Gericht werken aan opbrengsten in taal- en leesonderwijs

Een systematische review naar toetsvormen

Femke Scheltinga

Jos Keuning

Hans Kuhlemeier

Expertisecentrum Nederlands / Cito
Nijmegen, 2014

Dit onderzoek is gefinancierd door de Programmaraad Praktijkgericht Onderzoek (PPO) van het Nationaal Regieorgaan Onderwijsonderzoek (NRO), dossiernummer 405-14-533

Inhoud

Samenvatting

Door anders te toetsen in het taal- en leesonderwijs kunnen leraren hun leerlingen gerichter vooruit helpen. Dat is de voorzichtige conclusie van een literatuurstudie die is uitgevoerd door het Expertisecentrum Nederlands en Cito. In een systematische literatuurstudie is gezocht naar toetsvormen die leraren in staat stellen het leerproces te verbeteren, het leerpotentieel vast te stellen of een gedetailleerde diagnose te geven van de sterke en zwakke punten van leerlingen. Het ging daarbij om het toetsen van leesvaardigheid, woordenschat, schrijfvaardigheid en mondelinge taalvaardigheid. Om bruikbaar te zijn voor leraren moesten de toetsvormen bewezen effectief zijn. Dit betekent dat het onderzoek waarin de effectiviteit wordt onderzocht, moest voldoen aan de gebruikelijke wetenschappelijke kwaliteitsstandaarden.

Waarom zijn andere toetsvormen wenselijk?
Om goed taal- en leesonderwijs te geven, moet de leraar weten hoe leerlingen leren. Nu bieden toetsen en leerlingvolgsystemen hierover vaak nog weinig informatie, doordat de taalvaardigheid van leerlingen op een statische manier wordt getoetst. Op een vast moment kijken leraren bijvoorbeeld of hun leerlingen teksten goed hebben begrepen en of hun woordenschat op het gewenste niveau is. De leerling moet een statische toets volledig zelfstandig maken. Als de leerling tijdens de afname een probleem tegenkomt, mag de docent niet helpen, bijvoorbeeld door hints, uitleg, feedback of instructie te geven. Daardoor krijgt de docent weinig informatie over hoe leerlingen leren en wat nodig is om het leren te verbeteren. Het is daarom wenselijk om ook andere toetsen te gebruiken, zoals dynamische en diagnostische toetsen.

Wat is dynamisch toetsen?
Bij een dynamische toets mag de leraar de leerling tijdens de afname helpen. De leraar geeft instructie en feedback. De toets meet niet zozeer wat een leerling tot dan toe geleerd heeft, maar vooral het leerpotentieel. De afname geeft leraren een beeld van wat de leerling maximaal kan bereiken en hoeveel en welke instructie en begeleiding daarvoor nodig is. Bij intelligentiemetingen worden dynamisch toetsen al lang met veel succes toegepast. Uit onze literatuurstudie blijkt dat dynamische toetsen nog nauwelijks in het taalonderwijs worden gebruikt. Wij vonden slechts negen onderzoeken met voorbeelden van dynamische toetsing die mogelijk relevant zijn voor leraren. Zo vonden we succesvolle voorbeelden van toetsing van begrijpend lezen waarbij leerlingen tijdens de toetsafname feedback krijgen. De feedback bestaat onder meer uit hints over een geschikte leesstrategie om de tekst te begrijpen. De leerlingen krijgen bijvoorbeeld een hint als het gegeven antwoord op een tekstbegripsvraag onjuist is. Om de leerling tot het juiste antwoord te laten komen, volgen er zo nodig meer hints die steeds explicieter worden. Zo kan een eerste hint zijn dat de leerling moet proberen de hoofdgedachte van een passage te vinden. Als de leerling het antwoord dan niet weet, kan worden verwezen naar een specifieke zin, een woord, of kan worden gezegd wat het juiste antwoord is en hoe dit antwoord uit de tekst is af te leiden. In een dergelijke toets geeft het aantal hints dat een leerling nodig heeft, niet alleen zicht op het vaardigheidsniveau, maar ook op de manier waarop de docent de taalvaardigheid verder kan ontwikkelen door het geven van instructie en ondersteuning.

