

**Test-Hertest Betrouwbaarheid van  
Meetinstrumenten voor het Meten van Executieve  
Functies bij Normaal Ontwikkelde Kleuters**

Thesis Pedagogische Wetenschappen (200600042)

Universiteit Utrecht

Werkgroep: 19

Datum: 25 juni 2017

Begeleider: Lianne van Gelder

Tweede beoordelaar: Marloes Röttger

Door: Anna Warmenhoven - 5730767

Helma van Nuenen - 5673232

### **Abstract**

This study aimed to give insight into the test-retest reliability of a testing battery for the executive functions (EF) used for normal developed pre-schoolers in the Netherlands. This testing battery measured the EF inhibition, cognitive flexibility and working memory. The measuring instruments used in this testing battery are the Corsi block-tapping task, Dimensional Change Cart Sort, Updating Number task and the Head-Shoulders-Knees-Toes. This research is relevant for daily practice because pre-schoolers with EF-problems can be trained to improve their EF (Alloway, 2006). To be able to identify these problems, reliable measuring instruments are of great importance (Alloway, 2006; Diamond, 2013). With a repeated measurement with an interval of one week, 38 pre-schoolers from 48 to 84 months old were tested. The results indicated good test-retest reliability for the testing battery. Also the reliability for the individual tests has shown to be sufficient. When splitting the sample into three age groups, for a number of measuring instrument's reliability was found to be insufficient. Noteworthy, the test-retest reliability for the youngest pre-schoolers (48-60 months) was insufficient for all measuring instruments. This raises the question whether a stable measurement of EF is possible in preschool children of this age. Because of the limited presence of significant correlations between the measuring instruments itself, the measuring instruments cannot be viewed as one overarching testing battery. Recommendations for future research are to replicate this research on a larger random sample and to examine a different composition of EF measuring instruments to obtain a reliable testing battery.

*Keywords:* executive functions, inhibition, cognitive flexibility, working memory, tests, test-retest reliability, normally developed pre-schoolers in the Netherlands

### **Samenvatting**

Dit onderzoek is gericht op het geven van inzicht in de test-hertest betrouwbaarheid van een testbatterij voor executieve functies (EF) voor normaal ontwikkelde kleuters in Nederland. De onderzochte EF in deze testbatterij zijn inhibitie, cognitieve flexibiliteit en werkgeheugen. De meetinstrumenten in deze testbatterij zijn de Corsi-blokkentest, Dimensional Change Cart Sort, Updating Cijfertaak en de Hoofd-Schouders-Knieën-Tenen. Dit onderzoek is van belang omdat problemen met EF al in de kleutertijd getraind en verbeterd kunnen worden (Alloway, 2006). Om deze problemen te kunnen signaleren zijn betrouwbare meetinstrumenten van groot belang (Alloway, 2006; Diamond, 2013). Bij een herhaalde meting met een interval van één week werden 38 kleuters van 48 tot 84 maanden oud onderzocht. De resultaten tonen een goede test-hertest betrouwbaarheid van de gehele testbatterij en voldoende tot goed bij alle meetinstrumenten. Bij het splitsen van de steekproef in drie leeftijdsgroepen is voor verschillende meetinstrumenten onvoldoende betrouwbaarheid gevonden. Opvallend is dat de betrouwbaarheid voor de jongste kleuters (48-60 mnd.) bij ieder meetinstrument

onvoldoende was. Dit heeft de vraag opgeroepen of een stabiele meting van EF mogelijk is bij deze leeftijdsgroep. Door de beperkte aanwezigheid van significante correlaties kon niet worden gesproken van één overkoepelende testbatterij. Aanbevelingen voor vervolgonderzoek zijn om een grotere steekproef op een aselechte manier te werven, specifiek onderzoek te doen naar de jongste leeftijdsgroep en naar een testbatterij met EF-meetinstrumenten die beter met elkaar correleren.

*Trefwoorden:* executieve functies, inhibitie, cognitieve flexibiliteit, werkgeheugen, test-hertest betrouwbaarheid, normaal ontwikkelde kleuters in Nederland

#### Test-Hertest Betrouwbaarheid van Meetinstrumenten voor het Meten van Executieve Functies bij Normaal Ontwikkelde Kleuters

Er is de afgelopen jaren veel vooruitgang geboekt in het begrijpen van EF bij jonge kinderen. Uit onderzoek in 2002 van Casey, Tottenham en Rossella bleek al dat problemen in de ontwikkeling van EF vaak comorbide voorkomen met sociale en academische problemen. In 2007 verklaarden Blair en Razza dat cognitieve processen die betrokken zijn bij EF een grote rol spelen in het toe-eigenen van kennis en bij kleuters een goede voorspeller zijn van toekomstige schoolprestaties. In 2013 concludeerde Diamond dat EF cruciaal zijn om te kunnen leren.

EF zijn processen die nodig zijn voor het uitvoeren van complexe taken zoals het oplossen van problemen, bepalen van een strategie, richten van de aandacht en vooruit denken (Smidts, 2003). EF kennen drie kernfuncties. Inhibitie is de vaardigheid om dominante responses te onderdrukken. Cognitieve flexibiliteit is het vermogen om van aanpak te wisselen als de omstandigheden veranderen of wanneer opgemerkt wordt dat de aanpak niet succesvol is (Miyake et al., 2000). Het werkgeheugen kan korte tijd een beperkte hoeveelheid informatie onthouden, bewerken en vervangen (Gathercole & Alloway, 2007). EF bevinden zich in de frontale kwab van de hersenen en maken in de kleutertijd een versnelde ontwikkeling door (Carlson, 2005). Dit komt volgens Garon, Bryson en Smith (2008) door de ontwikkeling van gerichte aandacht en de integratie van verschillende componenten van EF in die periode. Volgens Alloway, Gathercole en Pickering (2006) kunnen EF bij kleuters al vanaf 4 jaar betrouwbaar worden gemeten. In deze periode blijken EF een betere voorspeller voor toekomstige schoolprestaties dan het intelligentiequotiënt (Alloway & Alloway, 2010). Volgens Alloway (2006) zijn EF goed trainbaar met speciale lesprogramma's op school. Om EF bij kleuters in kaart te kunnen brengen en meetinstrumenten in te kunnen zetten voor diagnostiek is een betrouwbare beoordeling van EF-meetinstrumenten van groot belang (Beck, Schaëfer, Pang, & Carlson, 2011; Neuman, 2009).

