op theoretische aannames over taalvaardigheid enerzijds en de idee dat de subtests van de CELF-5-NL in staat zijn om te discrimineren anderzijds. Een exploratieve analyse van de interne structuur van de batterij, zoals exploratieve factoranalyse of principale componentenanalyse, zou toelaten om na te gaan in welke mate deze testbatterij in staat is om meer dan één algemene taalvaardigheidsfactor te onderscheiden. Het is jammer dat dergelijke exploratieve analyses niet werden uitgevoerd om de interne structuur van de CELF-5-NL mee te bepalen.

Wanneer verwachte **verschillen tussen relevante groepen** worden teruggevonden, biedt dit verdere evidentie voor de begripsvaliditeit van een test. In de Technische handleiding wordt ten eerste de ***samenhang met socio-demografische gegevens*** besproken: namelijk met leeftijd, geslacht, opleidingsniveau moeder, migratieachtergrond en thuistaal. De meeste relaties en verschillen liggen in lijn met de verwachtingen. Echter, met uitzondering van de verschillen tussen jongens en meisjes, wordt geen informatie verstrekt over de statistische significantie van de gevonden verbanden en verschillen. Het blijft dus onduidelijk of dit daadwerkelijke verschillen zijn. Met betrekking tot geslachtsverschillen tonen een aantal voorgaande studies aan dat meisjes gemiddeld beter presteren op taaltests dan jongens, maar andere studies vinden geen geslachtsverschillen. Op basis van deze tegenstrijdige bevindingen kan dus geen eenduidige hypothese worden gesteld over geslachtsverschillen bij de CELF-5-NL. Toch wordt door de testauteurs de verwachting geformuleerd geen significante geslachtsverschillen te vinden, dewelke niet volledig kan worden bevestigd door de resultaten uit de Nederlandse steekproef waar meisjes op de meeste aspecten hoger scoren dan jongens. Bovendien is onduidelijk waarom de geslachtsverschillen werden getoetst op het .01-significantieniveau. Wanneer het meer gebruikelijk .05-significantieniveau zou worden gehanteerd, zouden waarschijnlijk nog meer geslachtsverschillen worden teruggevonden. De testauteurs bevestigen dat er inderdaad meer geslachtsverschillen worden teruggevonden in de Nederlandse steekproef wanneer gebruik wordt gemaakt van het .05-significantieniveau. In de Vlaamse steekproef worden zowel bij het .01- als het .05-significantieniveau geen significante verschillen teruggevonden tussen jongens en meisjes. De testauteurs geven bovendien in een aparte communicatie aan dat de keuze om geen aparte normen te voorzien voor jongens en meisjes kan worden gemotiveerd omdat het bijvoorbeeld maatschappelijk niet wenselijk zou zijn om jongens en meisjes verschillend te behandelen om in aanmerking te komen voor doorverwijzing of extra begeleiding op school.

Ten tweede werden ***verschillen tussen klinische groepen*** (i.e., kinderen met een taalontwikkelingsstoornis (TOS) en dyslexie) en de normeringssteekproef onderzocht. Hierbij is het jammer dat enkel Nederlandse kinderen werden onderzocht en geen Vlaamse. De verwachting dat *kinderen met een TOS* lagere gemiddelde subtests- en indexscores behalen dan kinderen uit de normeringssteekproef (met grote effectgroottes) werd bevestigd door de resultaten. Bij dit onderzoek valt op dat ook kinderen met zeer ernstige begripsproblematieken deel uitmaken van de klinische groep (Siméa, 2014) wat de gemiddelde scores sterk naar beneden kan hebben getrokken. Bovendien werden enkel kinderen uit het buitengewoon onderwijs geïncludeerd en niet diegene met een (lichtere) TOS in het algemeen (inclusief) onderwijs. Hierdoor is het onduidelijk of de CELF-5-NL ook discrimineert tussen kinderen met en zonder minder ernstige TOS-problematieken. Finaal besluiten de testauteurs dat deze resultaten onderbouwen dat de CELF-5-NL kan worden ingezet om kinderen met een normale taalontwikkeling te onderscheiden van kinderen met een TOS. Deze uitspraak is pas gerechtvaardigd na het uitvoeren van onderzoek naar het diagnostisch onderscheidingsvermogen van de CELF-5-NL (cf. onder criteriumvaliditeit) en staat hier dus niet op zijn plaats. De resultaten bieden hier wel verdere evidentie voor de begripsvaliditeit van de CELF-5-NL.