Dynamische toetsen worden ook gebruikt om de taalvaardigheid van specifieke groepen leerlingen vast te stellen. We vonden studies die suggereren dat een dynamisch toets kan helpen een betrouwbaar onderscheid te maken tussen leerlingen met en zonder een taalstoornis. Leerlingen uit een anderstalige en/of zwak sociaal-economische omgeving presteren op een statische toets dikwijls even laag als een leerling met een taalstoornis, doordat zij onbekend zijn met de taak. Zij presteren zwak, niet door een taalstoornis, maar door een achterstand in ervaring. De leerlingen zijn bijvoorbeeld gewend de woorden naar hun functie te omschrijven (‘Je kunt het eten’) in plaats van te benoemen (‘Een appel’). Hoewel deze leerlingen de woorden wel kennen, behalen ze dus toch een lage score op een statische woordenschattoets. Van de studies naar dynamisch toetsen voldeden er slechts twee aan gangbare wetenschappelijke kwaliteitsstandaarden. De studies bij specifieke groepen, om onderscheid te maken tussen taalstoornis en achterstand, voldeden hier niet aan, zodat we de bewijskracht als zwak moeten beschouwen. Hetzelfde geldt voor de studies die claimen dat dynamische toetsen kunnen helpen om beter onderscheid te maken tussen leerlingen die wél en leerlingen die geen speciale onderwijsvoorzieningen nodig hebben.

Wat is diagnostisch toetsen?
De gebruikelijke toetsen bieden leraren vaak weinig informatie over de sterke en zwakke punten van leerlingen. Bovendien is meestal niet duidelijk hoe het komt dat de leerling op één of meer onderdelen zwak presteert en wat er precies aan gedaan kan worden. Diagnostische toetsen zijn gericht op het verstrekken van gedetailleerde informatie waarmee het leerproces verder ingericht kan worden. Net als dynamische toetsen worden ook diagnostische toetsen nog weinig in het taalonderwijs gebruikt. We vonden slechts vijf studies die mogelijk relevant zijn voor leraren. Daarvan voldeed er één aan de gangbare wetenschappelijke kwaliteitscriteria. Er is onder meer onderzoek gedaan naar de effectiviteit van geautomatiseerde feedback op het gebied van

schrijfvaardigheid. De leerlingen maken bij het schrijven gebruik van de computer en de feedback wordt door het computerprogramma gegenereerd. Daarbij bleek dat leerlingen die gedetailleerde feedback kregen veel meer tijd besteedden aan het verbeteren van hun schrijfproducten dan degenen die geen feedback kregen. Daarnaast was de inhoudelijke, organisatorische en stilistische kwaliteit van de geschreven teksten met feedback een stuk beter dan zonder. Er zijn dus aanwijzingen dat het geven van geautomatiseerde feedback tijdens de toetsafname een positieve uitwerking kan hebben op de schrijfprestaties van de leerlingen.

Wat betekent dit voor leraren?
Onze literatuurstudie heeft enkele alternatieve toetsvormen aan het licht gebracht die leraren kunnen ondersteunen bij het verder ontwikkelen van de taalvaardigheid van leerlingen. De studies naar dynamisch toetsen doen vermoeden dat het mogelijk is om de taalvaardigheid van leerlingen door middel van dynamisch toetsen te verhogen. Gemeenschappelijk in de diverse toepassingen is de vergaande integratie van onderwijzen, leren en toetsen en het verstrekken van formatieve feedback tijdens de toetsafname (door de leraar). Om bruikbaar te zijn voor het onderwijs, moeten deze toepassingen eerst voor de Nederlandse situatie geschikt gemaakt worden. Daartoe is onderzoek noodzakelijk waarin veelbelovende toetsvormen in samenwerking met leraren op uitvoerbaarheid en effectiviteit getest worden.