#### **Leereffect, Effect van Leeftijd en Test-Hertest Betrouwbaarheid**

Een betrouwbaar meetinstrument meet nauwkeurig en genereert stabiele scores

bij een herhaalde testafname. De betrouwbaarheid van een meetinstrument kan worden gemeten met de test-hertest methode (Tak, Bosch, Begeer & Albrecht, 2014). Hierbij wordt gekeken of er voldoende overeenkomst is tussen de scores van de eerste testafname (T1) en de tweede testafname (T2). Wanneer scores van T1 en T2 in hoge mate overeenkomen heeft een meetinstrument een hoge test-hertest betrouwbaarheid (Gray, 2003). Bij een herhaalde meting kan sprake zijn van een leereffect, dit vertekent de onderzoeksresultaten en heeft een negatieve invloed op de test-hertest betrouwbaarheid van een meetinstrument. Het treedt op als de prestatie van de participant wordt beïnvloed vanwege bekendheid met het meetinstrument door T1 waardoor de resultaten van T2 niet meer zuiver zijn. Een ander veel voorkomend systematisch verschil bij een herhaalde meting is het effect van verveling (Field, 2013). Lezak (1995) vindt bij herhaalde testafnamen van EF-meetinstrumenten een leereffect. Het Effect van Leeftijd ontstaat wanneer participanten uit één leeftijdsgroep significant anders scoren dan participanten uit een andere leeftijdsgroep. Carlson (2005) en Garon en collega's (2008) stellen dat door de versnelde ontwikkeling van EF in de kleutertijd, leeftijd mogelijk van invloed is op de betrouwbaarheid. Volgens Willoughby en Blair (2011) is deze invloed er niet. De bestaande literatuur is niet eenduidig over de invloed van leeftijd op de betrouwbaarheid van EF-meetinstrumenten voor kleuters (Beck et al., 2011; Müller, Kerns & Konkin, 2012; Willoughby & Blair, 2011).

Ondanks dat EF-meetinstrumenten en testbatterijen wereldwijd worden gebruikt bij kinderen vanaf 4 jaar (Gathercole & Alloway, 2007) heeft eerder onderzoek naar de betrouwbaarheid van deze meetinstrumenten zich met name gericht op oudere kinderen, volwassenen en klinische patiënten. Daarnaast is onderzoek naar EF bij normaal ontwikkelde kleuters voornamelijk uitgevoerd met kleine steekproeven. In Nederland is in beperkte mate onderzoek gedaan naar EF-meetinstrumenten bij normaal ontwikkelde kleuters (Smidts, 2003) en is ook nog geen testbatterij beschikbaar. De Universiteit Utrecht is het project Kleuter Extra gestart. Dit is een ondersteuningsaanbod voor kleuters die moeite hebben met het volgen van het lesaanbod of het functioneren in de klas. Om bij deze kleuters een sterkte-zwakteanalyse te kunnen maken heeft Kleuter Extra een aantal meetinstrumenten voor EF aangepast en nieuw ontwikkeld voor kleuters. De EF meetinstrumenten zijn de Corsi-blokkentest (CBT), Dimensional Change Cart Sort (DCCS) en de Hoofd-Tenen- Knieën-Schouders (HTKS). De nieuw ontwikkelde taak is de Updating Cijfertaak (UC).

**UC.** De UC meet het werkgeheugen en de cognitieve flexibiliteit. Omdat de UC een nieuw meetinstrument is, zijn hier nog geen psychometrische kwaliteiten van bekend. Dit onderzoek wil hier meer inzicht in krijgen.

**CBT.** De CBT is ontworpen door Corsi (1972) en meet het visueel-ruimtelijk kortetermijngeheugen en het werkgeheugen (Baddeley, 2001). Het is een bestaande

subtest uit de Wechsler Nonverbal Scale of Ability (WNV-NL) voor kinderen vanaf 8 jaar. Volgens Diamond (2013) meet de voorwaartse CBT-subtest alleen het kortetermijngeheugen en niet het hele werkgeheugen. Onderzoek van Vandierendonck, Kemps, Fastame en Szmalec (2004) toont aan dat bij een test-hertest met een interval van één week, vooral bij de voorwaartse CBT-subtest, duidelijk sprake is van een leereffect. Volgens de Commissie Testaangelegenheden Nederland (COTAN) is de betrouwbaarheid van de WNV-NL voldoende. De subtests, waaronder de CBT, zijn niet apart onderzocht en daarom als onvoldoende beoordeeld (Jonkman, Kooij, Wechsler & Naglieri 2008). Een doel van dit onderzoek is dan ook om meer kennis te verkrijgen over de betrouwbaarheid van de CBT en de voor- en achterwaartse subtests.

**DCCS.** De DCCS is ontworpen door Zelazo (2006) en meet cognitieve flexibiliteit. De betrouwbaarheid van de DCCS is niet door COTAN beoordeeld, maar Beck en collega's (2011) vinden bij 2;6-4;6 jarigen een hoge betrouwbaarheid. Lezak (1995) stelt dat de betrouwbaarheid van instrumenten zoals de DCCS die cognitieve flexibiliteit meten moeilijk te beoordelen is, omdat het succes van de test afhangt van de mogelijkheid tot het ontdekken van de sorteerregel. Wanneer een participant de verandering van regels eenmaal heeft ontdekt is het onwaarschijnlijk dat deze een volgende keer opnieuw dezelfde fout maakt. Er zal worden onderzocht of er sprake is van dit leereffect.

**HTKS.** De HTKS is ontworpen door Ponitz en collega's (2008) en meet inhibitie en cognitieve flexibiliteit. De HTKS wordt in onderzoek van Ponitz, McClelland, Matthews en Morrison (2009) als betrouwbaar instrument beoordeeld. Doordat de HTKS niet is beoordeeld door COTAN en er mogelijk sprake is van belangenverstremgeling van de auteur, is er onvoldoende inzicht in de betrouwbaarheid van de HTKS. Dit onderzoek zal een bijdrage leveren aan het vergroten van dit inzicht.

**Testbatterij (TB).** Uit onderzoek van Willoughby en Blair (2011) blijkt dat bij kleuters de betrouwbaarheid van een EF-testbatterij hoger is dan van losse EF-meetinstrumenten. Dit kan verklaard worden doordat de TB uit meer items bestaat dan de individuele meetinstrumenten en de betrouwbaarheid daarmee automatisch hoger uitvalt (Evers et al., 2010). De TB wordt meegenomen in dit onderzoek om deze resultaten mogelijk te ondersteunen.

**Leeftijd.** Beck en collega's (2011) en Willoughby en Blair (2011) vonden in hun onderzoeken voor alle leeftijdsgroepen van kleuters een goede betrouwbaarheid voor EF-meetinstrumenten. In het onderzoek van Müller, Kerns en Konkin (2012) wordt alleen voor inhibitie taken een lagere test-hertest betrouwbaarheid gevonden wanneer kleuters ouder worden. Zij geven hiervoor de verklaring dat de motivatie voor deelname aan deze testen mogelijk afneemt. Door het effect van verveling (Field, 2013) kunnen de testscores negatief worden beïnvloed. Mogelijk speelt ook een leereffect bij deze oudste participanten een rol. Er zal worden onderzocht wat de invloed van leeftijd is en of de

test-hertest betrouwbaarheid bij alle leeftijdsgroepen voldoende is.

### **Huidig Onderzoek**

Op basis van bovenstaande literatuur is in dit onderzoek gekeken naar de test-hertest betrouwbaarheid van vier meetinstrumenten, de voor- en achterwaartse subtest van de CBT en de gehele TB (variabelen) door middel van een herhaalde meting met een test interval van één week. Voor alle variabelen werd op basis van bovenstaande literatuur een leereffect verwacht. Ook werd verwacht dat de betrouwbaarheid van de TB hoger zou zijn dan van de meetinstrumenten (Willoughby & Blair, 2011). Tevens is onderzocht of de test-hertest betrouwbaarheid van de meetinstrumenten bij alle leeftijden voldoende was. Dit fundamentele onderzoek heeft als doel gehad kennis te vergroten (Tijmstra & Boeije, 2011) over de test-hertest betrouwbaarheid van EF-meetinstrumenten bij normaal ontwikkelde kleuters van 48 tot 84 maanden oud in Nederland. Het doel van dit onderzoek was voor de bestaande EF-meetinstrumenten evaluerend en voor de nieuwe Updatingtaken explorerend (Neuman, 2009).