Naast kinderen met een TOS werden ook *kinderen met dyslexie* onderzocht. Ook hier werd gevonden dat kinderen met dyslexie systematisch lagere scores behalen op de CELF-5-NL subtests dan kinderen uit de normeringssteekproef, met uitzondering van de scores op het Pragmatiek Profiel en Tekstbegrip. De testauteurs halen hierbij terecht aan dat kinderen met lees-en-spellingsproblemen vaak ook een taalstoornis

hebben, en geven dit als rationale voor de gestelde hypothese dat kinderen met dyslexie lagere scores hebben op de CELF-5-NL. Als de verschillen inderdaad te wijten zijn aan een comorbide taalstoornis, kan de meerwaarde van deze studie in vraag worden gesteld. Bovendien zou een waarschuwing waarin wordt aangegeven dat de CELF-5-NL niet bruikbaar is voor het stellen van een diagnose dyslexie niet overbodig zijn.

In de Technische handleiding wordt ook verwezen naar een aantal Amerikaanse studies bij klinische groepen, namelijk bij kinderen met een taalstoornis, lees-en-spellingsproblemen en autismespectrumstoornis (ASS). Buitenlandse studies kunnen de validiteit verder onderbouwen via het principe van validiteitsgeneralisatie (Evers, Boxtel, et al., 2010), dus algemeen wordt deze aanvulling als positief beoordeeld. In lijn met het Nederlands onderzoek stellen we ons echter ook hier de vraag wat de meerwaarde is van de studies bij kinderen met lees-en-spellingsproblemen of ASS. Opnieuw bestaat hier de mogelijkheid dat verschillen in gemiddelde CELF-5-NL scores eerder te wijten zijn aan een comorbide taalstoornis dan aan de lees-en-spellingsproblemen of de ASS. Bovendien kunnen deze studies onterecht het idee geven dat de CELF-5-NL kan gebruikt worden voor het stellen van een diagnose van deze stoornissen. Ook hier zou een waarschuwing gepast zijn.

Het is jammer dat in het validiteitsonderzoek geen (klinische) groepen werden onderzocht met een achterstand in één of enkele specifieke domeinen van taalvaardigheid. Taalvaardigheid ontwikkelt in de algemene populatie immers in al zijn facetten gelijkmatig, wat een sterke samenhang tussen de facetten veroorzaakt. Deze samenhang is verstoord in klinische groepen waarbij sprake is van achterstand op één of meerdere facetten. Onderzoek bij deze groepen had de validiteit van de CELF-5-NL en voornamelijk het nut van en de nood aan de verschillende subtests en indexen verder kunnen onderbouwen.

Tot slot werd onderzoek gedaan naar de relatie tussen CELF-5-NL resultaten en scores op **andere, gerelateerde taaltests**. Wanneer verwachte verbanden worden bevestigd, biedt dit evidentie voor de congruente validiteit, een specifiek onderdeel van begripsvaliditeit. Ten eerste werd in een gecombineerde Nederlands-Vlaamse steekproef het verband onderzocht met de *CELF-4-NL*. Er werden matige tot zeer sterke verbanden teruggevonden. Ten tweede werden verbanden met de *Peabody Picture Vocabulary Test (PPVT-III-NL)* onderzocht. Opvallend is hier de matige correlatie met de subtest Linguïstische Concepten (LC; *r*=.56). Aangezien de PPVT-III-NL receptieve woordenschat meet, werden de sterkste verbanden verwacht met CELF-5-NL subtests zoals LC die ook deels begrip meten. Deze onverwacht matige correlatie wordt door de testauteurs verklaard door te stellen dat LC waarschijnlijk voornamelijk geheugen meet. Dit lijkt een heel speculatieve verklaring. De subtest LC pretendeert immers het meten van woordkennis. Bovendien wordt er in de Technische handleiding van de CELF-5-NL niet beschreven dat de subtest LC een beroep doet op het geheugen, zoals dat wel wordt gedaan bij andere subtests (Technische handleiding pg. 23). Tenslotte werden verbanden onderzocht met *de Schlichting tests voor taalproductie en taalbegrip*. De testauteurs stellen dat de meeste verbanden volgens verwachting zijn. Deze conclusie lijkt echter iets te sterk, gezien toch een noemenswaardig aantal onverwachte verbanden werd gevonden, zoals de relatief lage correlaties met de subtest Woordcategorieën (WC). Ook hier zijn de gegeven verklaringen vaak speculatief van aard, zo wordt bij hoge correlaties bijvoorbeeld verwezen naar overeenkomsten in hoe de testen worden aangeboden (bv. respons via prenten aanwijzen versus objecten manipuleren tijdens taalbegrip meting). Verder is de erg kleine overlap in doelgroep tussen de CELF-5-NL en de Schlichting testen (enkel 5-7 jarigen) opvallend. Hierbij kan men zich afvragen of de Schlichting tests wel de goede keuze waren voor het onderzoeken van de congruente validiteit van de CELF-5-NL.