De studies naar diagnostisch toetsen laten zien dat leerlingen van gecomputeriseerde feedback tijdens de afname veel kunnen leren. Voor het Nederlandse taalgebied zijn er nog vrijwel geen computergebaseerde systemen beschikbaar die leerlingen gedetailleerde feedback geven over hoe zij hun lees- of schrijfvaardigheid kunnen verbeteren. Het verdient aanbeveling na te gaan in hoeverre dergelijke systemen ook voor het onderwijs in Nederland ontwikkeld kunnen worden. Voor andere toetsdoelen dan het ondersteunen van het leren is het ‘effectiviteitsbewijs’ op zijn zachts gezegd veel minder krachtig. Dat geldt onder meer voor het maken van een betrouwbaar onderscheid tussen leerlingen met en zonder taalstoornis en tussen leerlingen die wél en geen speciale onderwijsvoorzieningen nodig hebben. Hier spreken de resultaten elkaar nog te vaak tegen en is de meerwaarde van de nieuwe toetsvormen vooralsnog onduidelijk. Voordat leraren hier iets mee kunnen, lijkt meer en vooral beter onderzoek geboden.

1 Inleiding

1.1 Achtergrond en kader

Verbetering van het taal- en leesonderwijs is een belangrijke doelstelling van het Ministerie van OCW. Om deze verbetering te kunnen realiseren worden scholen aangespoord om opbrengstgericht te werken (OCW, 2007, 2011). In 2015 zou 50 tot 60 procent van de scholen voor het primair en voortgezet onderwijs opbrengstgericht moeten werken. In de daaropvolgende jaren moet dat percentage 75 tot 90 zijn. Opbrengstgericht werken houdt in dat scholen in hun onderwijs van leerstandaarden en lesdoelen uitgaan, op basis van toetsen systematisch informatie over het leerproces verzamelen, deze informatie voor nadere analyse en interpretatie vastleggen, en op basis hiervan beslissingen nemen over de invulling van het onderwijs. In de praktijk blijken leraren het lastig te vinden om opbrengstgericht te werken (Inspectie van het Onderwijs, 2013). Zij hebben kennelijk onvoldoende mogelijkheden om toetsresultaten te relateren aan de leerlijnen voor het taal- en leesonderwijs. Daardoor is het voor hen moeilijk om vast te stellen welke leerstof en didactische benadering het meest geschikt is om leerlingen te begeleiden. Om het onderwijs te kunnen vormgeven willen leraren wel graag informatie over het leerproces en het leerpotentieel van leerlingen hebben. Zij vinden dergelijke informatie belangrijker dan toetsresultaten die laten zien hoe een leerling presteert in vergelijking met leeftijdgenoten (cf. Bosma, Hessels & Resing, 2012). In wetenschappelijke literatuur wordt erop gewezen dat alternatieve toetsvormen een bijdrage kunnen leveren aan het opbrengstgericht werken. Zulke toetsvormen worden echter nog weinig toegepast in de onderwijspraktijk (Elliot, 2003; Grigorenko, 2009). In deze literatuurstudie is gezocht naar alternatieve vormen van toetsen die houvast kunnen geven bij het vormgeven en evalueren van het taal- en leesonderwijs.

1.2 Leerlijnen voor het taal- en leesonderwijs

De doelen waaraan in het onderwijs opbrengstgericht gewerkt moet worden bestaan uit tussendoelen, kerndoelen en referentieniveaus. De referentieniveaus vormen het *Referentiekader Nederlandse taal* dat in schooljaar 2010/2011 is ingevoerd (Expertgroep Doorlopende Leerlijnen, 2007). Het referentiekader beschrijft wat leerlingen aan het eind van het primair, voortgezet en middelbaar beroepsonderwijs moeten kennen en kunnen op het gebied van taalvaardigheid. Daarbij worden 4 domeinen onderscheiden: 1) mondelinge taalvaardigheid, 2) leesvaardigheid, 3) schrijfvaardigheid en 4) begrippenlijst en taalverzorging. Voor elk domein worden vier verschillende fundamentele niveaus (1F t/m 4F) van opklimmende moeilijkheidsgraad beschreven. Deze niveaus beschrijven daarmee de ontwikkeling van leerlingen door de jaren heen. Elk niveau beschrijft wat leerlingen op een bepaald moment in de schoolloopbaan moeten kunnen en kennen afhankelijk van het onderwijstype dat gevolgd wordt. De niveaus sluiten op elkaar aan. Bij beheersing van een niveau, wordt het volgende niveau het streefniveau. Bij elk niveau wordt een beschrijving van kennis en vaardigheden gegeven die leerlingen moeten beheersen. Er wordt een algemene beschrijving gegeven, een beschrijving van de taak en kenmerken van de taakuitvoering.