### **Methode**

#### **Participanten**

De populatie van dit onderzoek bestond uit normaal ontwikkelde kleuters uit groep 1 en 2 van reguliere basisscholen uit vijf Nederlandse provincies. De participanten ( $n = 40$ ) maakten deel uit van een grootschalig onderzoek ( $n = 240$ ) naar EF-meetinstrumenten voor kleuters. De participanten zijn geworven middels een quota steekproef, een praktische en snelle selecte steekproefmethode waarbij is geselecteerd op basis van het inclusiecriteria: leerlingen tussen de 48 en 84 maanden oud. Door deze selecte steekproefmethode kan mogelijk sprake zijn van verminderde representativiteit van de populatie (Neuman, 2009).

Op basis van extreme scores zijn twee participanten uit het databestand verwijderd. De uiteindelijke steekproef bestond uit 38 participanten ( $M_{leeftijd} = 65.58$ ,  $SD_{leeftijd} = 7.30$ ), waarvan 19 meisjes (50%) en 19 jongens (50%). De participanten zijn ingedeeld in drie gelijke leeftijdsgroepen van ieder 12 maanden, namelijk: 48 tot 60 maanden ( $n = 9$ ), 60 tot 72 maanden ( $n = 17$ ) en 72 tot 84 maanden oud ( $n = 12$ ).

#### **Procedure**

Dit onderzoek is uitgevoerd door vier studenten Orthopedagogiek van de Universiteit Utrecht. Deelname kon alleen na informed consent van de ouders/verzorgers. Voordat het onderzoek werd gestart is een pilot uitgevoerd met drie door KleuterExtra ontwikkelde EF-meetinstrumenten: de Kleuren Updatingtaak, UC en de Complexe Updatingtaak. Naar aanleiding van deze pilot is de UC opgenomen in de testbatterij. Alle meetinstrumenten zijn met behulp van een gestandaardiseerde handleiding afgenomen. Hierin zijn de volgorde, manier van testafname, scores en reacties van de testleider op de antwoorden van de participanten gedetailleerd beschreven. Voorafgaand aan T1 zijn

van alle participanten demografische gegevens ingewonnen. T2 heeft één week na T1 plaatsgevonden. Ondanks de verhoogde kans op een leereffect (Ozonoff, 1995) is gekozen voor dit korte interval omdat de snelle ontwikkeling van EF bij kleuters (Carlson, 2005) geen invloed mocht hebben op de testcores. Beide testafnamen zijn uitgevoerd door dezelfde testleider in een aparte prikkelarme ruimte in de school en duurde ongeveer 35 minuten per participant. De vier meetinstrumenten zijn in vaste volgorde afgenomen: CBT, DCCS, UC en de HTKS. De gegevens zijn anoniem verwerkt met kindnummers en niet toegankelijk voor derden.

### **Meetinstrumenten**

**Demografische gegevens van participanten.** Per participant is de geboortedatum en het geslacht genoteerd. De leeftijd is gemeten in maanden.

**TB.** De TB heeft één algemene totaalscore die is samengesteld uit de totaalscores van de CBT, DCCS, HTKS, en de UC. Cronbach's Alpha van T1 ( $\alpha = .89$ ) en T2 ( $\alpha = .93$ ) zijn zeer betrouwbaar.

**CBT.** De CBT bestaat uit een plank met negen blokken en bevat twee oefeningen, vier trials en vier tests. Het aantal blokpatronen loopt per test op van twee tot zes of negen blokken. Het afkappunt is twee opeenvolgende onjuiste reeksen binnen één patroonlengte. Alle juiste antwoorden worden met één punt gescoord. De testcores vormen samen de totaalscore (Kessels, van den Berg, Ruis, & Brands, 2008). De testleider heeft blokken aangewezen en de participant gevraagd dit voorwaarts of achterwaarts te herhalen. De participant heeft voor iedere blok lengte twee pogingen gekregen. Bij T1 is Cronbach's Alpha voor de CBT ( $\alpha = .82$ ), de voorwaartse- ( $\alpha = .71$ ) en achterwaartse ( $\alpha = .77$ ) CBT-subtest voldoende betrouwbaar. Dit geldt ook voor T2 CBT ( $\alpha = .84$ ), de voorwaartse- ( $\alpha = .73$ ) en achterwaartse ( $\alpha = .76$ ) CBT-subtest.

**DCCS.** De DCCS bestaat uit drie testronden. Voor de derde testronde is het afkappunt participanten die jonger zijn dan 5 jaar en die in de eerste twee tests meer dan twee onjuiste antwoorden hebben gegeven. Alle juiste antwoorden worden met één punt gescoord. De testcores vormen samen de totaalscore. In de eerste testronde is de participant gevraagd kaarten te sorteren op kleur en in de tweede testronde op vorm. In de derde testronde is aan de hand van een zwarte rand op het kaartje aangegeven of de participant de sorteerregel van kleur of vorm dient te gebruiken. Door onvoldoende variantie in testcores is bij T1 alleen Cronbach's Alpha voor de gevorderde testitems berekend. Zowel T1 ( $\alpha = .63$ ) als T2 ( $\alpha = .58$ ) zijn onvoldoende betrouwbaar.

**UC.** De UC heeft zes niveaus met in totaal 26 items, waarvan vijf oefenitems. Het afkappunt is meer dan twee onjuiste antwoorden binnen een niveau. De scores per niveau vormen samen de totaalscore. De testleider heeft kaarten met stippen laten zien die na het tonen aan de participant omgekeerd op tafel werden gelegd. Er zijn kaarten toegevoegd en weggenomen. Een juist antwoord op de vraag: 'Hoeveel stippen liggen

hier?' is met één punt gescoord. Cronbach's Alpha van T1 ( $\alpha = .95$ ) en T2 ( $\alpha = .95$ ) zijn zeer betrouwbaar.

**HTKS.** De test bestaat uit drie testrondes met tien imitatie oefeningen, acht oefenitems en 20 testvragen. Tijdens de imitatie oefeningen en oefenitems mag de uitleg drie keer worden herhaald. Er is geen afkappunt. Alle juiste antwoorden worden met twee punten gescoord en zelfcorrectie met één. De scores per niveau vormen samen de totaalscore. De participant is gevraagd een spel te spelen waarin deze het tegenovergestelde lichaamsdeel aan moest raken dan de testleider heeft gevraagd. Wanneer de testleider de participant vroeg het hoofd aan te raken, diende deze de tenen aan te raken en andersom. Wanneer de testleider de participant vroeg de schouders aan te raken, diende deze de knieën aan te raken en andersom. Cronbach's Alpha van T1 ( $\alpha = .82$ ) en T2 ( $\alpha = .82$ ) zijn voldoende betrouwbaar.