Naast de bovenstaande bemerking kunnen nog een aantal zaken worden opgemerkt met betrekking tot de gehanteerde methodologie. Ten eerste werd bij alle drie de studies gebruikgemaakt van een sterk variabel tijdsinterval tussen de afnames van beide testen (i.e., 0/1-49/56 dagen). Vooral bij de vergelijking met de CELF-4-NL, waar de inhoudelijke overlap met de CELF-5-NL groot is, kan een kort tijdsinterval problematisch zijn in termen van leer- en geheugeneffecten. Een tweede opmerking betreft de gehanteerde steekproeven. Enkel de CELF-4-NL werd bij zowel Nederlandse als Vlaamse kinderen afgenomen, de andere testen enkel bij Nederlandse kinderen. Bovendien, vooral voor de samenstelling van de Vlaamse steekproef, valt het op dat de gerealiseerde samenstelling van de steekproef duidelijk afwijkt van de populatiepercentages voor bepaalde variabelen (bv. meisjes: steekproefpercentage 68.2% versus streefpercentage 50%;

migratieachtergrond: steekproefpercentage 50% versus streefpercentage 24.1%). Verder zijn alle steekproeven klein. Dergelijke kleine steekproeven (*N*<100) worden in de literatuur als onvoldoende beoordeeld (Evers et al., 2013).

Criteriumvaliditeit

Criteriumvaliditeit wordt onderzocht door de samenhang na te gaan tussen de test en een 'real-world' criterium meting. De Technische handleiding van de CELF-5-NL rapporteert over verbanden met schoolresultaten en het diagnostisch onderscheidingsvermogen van de CELF-5-NL voor het vaststellen van een TOS. De samenhang tussen de CELF-5-NL en *schoolresultaten* wordt door de testauteurs benoemd als ondersteuning van de begripsvaliditeit. Wij plaatsen dit echter onder criteriumvaliditeit, aangezien schoolresultaten eerder een real-world meting zijn dan testgedrag (Evers, Boxtel, et al., 2010) en bovendien een weergave zijn van veel bredere constructen dan enkel taalvaardigheid. Over het algemeen bevestigt het onderzoek de validiteit van de CELF-5-NL. Wel is het jammer dat enkel Nederlandse kinderen uit de basisschool werden onderzocht. Daarnaast is het vreemd dat de bevinding dat technisch lezen en rekenen/wiskunde de laagste verbanden vertonen met de CELF-5-NL subtests wordt benoemd als een bevestiging van de verwachtingen. De enige hypothese die werd gesteld was immers dat 'alle schoolse vaardigheden redelijk zouden samenhangen met taalvaardigheid' (Technische handleiding, pg. 131).