### Analyseplan

Alle data is gecontroleerd en fouten zijn gecorrigeerd. Na controle op de aanwezigheid van extreme scores, die voor vertekening van de analyse zorgen, zijn twee participanten in de boxplots van de HTKS naar voren gekomen als significante uitschieters en uit het databestand verwijderd. Om meer inzicht te krijgen in de interne consistentie van de variabelen is Cronbach's Alpha berekend. Een meetinstrument heeft voldoende interne betrouwbaarheid bij een waarde van .7 of hoger (Field, 2013). Voor controle van de assumpties voor parametrische correlatieanalyse is het meetniveau, de lineariteit en de mate van scheefheid van de scores getoetst. Alle data is van interval meetniveau met een lineaire relatie tussen T1 en T2. Voor berekening van de scheefheid van de verdeling van de variabelen is de skewness ( $SK$ ) gedeeld door de standaardfout ( $SE$ ). Bij de CBT T1 ( $SK = .58, SE = .38$ ) en T2 ( $SK = .57, SE = .38$ ), HTKS T1 ( $SK = -1.3, SE = .38$ ), UC T1 ( $SK = -0.4, SE = .38$ ) en T2 ( $SK = -1.7, SE = .38$ ) en de TB T1 ( $SK = -.33, SE = .38$ ) en T2 ( $SK = -.28, SE = .38$ ) vallen de  $z$ -waarden binnen de  $\pm 1.96$  en is voldaan aan de assumptie van normaliteit. De scores van HTKS T2 ( $SK = -2.06, SE = .38$ ), DCCS T1 ( $SK = .90, SE = .38$ ) en de DCCS T2 ( $SK = .95, SE = .38$ ) voldoen niet aan deze assumptie. Desondanks mogen ze volgens de centrale limietstelling, bij een steekproef van 38 participanten, wel beschouwd worden als normaal verdeeld (Field, 2013).

Van alle variabelen is het gemiddelde ( $M$ ), de standaarddeviatie ( $SD$ ) en de  $z$ -score ( $z$ ) berekend. Om te toetsen of de gehele TB mee mocht worden genomen in de test-hertest betrouwbaarheid is gecontroleerd of de meetinstrumenten onderling significant correleren met de Pearson correlatie analyse. Met een gepaarde  $t$ -toets (tweezijdig) is de significantie van het verschil in gemiddelden tussen T1 en T2 per variabele en leeftijdsgroep getoetst en gekeken of er sprake was van een leereffect. Met de Eenweg ANOVA (eenzijdig) en de Turkey Post Hoc toets is getoetst of er significante



verschillen in gemiddelden zijn tussen de leeftijdsgroepen, een effect van leeftijd, en welke leeftijdsgroepen voor dit verschil hebben gezorgd. Voor het bepalen van de test-hertest betrouwbaarheid zijn de richting van de lineaire relatie en de correlatie tussen T1 en T2 van de variabelen berekend met de Pearson correlatie analyse (eenzijdig). Bij het beoordelen van de test-hertest betrouwbaarheid is gebruik gemaakt van de criteria van COTAN voor meetinstrumenten op individueel niveau. Een correlatiecoëfficiënt ( $r$ ) tussen .70 en .80 is voldoende, boven de .80 is goed (Evers, Lucassen, Meijer & Sijtsma, 2010). Om het percentage overeenstemming in scores tussen T2 en T2 vast te stellen is de effectgrootte bepaald ( $R^2$ ). Met de Fisher's  $r$ -to- $z$  transformatie is gecontroleerd of er een significant verschil is in de betrouwbaarheid bij de verschillende leeftijdsgroepen.

## Resultaten

### Beschrijvende statistieken

In Tabel 1 staan de beschrijvende statistieken van alle variabelen in de analyses voor T1 en T2 per leeftijdsgroep. Daarnaast is de significantie van het verschil in gemiddelden ( $F$ ) tussen de leeftijdsgroep van T1 en T2 opgenomen.

Tabel 1

*Beschrijvende Statistieken en Verschillen in Gemiddelden tussen T1 en T2 van EF-Meetinstrumenten, CBT-subtests en TB per Leeftijdsgroep (LG)*

Meetinstrumenten	T1	T2	$F_{T1}$	$F_{T2}$
	$M$ ( $SD$ )	$M$ ( $SD$ )		
CBT	9.47 (2.98)	9.71 (3.17)	7.70**	9.01**
LG1 <sup>a</sup>	7.11 (1.69)	7.56 (1.13)		
LG2 <sup>b</sup>	9.29 (2.52)	9.06 (2.97)		
LG3 <sup>c</sup>	11.50 (3.06)	12.25 (2.93)		
CBT voorwaarts	5.76 (1.70)	5.76 (1.84)	10.72***	9.17**
LG1 <sup>a</sup>	4.11 (1.05)	4.11 (1.36)		
LG2 <sup>b</sup>	5.82 (1.19)	5.76 (1.68)		
LG3 <sup>c</sup>	6.92 (1.78)	7.00 (1.41)		
CBT achterwaarts	3.71 (1.71)	3.87 (1.91)	2.76	5.85**
LG1 <sup>a</sup>	3.00 (1.23)	3.11 ( .93)		
LG2 <sup>b</sup>	3.47 (1.66)	3.29 (1.69)		
LG3 <sup>c</sup>	4.58 (1.83)	5.25 (1.09)		
UC	25.03 (9.77)	26.05 (9.31)	1.24	.93
LG1 <sup>a</sup>	26.78 (9.83)	25.89 (6.75)		
LG2 <sup>b</sup>	22.29 (9.87)	24.12 (9.12)		
LG3 <sup>c</sup>	27.58 (9.39)	28.92 (11.07)		

HTKS	42.89 (8.20)	46.42 (5.48)	2.49	3.76**
LG1 <sup>a</sup>	39.22 (9.44)	43.00 (8.23)		
LG2 <sup>b</sup>	42.12 (7.94)	46.29 (4.62)		
LG3 <sup>c</sup>	46.75 (6.41)	49.17 (1.64)		
DCCS	19.00 (2.43)	19.21 (2.55)	1.31	1.96
LG1 <sup>a</sup>	17.89 ( .93)	18.00 ( .71)		
LG2 <sup>b</sup>	19.47 (2.67)	19.18 (2.46)		
LG3 <sup>c</sup>	19.17 (2.73)	20.17 (3.24)		
TB	96.39 (16.29)	101.39 (14.91)	2.74	4.09*
LG1 <sup>a</sup>	91.00 (14.00)	94.44 (12.82)		
LG2 <sup>b</sup>	93.18 (16.89)	98.65 (14.75)		
LG3 <sup>c</sup>	105.00 (14.63)	110.50 (13.04)		

Noot.  $n = 38$ , <sup>a</sup> $n = 9$ , <sup>b</sup> $n = 17$ , <sup>c</sup> $n = 12$ . \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

### Leereffect

**Verschillen in gemiddelden tussen T1 en T2.** Door te berekenen of er een significant verschil in gemiddelden is tussen T1 en T2 is onderzocht of er sprake is van een leereffect. Eerst is gekeken naar de gehele steekproef. Bij de CTB ( $t(37) = -.88$ ,  $p = .38$ ), voorwaartse CBT-subtest ( $t(37) = .00$ ,  $p = 1.00$ ), achterwaartse CBT-subtest ( $t(37) = -.76$ ,  $p = .45$ ), DCCS ( $t(37) = -.60$ ,  $p = .55$ ) en de UC ( $t(37) = -.81$ ,  $p = .43$ ) blijkt geen significant verschil in gemiddelde, en dus geen leereffect, aanwezig tussen T1 en T2. Bij de HTKS ( $t(37) = -3.02$ ,  $p = .01$ ) en de gehele TB ( $t(37) = -3.80$ ,  $p < .01$ ) werd met een significant verschil wel een leereffect gevonden tussen T1 en T2.

**Verschil in gemiddelden tussen T1 en T2 per Leeftijdsgroep.** Per variabele is voor iedere leeftijdsgroep een gepaarde  $t$ -toets voor verschillen in scores tussen T1 en T2 uitgevoerd. Alleen bij de middelste leeftijdsgroep van de HTKS ( $t(16) = -2.77$ ,  $p = .01$ ) en de TB ( $t(16) = -2.96$ ,  $p = .01$ ) is met significante verschillen in gemiddelden tussen T1 en T2 een leereffect gevonden. Bij de jongste leeftijdsgroep en oudste leeftijdsgroep is geen significante verschillen in gemiddelden gevonden. Hier is dan ook geen sprake van een leereffect.

### Effect van Leeftijd

**Verschillen in gemiddelden tussen leeftijdsgroepen.** Met de Eenweg ANOVA is per variabele onderzocht of er sprake is van een effect van leeftijd. Er is gekeken of de verschillen in gemiddelden tussen de leeftijdsgroepen significant waren en de tests beter werden gemaakt naarmate de participanten ouder waren. Bij de CBT T1 ( $F(2,35) = 7.70$ ,  $p < .01$ ), CBT T2 ( $F(2,35) = 9.01$ ,  $p < .01$ ), voorwaartse CBT-subtest T1 ( $F(2,35) = 10.72$ ,  $p < .01$ ), voorwaartse CBT-subtest T2 ( $F(2,35) = 9.17$ ,  $p < .01$ ), achterwaartse CBT-subtest T2 ( $F(2,35) = 5.85$ ,  $p = .01$ ), HTKS T2 ( $F(2,35) = 3.76$ ,  $p = .03$ ) en de TB

T2 ( $F(2,35) = 4.09, p = .03$ ) zijn significante verschillen in gemiddelden gevonden tussen de drie leeftijdsgroepen. De verschillen zijn niet significant gebleken bij de achterwaartse CBT-subtest T1 ( $F(2,35) = 2.76, p = .08$ ), DCCS T1 ( $F(2,35) = 1.31, p = .28$ ), DCCS T2 ( $F(2,35) = 1.96, p = .16$ ), UC T1 ( $F(2,35) = 1.24, p = .30$ ), UC T2 ( $F(2,35) = .93, p = .40$ ), HTKS T1 ( $F(2,35) = 2.49, p = .10$ ) en de TB T1 ( $F(2,35) = 2.74, p = .08$ ). Er lijkt bij deze variabelen geen effect van leeftijd te zijn.

Met Turkey's Post-Hoc toets is gekeken naar de metingen die bij de eenweg ANOVA een significant verschil tussen leeftijdsgroepen toonde. Uitkomsten laten zien dat verschillen vooral zijn gevonden tussen de jongste en oudste leeftijdsgroep. Dit is gebleken bij CBT T1 ( $M_D = 4.39, p < .01$ ), CBT T2 ( $M_D = 4.69, p < .01$ ), voorwaartse CBT-subtest T1 ( $M_D = 2.80, p < .01$ ), voorwaartse CBT-subtest T2 ( $M_D = 2.89, p < .01$ ), achterwaartse CBT-subtest T2 ( $M_D = 2.14, p = .02$ ), HTKS T2 ( $M_D = 6.17, p = .03$ ) en de gehele TB T2 ( $M_D = 16.06, p = .03$ ). Bij de CBT T2 is een significant verschil gevonden tussen de middelste leeftijdsgroep en oudste leeftijdsgroep ( $M_D = 3.19, p = .01$ ). Bij de voorwaartse CBT-subtest T1 ( $M_D = 1.71, p = .01$ ) en de voorwaartse CBT-subtest T2 ( $M_D = 1.65, p = .03$ ) is een significant verschil gevonden tussen de jongste en de middelste leeftijdsgroep. Dit betekent dat bij de CBT als geheel en met name de voorwaartse CBT-subtest kleuters hoger scoren naarmate zij ouder zijn. Bij de achterwaartse CBT-subtest, HTKS en de gehele TB is dit effect van leeftijd ook gevonden, maar alleen bij T2.

### **Test-Hertest Betrouwbaarheid**

**CBT.** Uit Pearson's correlatie analyse van de CBT is gebleken dat er een positieve significante correlatie bestaat tussen T1 en T2 ( $r(36) = .86, p < .001, R^2 = .74$ ). Dit betekent dat de test-hertest betrouwbaarheid volgens de COTAN criteria goed is en 74% van de variabiliteit van de testcores van de CBT T2 kan worden verklaard door de variabiliteit in testcores op T1. Voor de jongste leeftijdsgroep is de betrouwbaarheid onvoldoende ( $r(7) = .42, p = .13, R^2 = .18$ ), maar voor de middelste leeftijdsgroep ( $r(15) = .86, p < .001, R^2 = .74$ ) en de oudste leeftijdsgroep ( $r(10) = .82, p < .01, R^2 = .67$ ) wordt de betrouwbaarheid volgens COTAN als goed beoordeeld.

**Voorwaartse CBT-subtest.** Uit Pearson's correlatie analyse van de voorwaartse CBT-subtest is gebleken dat er een positieve significante correlatie bestaat tussen T1 en T2 ( $r(36) = .78, p < .001, R^2 = .61$ ). Dit betekent dat de test-hertest betrouwbaarheid volgens de COTAN criteria voldoende is en 61% van de variabiliteit van de testcores van de T2 kan worden verklaard door de variabiliteit in testcores op T1. Bij de jongste leeftijdsgroep ( $r(7) = .69, p = .02, R^2 = .48$ ) en oudste leeftijdsgroep ( $r(10) = .54, p = .04, R^2 = .29$ ) is volgens COTAN de betrouwbaarheid onvoldoende. Alleen bij de middelste leeftijdsgroep, ( $r(15) = .80, p < .001, R^2 = .64$ ) is deze goed.

**Achterwaartse CBT-subtest.** Uit Pearson's correlatie analyse van de achterwaartse CBT-subtest is gebleken dat er een positieve significante correlatie bestaat

tussen T1 en T2 ( $r(36) = .75, p < .001, R^2 = .56$ ). Dit betekent dat de test-hertest betrouwbaarheid volgens de COTAN criteria voldoende is en 56% van de variabiliteit van de test scores van de T2 kan worden verklaard door de variabiliteit in test scores op T1. Bij de jongste leeftijdsgroep is de betrouwbaarheid volgens COTAN goed ( $r(7) = .88, p < .01, R^2 = .77$ ). Bij de middelste leeftijdsgroep is deze voldoende ( $r(15) = .77, p < .001, R^2 = .59$ ) en bij oudste leeftijdsgroep onvoldoende ( $r(10) = .62, p = .02, R^2 = .38$ ).