Om te onderzoeken of en aan de hand van welke grensscore(s) de CELF-5-NL kan discrimineren tussen kinderen met en zonder een TOS, werden de sensitiviteit en de specificiteit berekend voor een aantal grensscores. Een grensscore van -1.5 *SD*, of dus 77 (voor de Kernscore, ETI en RTI) levert met een sensitiviteit van 86% en specificiteit van 87% behoorlijk goede resultaten op. Bovendien wordt het als positief beoordeeld dat ook de positieve en negatieve voorspellende waarde (PVW en NVW) werden onderzocht voor verschillende prevalenties van een TOS (i.e., 10% voor screening en 50%, 60%, 70% en 80% voor verwijzing). Toch kunnen een aantal opmerkingen gegeven worden over dit onderzoek.

Ten eerste is het onduidelijk hoe de TOS werd gediagnosticeerd bij de kinderen uit de klinische groep. Wanneer hiervoor gebruik werd gemaakt van een CELF-variant, zoals de CELF-4-NL of de CELF-Preschool-2-NL, is het risico op criteriumcontaminatie groot. In dat geval wordt de CELF-5-NL score van een kind met een TOS – bij wie de TOS werd vastgesteld aan de hand van de CELF-4-NL – als bewijs aangehaald voor het goed discrimineren van de CELF-5-NL tussen kinderen met en zonder een TOS. Dit bewijs steunt dus op een cirkelredenering. De testauteurs bevestigen dat bij sommige kinderen uit de klinische groep wellicht een CELF-variant werd ingezet bij het stellen van de initiële TOS diagnose. Ze benadrukken echter dat de diagnose nooit enkel op basis van de CELF-test werd gesteld en zijn van mening dat de verschillen tussen de CELF-5-NL en de CELF-varianten voldoende groot zijn om het risico op criteriumcontaminatie te beperken.

Ten tweede lijkt volgende bewering onaanvaardbaar: "Als de tijd of middelen beperkt zijn, kan ervoor gekozen worden om *enkel* aan de hand van de Kernscore een taalstoornis vast te stellen" (Technische handleiding, pg. 136). Een enkele score is nooit voldoende voor het vaststellen van een stoornis. In Hoofdstuk 1 van de Technisch handleiding (pg. 13) wordt terecht gesteld dat de resultaten van de CELF-5-NL altijd in combinatie met andere (test)resultaten moeten geïnterpreteerd worden, zoals resultaten uit ander onderzoek en gedragsobservaties. De testauteurs zullen de betreffende bewering nuanceren in een volgende druk van de Technische handleiding.

Ten derde moet een opmerking worden gemaakt over de sensitiviteit en specificiteit van de CELF-5-NL. Een Kernscore van 77 heeft inderdaad een behoorlijk goede sensitiviteit en specificiteit. Toch is het opvallend dat met een prevalentie van TOS van 10% (10 kinderen met TOS per 100), een sensitiviteit van 86% impliceert dat 8 à 9 van die 10 correct worden geïdentificeerd, maar 11 à 12 van de 90 zonder TOS als valse positieven onterecht riskeren te worden opgezadeld met een TOS of minstens worden doorverwezen voor verdere diagnose (specificiteit 87% toegepast op de 90 kinderen zonder TOS). Terecht wijzen de auteurs erop dat bij gebruik van dit instrument in geval er een vermoeden van TOS is, de prevalentie in die groep hoger is dan 10% en het risico op valse positieven bijgevolg lager. Deze bedenking geeft echter wel aan dat de CELF-5-NL niet geschikt is om te screenen voor TOS in de algemene populatie. Testgebruikers moeten zich hiervan

terdege bewust zijn en uiterst voorzichtig omspringen met de CELF-5-NL om TOS te detecteren in de algemene populatie.

De hierboven beschreven resultaten leveren evidentie voor de gelijktijdige of retrospectieve criteriumvaliditeit van de CELF-5-NL (Evers et al., 2013). Meer onderzoek is noodzakelijk om ook de voorspellende criteriumvaliditeit van de CELF-5-NL te onderbouwen, waarbij bijvoorbeeld kan onderzocht worden of een CELF-5-NL score schoolresultaten kan voorspellen (in de toekomst).