**DCCS.** Uit Pearson's correlatie analyse van de DCCS is gebleken dat er een positieve significante correlatie bestaat tussen T1 en T2 ( $r(36) = .62, p < .001, R^2 = .39$ ). Dit betekent dat de test-hertest betrouwbaarheid volgens de COTAN criteria onvoldoende is en slechts 39% van de variabiliteit van de test scores van de T2 kan worden verklaard door de variabiliteit in test scores op T1. De betrouwbaarheid van de DCCS voor de jongste leeftijdsgroep ( $r(7) = .57, p = .054, R^2 = .32$ ), en oudste leeftijdsgroep ( $r(10) = .47, p = .062, R^2 = .22$ ) is volgens COTAN onvoldoende. Voor de middelste leeftijdsgroep is deze voldoende ( $r(15) = .76, p < .001, R^2 = .58$ ).

**UC.** Uit Pearson's correlatie analyse van de UC is gebleken dat er een positieve significante correlatie bestaat tussen T1 en T2 ( $r(36) = .66, p < .001, R^2 = .44$ ). Dit betekent dat de test-hertest betrouwbaarheid volgens de COTAN criteria onvoldoende is en slechts 44% van de variabiliteit van de test scores van de T2 kan worden verklaard door de variabiliteit in test scores op T1. Voor de jongste leeftijdsgroep is de betrouwbaarheid volgens COTAN onvoldoende ( $r(7) = .28, p = .24, R^2 = .08$ ), voor de middelste leeftijdsgroep is deze net onvoldoende ( $r(15) = .69, p < .01, R^2 = .48$ ), maar voor oudste leeftijdsgroep is de betrouwbaarheid goed ( $r(10) = .82, p < .01, R^2 = .66$ ).

**HTKS.** Uit Pearson's correlatie analyse van de HTKS is gebleken dat er een positieve significante correlatie bestaat tussen T1 en T2 ( $r(36) = .50, p < .01, R^2 = .26$ ). Dit betekent dat de test-hertest betrouwbaarheid volgens de COTAN criteria onvoldoende is en slechts 26% van de variabiliteit van de test scores van de T2 kan worden verklaard door de variabiliteit in test scores op T1. Voor de jongste leeftijdsgroep ( $r(7) = .32, p = .204, R^2 = .10$ ), de middelste leeftijdsgroep ( $r(15) = .62, p = .004, R^2 = .38$ ) en oudste leeftijdsgroep ( $r(10) = .30, p = .174, R^2 = .09$ ) is de betrouwbaarheid onvoldoende.

**TB.** Om van een TB te mogen spreken en deze mee te mogen nemen in de test-hertest betrouwbaarheid analyses moeten de meetinstrumenten onderling significant correleren. Uit Pearson's correlatie analyse komen bij T1 slechts twee significante correlaties voor. Dit is tussen de HTKS en de UC ( $r(36) = .29, p = .04, R^2 = .08$ ) en tussen de HTKS en de DCCS ( $r(36) = .34, p = .02, R^2 = .12$ ). Bij T2 is alleen de correlatie tussen de HTKS en de UC ( $r(36) = .35, p = .02, R^2 = .12$ ) significant. Door de beperkte aanwezigheid van significante correlaties kan niet worden gesproken van één overkoepelende TB.

Uit Pearson's correlatie analyse van de TB is gebleken dat er een positieve

significante correlatie bestaat tussen T1 en T2 ( $r(36) = .87, p < .001, R^2 = .76$ ). Dit betekent dat de test-hertest betrouwbaarheid volgens de COTAN criteria goed is en 76% van de variabiliteit van de test scores van de T2 kan worden verklaard door de variabiliteit in test scores op T1. Voor de jongste leeftijdsgroep ( $r(7) = .79, p = .01, R^2 = .62$ ) is de betrouwbaarheid volgens COTAN voldoende. Voor de middelste leeftijdsgroep ( $r(15) = .89, p < .001, R^2 = .79$ ) en oudste leeftijdsgroep ( $r(10) = .81, p < .01, R^2 = .64$ ) is deze goed. Uit deze analyses blijkt dat de TB een hoge test-hertest betrouwbaarheid heeft voor alle drie de leeftijdsgroepen.

In Tabel 2 staan Pearson's correlatiecoëfficiënten ( $r$ ) en de Fisher  $r$ -to- $z$  transformatie ( $z$ ). De  $z$ -waarden vallen tussen  $\pm 1.96$  en onderschrijven dat er geen significant verschil is in de betrouwbaarheid bij de verschillende leeftijdsgroepen.

Tabel 2

*Resultaten van Pearson's Correlatie Analyse en de Z-waarden tussen T1 en T2 van EF-Meetinstrumenten, de CBT-subtests en de TB per Leeftijdsgroep (LG)*

Meetinstrumenten	Totale steekproef	LG1 <sup>a</sup>		LG2 <sup>b</sup>		LG3 <sup>c</sup>	
		$r$	$z$	$r$	$z$	$r$	$z$
CBT	.86**	.42	.45	.86**	1.30	.82**	1.16
CBT-voorwaarts	.78**	.69*	.85	.80**	1.10	.54*	.60
CBT-achterwaarts	.75**	.88**	1.38	.77**	1.02	.62*	.73
DCCS	.62**	.57	.65	.76**	1.00	.47	.51
UC	.66**	.28	.29	.69**	.85	.81**	1.13
HTKS	.50**	.32	.55	.62**	.73	.30	.31
TB	.87**	.79**	1.07	.89**	1.42	.80**	1.10

*Noot.*  $n = 38, {}^a n = 9, {}^b n = 17, {}^c n = 12$ . \* $p < .05$ . \*\* $p < .01$ .

## Discussie en Conclusie

### Discussie

**CBT.** De test-hertest betrouwbaarheid van de gehele Corsi-blokkentest is goed en bij de voor- en achterwaartse subtests voldoende. Alleen bij de jongste leeftijdsgroep is de betrouwbaarheid onvoldoende. Het feit dat de voor- en achterwaartse subtests een lagere test-hertest betrouwbaarheid hebben komt doordat deze uit minder items bestaan dan de gehele Corsi-blokkentest (Evers et al., 2010). Het leereffect dat Vandierendonk en collega's (2004) verwachtte is niet gevonden. Dit kan verklaard worden doordat de participanten in hun onderzoek rond de 18 jaar oud waren. Het verder ontwikkelde

werkgeheugen van jongvolwassenen stelt hen mogelijk in staat beter te leren van de eerste testafname dan kleuters. Het gevonden effect van leeftijd tussen de jongste en oudste leeftijdsgroep kan verklaard worden door de snelle ontwikkeling van executieve functies in de kleutertijd (Carlson, 2005). Collins en Laski (2015) stellen dat de Corsi-blokkentest een abstracte patroon-taak is en jonge kleuters nog niet in staat zijn nauwkeurig verbanden te leggen om deze te kunnen onthouden en reproduceren. Dit zou een mogelijke verklaring kunnen zijn voor de lage testcores bij de jongste leeftijdsgroep in dit onderzoek en het uitblijven van het verwachte leereffect.

**DCCS.** Hoewel Beck en collega's (2011) voor de hele steekproef en alle leeftijdsgroepen voldoende betrouwbaarheid hebben gevonden, is dat in dit onderzoek alleen voor de middelste leeftijdsgroep het geval geweest. Het leereffect dat Lezak (1995) verwachtte is niet gevonden. Dit kan mogelijk worden verklaard doordat de testleider bij dit meetinstrument al bij de eerste testafname voor iedere testvraag de sorteerregel heeft genoemd. Hierdoor is niet gemeten in hoeverre de participant de sorteerregel heeft kunnen ontdekken en daar bij een tweede testafname profijt van heeft gehad, maar is alleen getest of de participant de sorteerregel heeft kunnen toepassen. Hiermee is de situatie, die Lezak (1995) voorspeld heeft, in dit onderzoek niet aanwezig.

**UC.** De test-hertest betrouwbaarheid van de Updating Cijfertaak is in dit onderzoek voor het eerst getest. De betrouwbaarheid is volgens de COTAN criteria onvoldoende gebleken. Er is geen leereffect of een effect van leeftijd naar voren gekomen. Alleen bij de oudste leeftijdsgroep was de betrouwbaarheid goed. Het lijkt erop dat dit meetinstrument aanspraak doet op executieve functies die bij oudere kleuters stabiel aanwezig zijn dan bij jongere kleuters. Mogelijk wordt dit verklaard door de snelle ontwikkeling van executieve functies in deze leeftijdsfase (Carlson, 2005).

**HTKS.** Uit de analyses blijkt dat de test-hertest betrouwbaarheid tegen de verwachtingen van Ponitz en collega's (2008) in, voor de gehele steekproef en alle leeftijdsgroepen onvoldoende is. Een mogelijke verklaring hiervoor kan zijn dat alle leeftijdsgroepen bij de tweede testafname beter hebben gescoord. De betere score bij de tweede testafname kan mogelijk worden verklaard doordat participanten bij de tweede testafname wisten wat er van hen werd verwacht bij het uitvoeren van de handelingen van de test. Alleen bij de middelste leeftijdsgroep was sprake van een leereffect. Door de kleine steekproef is moeilijk te bepalen waarom het leereffect alleen bij deze leeftijdsgroep is vastgesteld. Het effect van leeftijd tussen de jongste en oudste leeftijdsgroep bij de tweede testafname zou verklaard kunnen worden doordat inhibitie bij kleuters zich ontwikkelt naarmate zij ouder worden (Carlson, 2005).

**TB.** De verwachting van Willoughby en Blair (2011) dat de betrouwbaarheid van de testbatterij hoger is dan van een individueel meetinstrument is grotendeels overeen gekomen met de resultaten van dit onderzoek bij de gehele steekproef. Per leeftijdsgroep

hebben enkele meetinstrumenten wel een hogere betrouwbaarheid dan de testbatterij. De scores bleken op te lopen naarmate kleuters ouder worden, wat inhoudt dat de gemeten executieve functies zich in deze periode ontwikkelen (Carlson, 2005). De lagere betrouwbaarheid bij de jongste leeftijdsgroep kan verklaard worden doordat alle meetinstrumenten van de testbatterij bij deze groep onvoldoende betrouwbaar zijn.

### **Kritische Kanttekeningen en Aanbevelingen voor Vervolgonderzoek**

Een sterke kant van dit onderzoek is dat de testbatterij is afgenomen aan de hand van gestandaardiseerde handleidingen en dat beide testafnames door dezelfde testleider onder dezelfde omstandigheden zijn uitgevoerd. Ook kent dit onderzoek geen drop-outs. Deze factoren verhogen de betrouwbaarheid van de resultaten. Met een test interval van één week is getracht om de test-hertest betrouwbaarheid zuiver te meten en niet ook de snelle ontwikkeling van executieve functies (Anderson, 2002). Dit heeft een positief effect op de inhoudsvaliditeit gehad.

Dit onderzoek kent ook een aantal limitaties, waardoor de testresultaten moeilijk generaliseerbaar zijn en met voorzichtigheid geïnterpreteerd dienen te worden. Ten eerste bleek de handleidingen van de Updating Cijfertaak enkele onduidelijkheden te bevatten waardoor niet alle testleiders hetzelfde afkappunt hebben gehanteerd. Dit heeft ervoor gezorgd dat niet alle testresultaten van deze taak op dezelfde manier zijn gescoord en de interbeoordelaarsbetrouwbaarheid negatief is beïnvloed. Tevens is in dit onderzoek, ondanks de beperkte correlatie van de meetinstrumenten, de gehele testbatterij meegenomen in de analyses. Een aanbeveling is om aanvullend onderzoek te doen naar een combinatie van meetinstrumenten voor executieve functies die onderling sterker correleren. Ook kan worden afgevraagd of de voorwaartse Corsi-blokken subtest wel thuishoort in een testbatterij naar executieve functies. Diamond (2013) stelt dat deze subtest alleen het korte termijn geheugen meet en geen beroep doet op de informatiebewerking van het werkgeheugen. Vervolgonderzoek naar de inhoudsvaliditeit van deze subtest is gewenst. Een aanbeveling is om tot die tijd de voorwaartse Corsi-blokken subtest mogelijk alleen te gebruiken als oefentaak. Daarnaast is in dit onderzoek op basis van eerdere literatuur alleen gebruik gemaakt van een test interval van één week. In toekomstig onderzoek zou verder gekeken kunnen worden naar het effect dat langere en kortere test intervallen hebben op het meten van leereffect en de ontwikkeling van EF. Tot slot is bij de werving van participanten sprake geweest van een selecte steekproefmethode en zijn alleen participanten geworven wiens school en ouders geïnteresseerd waren in deelname aan dit onderzoek. Ook waren participanten slechts afkomstig uit vijf Nederlandse provincies en is in dit onderzoek gebruik gemaakt van een kleine steekproef. In 2016 woonden er ongeveer 554.000 kleuters van 4 tot 7 jaar oud in Nederland (CBS, 2016). Op basis van dit aantal geeft volgens de berekening beschreven door Field (2013) een steekproefgrootte van minimaal 385 participanten een

betrouwbare afspiegeling van de populatie. Tevens heeft de verdeling van de 38 participanten in drie leeftijdsgroepen tot nog kleinere steekproeven geleid. Deze factoren hebben een negatief effect op de externe validiteit en de statistische power van de resultaten van deze groepen (Gravetter & Wallnau, 2013). Omdat de steekproef van dit onderzoek niet groot genoeg is moeten de resultaten met voorzichtigheid worden geïnterpreteerd. Voor vervolgonderzoek wordt volgens de centrale limietstelling (Field, 2013) aangeraden om een aselechte steekproef met minimaal 30 participanten per leeftijdsgroep te hanteren.