Aandachtspunten

- Het onderzoek naar de interne structuur van de CELF-5-NL en de confirmatorische factoranalyse wijzen erop dat de CELF-5-NL vooral één grote factor "taalvaardigheid" meet en niet zonder meer in staat is om deelaspecten van die taalvaardigheid te onderscheiden.
- De resultaten van het validiteitsonderzoek kunnen niet overtuigend bevestigen dat er geen geslachtsverschillen bestaan in taalvaardigheid (de scores op de CELF-5-NL). In de Nederlandse steekproef werd gevonden dat meisjes op verschillende onderdelen hoger scoren dan jongens. Verder onderzoek naar mogelijke geslachtsverschillen in voldoende grote steekproeven is noodzakelijk.
- De resultaten met betrekking tot kinderen met dyslexie, lees- en spellingsproblemen en ASS zijn exemplarisch en dienen als ondersteuning van de validiteit. De gevonden groepsverschillen op de CELF-5-NL en CELF-5 (Amerikaanse versie) mogen niet gebruikt worden als rechtvaardiging om de test te gebruiken bij het stellen van een diagnose dyslexie, lees- en spellingsproblemen of ASS.
- Bij gebruik van de CELF-5-NL, met als grensscore 77 voor de Kernscore, de RTI en de ETI, als screeningsinstrument voor het detecteren van taalstoornissen in de algemene populatie is grote voorzichtigheid geboden aangezien in dat geval sprake is van een te hoog aantal vals positieven (bv. 11 à 12 % vals positieven wanneer sprake is van een prevalentie van taalstoornissen van 10%).
- De (voorspellende) criteriumvaliditeit dient nog verder onderzocht te worden.

De **inhoudsvaliditeit** werd als **goed** beoordeeld voor zowel het nemen van *belangrijke* als *minder belangrijke beslissingen op individueel niveau*.

Omwille van bovenstaande aandachtspunten werden zowel de **begripsvaliditeit** als de **criteriumvaliditeit** als **voldoende** beoordeeld voor zowel het nemen van *belangrijke* als *minder belangrijke beslissingen op individueel niveau*.

Kwaliteit van computergegenereerde rapporten

Wanneer de CELF-5-NL digitaal wordt afgenomen via Q-Interactive, of wanneer de ruwe scores verkregen door middel van een papier-en-potlood afname worden ingegeven in Q-Global, kan een digitaal testrapport of voortgangsrapport worden verkregen. Een dergelijk rapport geeft op een objectieve wijze de resultaten weer en leest dus hoofdzakelijk als een automatische berekening van de resultaten, eerder dan als een klinische rapportage of verslag.

Een CELF-5-NL digitaal testrapport bevat vier onderdelen: 1) een samenvatting, 2) een beschrijvende rapportage, 3) een analyse van de opgaven per subtest en 4) een samenvatting van de Checklist Pragmatiek in Activiteiten. De samenvatting geeft tabellen en grafieken voor alle subtests en indexen die rapporteren over de ruwe scores, geschaalde scores of standaardscores, percentielscores, betrouwbaarheidsintervallen, leeftijdsequivalenten en groeiscore-equivalenten. Daarnaast wordt ook een vergelijking gemaakt tussen verschillende indexscores, waarbij wordt weergegeven of de verschilscores significant zijn en wat de prevalentie ervan is in de normgroep. Het tweede onderdeel, de beschrijvende rapportage, bevat gelijkaardige informatie in tekstvorm. Eerst worden de Kern- en indexscores besproken. Er wordt kort beschreven welke taalvaardigheden worden gemeten door een bepaalde index en welke subtests daarvoor werden afgenomen. Nadien wordt aangegeven welke standaardscore het kind behaalde en hoe deze zich verhoudt tegenover de normgroep (bv. “Deze score wijst op prestaties in het gemiddelde gebied van talig functioneren”). Vervolgens wordt ook voor elke subtest beschreven welke taalvaardigheid deze in kaart brengt en welke geschaalde score het kind behaalde. Voor de Checklist Pragmatiek in Activiteiten wordt aangegeven welke score het kind behaalde, en of deze al dan niet voldoet aan ‘het criterium’. Hierbij had het duidelijker geweest wanneer werd gesproken over ‘het criterium voor adequate ontwikkeling’. Het derde onderdeel, de analyse van de opgaven per subtests, kan helpen bij het opsporen van foutpatronen. Zo kan het bijvoorbeeld zijn dat een kind bij de subtest Woordstructuur voornamelijk fouten maakte tegen verkleinwoorden. Tot slot wordt een samenvatting gegeven van de Checklist Pragmatiek in Activiteiten, waarbij de gedragingen die het kind stelde, worden gecategoriseerd als non-verbaal of verbaal (wijze, relevantie of kwaliteit/kwantiteit van communiceren).