### **Conclusie**

Tot nu toe is in Nederland weinig bekend over de betrouwbaarheid van meetinstrumenten voor executieve functies bij kleuters. Om problemen met executieve functies vast te kunnen stellen en te behandelen zijn betrouwbare meetinstrumenten essentieel. Het primaire doel van dit onderzoek is dan ook het meten van de test-hertest betrouwbaarheid van een testbatterij naar executieve functies voor kleuters tussen de 48 en 84 maanden oud geweest. Hoewel de meeste meetinstrumenten uit dit onderzoek in de praktijk al worden gebruikt, is alleen bij de Corsi-blokkentest en de gehele testbatterij voldoende test-hertest betrouwbaarheid gevonden. Opvallend is dat bij de jongste leeftijdsgroep, ondanks de afwezigheid van een significant leereffect, de test-hertest betrouwbaarheid bij alle meetinstrumenten onvoldoende is. In dit onderzoek is gebleken dat met de gebruikte meetinstrumenten bij kleuters jonger dan 60 maanden moeilijk een betrouwbare beoordeling van executieve functies kan worden verkregen. Het is niet duidelijk of 60 maanden de ondergrens is voor het meten hiervan, of dat andere meetinstrumenten mogelijk wel een betrouwbare meting kunnen doen bij deze leeftijdsgroep. Wel is gebleken dat er nauwelijks sprake was van een leereffect tussen de eerste en tweede testafname. De ontwikkeling van de executieve functies bij kleuters is met name bij de tweede testafname naar voren gekomen door een effect van leeftijd tussen jongste en oudste kleuters. Tot slot lijken de vier gebruikte meetinstrumenten in dit onderzoek, door een beperkte onderlinge correlatie, geen geschikte testbatterij te vormen.



## Referenties

- Alloway, T. P. (2006). How does working memory work in the classroom? *Educational Research and Reviews, 1*, 134-139.  
<http://www.academicjournals.org/journal/ERR/article-abstract/475D3083412>
- Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology, 106*, 20-29. doi:10.1016/j.jecp.2009.11.003
- Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? *Child Development, 77*, 1698-1716. doi:10.1111/j.1467-8624.2006.00968.x
- Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology, 8*, 71-82. doi:10.1076/chin.8.2.71.8724
- Baddeley, A. (2001). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*, 417-423. doi:10.1016/S1364-6613(00)01538-2
- Beck, D. M., Schaefer, C., Pang, K., & Carlson, S. M. (2011). Executive function in preschool children: Test-retest reliability. *Journal of Cognition and Development, 12*, 169-193. doi:10.1080/15248372.2011.563485
- Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development, 78*, 647-663. doi:10.1111/j.1467-8624.2007.01019.x
- Carlson, S. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology, 28*, 595-616. doi:10.1207/s15326942dn2802\_3
- Casey, B. J., Tottenham, N., & Fossella, J. (2002). Clinical, imaging, lesion and genetic approaches toward a model of cognitive control. *Developmental psychobiology, 40*, 237-254. doi:10.1002/dev.10030
- CBS. (2016). *Bevolking; geslacht, leeftijd en burgerlijke staat, 1 januari*. Verkregen via: <http://statline.cbs.nl>
- Collins, M. A., & Laski, E. V. (2015). Preschoolers' strategies for solving visual pattern tasks. *Early Childhood Research Quarterly, 32*, 204-214. doi:org/10.1016/j.ecresq.2015.04.004
- Corsi, P. M. (1972). *Human memory and the medial temporal region of the brain*. (master thesis). Verkregen via [mccgill.ca](http://mccgill.ca)
- Diamond, A. (2013). Executive Functions. *Annual Review of Psychology, 64*, 135-168. doi:10.1146/annurev-psych-113011-143750
- Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *Cotan beoordelingssysteem voor de kwaliteit van tests*. Amsterdam: Nederlands Instituut van Psychologen

- Field, A. (2013). *Discovering Statistics using IBM SPSS Statistics*. Londen: SAGE.
- Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers; A review using an integrative framework. *Psychological Bulletin*, *134*, 31-60. doi:10.1037/0033-2909.134.1.31
- Gathercole S. E., & Alloway, T. P. (2007). *Understanding working memory, A classroom guide*, United Kingdom: Harcourt Assessment.
- Gravetter, F. J. & Wallnau, L. B. (2013). *Statistics for the Behavioral Sciences*. Negende editie, London: Thomson Wadsworth.
- Gray, S. (2003). Diagnostic accuracy and test-retest reliability of nonword repetition and digit span tasks administered to preschool children with specific language impairment. *Journal of Communication Disorders*, *36*, 129-151. doi:10.1016/S0021(03)00003-0
- Jonkman, K., Kooij, A., Wechsler, D., & Naglieri, J. A. (2008). *Researchbeschrijving, Wechsler nonverbal scale of ability*. Amsterdam: Boom test uitgevers.
- Kessels, R. P. C., Berg, E. van den, Ruis, C., & Brands, A. M. A. (2008). The backward span of the corsi block-tapping task and it's association with the WAIS-III digit span. *Sage Journals*, *15*, 426-434. doi:10.1177/1073191108315611
- Lezak, M. D. (1995). *Neuropsychological assessment (3rd ed.)*. New York: Oxford University Press
- Miyake, A., Friedman, N. P., Emmerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D., (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49-100. doi:10.1006/cogp.1999.0734
- Müller, U., Kerns, K. A., Konkin, K. (2012). Test-retest reliability and practice effects of executive function tasks in preschool children. *The Clinical Neuropsychologist*, *26*, 271-287. doi:10.1080/13854046.2011.645558
- Neuman, W. L. (2009). *Understanding research*. Londen, United Kingdom: Pearson Education, Inc.
- Ozonoff, S., (1995). Reliability and validity of the Wisconsin Card Sorting Test in studies of autism. *Neuropsychology*, *9*, 491-500. doi:10.1037/0894-4105.9.4.491
- Ponitz, C. E., McClelland, M. M., Jewkes, A. M., McDonald-Connor, C., Farris, C. L., & Morrison, F. J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Quarterly*, *23*, 141-158. doi:10.1016/j.ecresq.2007.01.004
- Smidts, D. P. (2003). Executieve functies van geboorte tot adolescentie: Een literatuuroverzicht. *Neuropraxis*, *7*, 113-119. doi:10.1007/BF03099824
- Tak, J. A., Bosch, J. D., Begeer, S., & Albrecht, G. (2014). *Handboek psychodiagnostiek voor hulpverlening aan kinderen en adolescenten*. Utrecht: De Tijdstroom.

- Tijmstra, J. & Boeije, H. (2011). *Wetenschapsfiliosifie in de contexte van de sociale wetenschap*. Den Haag: Boom Lemma Uitgevers.
- Vandierendonck, A., Kemps, E., Fastame, M. C., & Szmalec, A. (2004). Working memory components of the corsi blocks task. *British Journal of Psychology*, *95*, 57–79. doi:10.1348/000712604322779460
- Willoughby, M., & Blair, C. (2011). Test-retest reliability of a new executive function battery for use in early childhood. *Child Neuropsychology*, *17*, 564–579. doi:10.1080/09297049.2011.554390
- Zelazo, P. D. (2006). The dimensional change card sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, *1*, 297-301. doi:10.1038/nprot.2006.46