Het digitaal testrapport is een handig hulpmiddel. Het geeft een duidelijk en volledig overzicht van de scores van het kind, inclusief grafische voorstelling. Bovendien vereenvoudigt de automatische scoring het administratieve werk van de testleider aanzienlijk. Toch dienen een aantal opmerkingen te worden meegegeven:

- Het rapport is erg uitgebreid. De lengte, en voornamelijk de herhaling van informatie op verschillende plaatsten in het rapport vermindert de gebruiksvriendelijkheid.
- De enige interpretatie van de prestaties waarin het rapport voorziet, is een beperkte vergelijking van de kern- en indexscores van het kind met de normgroep (bv. “Deze score wijst op prestaties in het gemiddelde gebied van talig functioneren”). Een overzicht van welke scores op welke interpretatie wijzen (bv. standaardscore tussen 85 en 115 = gemiddeld) zou meer transparantie bieden. Het digitaal testrapport kan in geen enkele zin een logopedisch verslag vervangen.
- Het rapport kan de schijn wekken dat de CELF-5-NL een (sterk) discriminerend vermogen heeft wat betreft verschillende aspecten van taalvaardigheid (subtest en indexscores). Zoals hoger beschreven toont het onderzoek aan dat er voornamelijk sprake is van één overkoepelende taalfactor en de CELF-5-NL niet zonder meer in staat is verschillende deelaspecten te onderscheiden. Een korte waarschuwing in het rapport die hierop wijst, is welkom.
- Er wordt niet gerapporteerd hoe groot de consistentie is tussen manuele en digitale scoring en op welke wijze het systeem wordt gecontroleerd, vermoedelijk vanwege de behoorlijk eenvoudige verwerking. Met name, ruwe scores worden door de digitale applicatie opgeteld en omgezet naar normscores en percentielen. De testauteurs lieten weten dat deze omzetting uitvoerig gecontroleerd werd in een geautomatiseerd kwaliteitscontroleproces.

De **kwaliteit van computergegeneerde rapporten** werd als **goed** beoordeeld voor zowel het nemen van *belangrijke* als *minder belangrijke beslissingen op individueel niveau*. Het is hierbij wel belangrijk op te merken

dat het rapport louter beschrijvend is en geenszins vervanging kan bieden voor een logopedisch/klinisch verslag of advies.

Samenvatting beoordeling

Het voorliggende adviesrapport handelt over de evaluatie van de Clinical Evaluation of Language Fundamentals – versie 5 – Nederlandstalige versie (CELF-5-NL). Het rapport heeft voornamelijk betrekking op het gebruik en de normering van de CELF-5-NL in Vlaanderen. De beoordeling voor het gebruik en de normering in Nederland gebeurde door de Commissie Testaangelegenheden Nederland (COTAN) en kan worden geraadpleegd via: https://www.cotandocumentatie.nl/beoordelingen/.

Onderstaande kwaliteitslabels werden toegekend aan de CELF-5-NL voor gebruik in Vlaanderen door het Kwaliteitscentrum voor Diagnostiek vzw na een grondige beoordeling van de test door twee onafhankelijke experten en een wetenschappelijk onderzoeker van het Kwaliteitscentrum zelf. Voor de beoordeling werd gebruik gemaakt van het beoordelingsmodel voor de beschrijving en evaluatie van psychologische en educatieve testen van de European Federation of Psychologists' Associations (EFPA; Evers, et al., 2013), naar het Nederlands vertaald door het Kwaliteitscentrum voor Diagnostiek vzw.

Beoordeling CELF-5-NL voor gebruik in Vlaanderen

	*Relatief minder belangrijke beslissingen**	*Belangrijke beslissingen***
Kwaliteit uitgangspunten, presentatie en beschikbare informatie	Goed	Goed
Kwaliteit van het testmateriaal	Goed	Goed
Vlaamse normen	Goed	Voldoende
Criteriumgerichte normering van de PAC	Onvoldoende	Onvoldoende
Betrouwbaarheid	Goed	Goed
Inhoudsvaliditeit	Goed	Goed
Begripsvaliditeit	Voldoende	Voldoende
Criteriumvaliditeit	Voldoende	Voldoende
Kwaliteit van computergegenereerde rapporten	Goed	Goed

* Voorbeelden van relatief minder belangrijke beslissingen zijn: voortgangscontrole, beschrijvend gebruik van de testresultaten, therapie-indicatie en beroepskeuzebegeleiding (Evers, Boxtel, et al., 2010, pg. 22).
** Met belangrijke beslissingen wordt bedoeld: beslissingen die op basis van de testscores worden genomen, die in principe, of op korte termijn, onomkeerbaar zijn, en die voor een belangrijk deel buiten de geteste persoon om worden genomen. Voorbeelden van belangrijke beslissingen zijn: personeelsselectie, verwijzing naar speciaal onderwijs, opname/ontslag kliniek, certificering (Evers, Boxtel, et al., 2010, pg. 22).

Errata

Technische handleiding

- pg. 14 (1.1): In de volgende zin moet 'de kind' vervangen worden door 'het kind': "*ook worden per subtest mogelijkheden gegeven voor het doen van aanvullend onderzoek en suggesties voor dynamische afnameprocedures om te achterhalen op welk niveau de kind het best presteert.*"
- pg. 28 (1.5.7): De leeftijdsrange die wordt vermeld bij de subtest Zinnen formuleren (i.e., 5-8 jaar) is niet correct. Deze subtest kan immers worden afgenomen bij kinderen en jongeren van 5 tot en met 18 jaar.
- pg. 28/29 (1.5.7): De zinsconstructie van de volgende zin is niet correct: "*Toegevoegd werden een voornaamwoord en een voorzetsel…*" Een correctere zinsconstructie zou zijn: "Een voornaamwoord en een voorzetsel werden toegevoegd…"
- pg. 33 (1.5.12): De volgende zin is moeilijk te begrijpen: *"Bij Semantische Relaties gaat het om de verwerking en interpretatie van contrasterende zinnen op basis van voor de vergelijking kritische woorden."*
- pg. 35 (1.5.14): Bij de beschrijving van de subtests staat telkens het aantal items vermeld, met uitzondering van bij de Checklist Pragmatiek in Activiteiten (PAC). Het aantal items van de PAC is 35.
- pg. 50 (3.5): In de volgende zin is het woord 'voor' overbodig: *"… van voldoende grootte voor om betrouwbare statistische analyses mogelijk te maken."*
- pg. 50 (3.5): In de volgende zin ontbreekt er een komma tussen 'Vlaanderen: 3e graad secundair onderwijs' en 'getuigschrift bso': "*Als indicator voor sociaaleconomische status is het hoogst afgeronde opleidingsniveau van de moeder gehanteerd, verdeeld in de niveaus Laag (Nederland: lager onderwijs, vmbo, mbo 1; Vlaanderen: geen, basisonderwijs, 1e graad secundair onderwijs, 2e graad secundair onderwijs), Midden (Nederland: havo 3-5, vwo 3-6, mbo 2-4; Vlaanderen: 3e graad secundair onderwijs getuigschrift bso, 4e graad secundair onderwijs) en Hoog (Nederland: hbo, universiteit; Vlaanderen: professioneel gerichte bacheloropleiding (hogeschool), academisch gerichte bacheloropleiding (hogeschool of universiteit), masteropleiding (hogeschool of universiteit)).*"
- pg. 76 (3.8.4): Eén van de referenties waarnaar wordt verwezen bij inferentiële normering, namelijk Zhu & Chen (2011), wordt niet weergegeven in de referentielijst (pg. 156).
- pg. 87 (3.8.10): De volgende term is niet correct geschreven: "*Rash-model*" moet Rasch-model zijn.
- pg. 140 (5.14.2): De volgende term is niet correct geschreven: "*paragraf*" moet 'paragraaf' zijn.

Afnamehandleiding

- pg. 110 (3.9.4): Er ontbreekt een 'en' in de volgende zin: *"U berekent de boven- en ondergrens van het betrouwbaarheidsintervallen door deze waarden op te tellen bij en respectievelijk af te trekken van de geschaalde subtestscore."*
- pg. 137 (4.5.1): De volgende zinsconstructie is niet correct: *"Het kan zijn dat een kind nog niet lang genoeg in behandeling is geweest om een verandering zien te geven als gevolg van de interventie."* Een correctere zinconstructie zou zijn: "Het kan zijn dat een kind nog niet lang genoeg in behandeling is geweest om een verandering te zien als gevolg van de interventie."

Online

- Op de website van Pearson (https://www.pearsonclinical.nl/tests/celf-5-nl-test-diagnose-evaluatie-taalproblemen) staat bij de samenvatting van de informatie van de CELF-5-NL vermeld: *Leeftijd: 4-12 jaar (Basisschool), 12+ jaar (Jongeren/Adolescenten)*. Dit komt niet volledig overeen met de informatie uit de handleidingen, waaruit duidelijk blijkt dat de CELF-5-NL een test is die gebruikt kan worden bij kinderen en jongeren met een leeftijd van 5 tot en met 18 jaar.

Referenties

Bechger, T., Hemker, B., & Maris, G. (2009). *Over het gebruik van continue normering*. Cito.

Daniel, M. H., Wahlstrom, D., & Zhou, X. (2014). *Equivalence of Q-interactive ® and Paper Administrations of Language Tasks: Selected CELF ®-5 Tests Q-interactive Technical Report 7*.

Evers, A., Boxtel, H. W. Van, Dinger, M. E., Hemker, B. T., Kersten, W. W., Lucassen, W. I., Meijer, R. R., Evers, A., & Lucassen, W. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests*.

Evers, A., Hagemeister, C., Høstmaelingen, A., Lindley, P., Muñiz, J., & Sjöberg, A. (2013). *EFPA REVIEW MODEL FOR THE DESCRIPTION AND EVALUATION OF PSYCHOLOGICAL AND EDUCATIONAL TESTS TEST*.

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch Review Process for Evaluating the Quality of Psychological Tests: History, Procedure, and Results. *International Journal of Testing*, *10*, 295–317. https://doi.org/10.1080/15305058.2010.518325

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (Thirth). The Guilford Press.

Lenhard, A., Lenhard, W., & Gary, S. (2019). Continuous norming of psychometric tests: A simulation study of parametric and semi-parametric approaches. *PloS One*. https://doi.org/10.1371/journal.pone.0222279

Leonard, L. B. (2009). Is Expressive Language Disorder an Accurate Diagnostic Category? *Journal of Speech-Language Pathology*, *8*, 115–112. https://doi.org/10.1044/1058-0360(2008/08-0064

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). Jerome M. Sattler, Publisher, Inc.

Siméa. (2014). *Indicatiecriteria auditief en / of communicatief beperkte leerlingen*. 1–20. http://www.simea.nl/dossiers/passend-onderwijs/brochures-po/

Statistics Belgium (Statbel). (2017). *Structuur van de bevolking*. http://statbel.fgov.be/nl/statistieken/Cijfers.

Tomblin, J. B., & Zhang, X. (2006). The Dimensionality of Language Ability in School-Age Children. *Journal of Speach, Language, and Hearing Research*, *49*, 1193–1208. https://doi.org/10.1044/1092-4388(2006/086

Vlaams Departement Onderwijs en Vorming, V. (2015). *Vlaams onderwijs in cijfers 2014-2015*. https://www.vlaanderen.be/publicaties/vlaams-onderwijs-in-cij%0Afers-2014-2015 Het Vlaams Ministerie van Onderwijs en Vorming (2017a). Leerlingkenmerken.

Wang, K., Xu, Y., Wang, C., Tan, M., & Chen, P. (2019). A Corrected Goodness-of-Fit Index (CGFI) for Model Evaluation in Structural Equation Modeling. *Structural Equation Modeling*, 1–15. https://doi.org/10.1080/10705511.2019.1695213