In diverse publicaties, zoals in concretiseringen van de referentieniveaus (Meestringa, Ravesloot & de Vries, 2010), wordt beschreven hoe toegewerkt kan worden naar beheersing van deze niveaus. Voor het basisonderwijs en de onderbouw van het voortgezet onderwijs zijn bovendien tussendoelen geformuleerd die een leerlijn vormen om stapsgewijs tot het gewenste einddoel te komen (zie bijvoorbeeld: Aarnoutse & Verhoeven, 1999; Aarnoutse, Verhoeven, van het Zandt, & Biemond, 2003; Punt & De Krosse, 2012; Verhoeven, Biemond, & Litjens 2007). Hoe deze tussendoelen op het gebied van taal zich tot de referentieniveaus verhouden, wordt inzichtelijk gemaakt op de website www.leerlijnentaal.nl. Om vast te stellen of de doelen waaraan is gewerkt ook behaald zijn, worden toetsen afgenomen. Dat betekent dat de tussendoelen, einddoelen en referentieniveaus verwerkt zijn in de toetsen en examens die tussentijds of aan het einde van een schoolperiode worden afgenomen. Leraren bepalen aan de hand van toetsen of de tussendoelen zijn behaald op weg naar het gewenste referentieniveau of einddoel. Verschillende toetsen vormen zo samen een doorlopende toetslijn. Het gaat om methode-onafhankelijke en methode-afhankelijke toetsen aan het einde van een onderwijsperiode waarmee doelen met betrekking tot kennis en vaardigheden, zoals in de referentieniveaus beschreven, kunnen worden getoetst.

Het werken met doelen en toetsen zorgt voor een opbrengstgerichte aanpak. Dit betekent dat scholen uitgaan van leerdoelen en leerstandaarden en dat zij het leerproces systematisch en cyclisch volgen aan de hand van toetsen.

De toetsinformatie wordt gebruikt om het vervolg in onderwijsaanbod vast te stellen. Het onderwijsaanbod in de verschillende taaldomeinen moet op de tussen- en einddoelen worden afgestemd en verschilt vanzelfsprekend in de verschillende fasen en niveaus van het onderwijs. Zo zullen bepaalde (deel)vaardigheden als technisch lezen vooral in het basisonderwijs aandacht krijgen. Dit is terug te zien in de beschrijving van het referentiekader waarin de techniek van het lezen alleen deel uitmaakt van het laagste referentieniveau, maar de ontwikkeling van leesbegrip tot en met 4F beschreven wordt. Ter illustratie hebben we in de bijlage tabellen opgenomen waarin we een aantal kenmerken laten zien van de taakuitvoering bij de referentieniveaus en welke tussendoelen daaraan vooraf gaan binnen de domeinen leesvaardigheid (technisch en begrijpend) en schrijfvaardigheid. We hebben de doelen met betrekking tot woordenschat afzonderlijk in een tabel opgenomen, hoewel ze in het referentiekader deel uitmaken van de andere domeinen.

Hoewel de doelen voor de afzonderlijke domeinen worden beschreven, staan de vaardigheden en doelen niet los van elkaar. Het is van belang te beseffen dat de domeinen met elkaar samenhangen. Bepaalde vaardigheden zijn voorspellers voor latere vaardigheden. Letterkennis is bijvoorbeeld nodig om woorden te kunnen decoderen (Ehri, 2005) en decodeervaardigheid is een belangrijke voorwaarde om tot begrijpend lezen te komen (Verhoeven & Van Leeuwe, 2008). De omvang van de woordenschat hangt ook samen met begrijpend lezen (Muter, Hulme, Snowling & Stevenson, 2004) en in iets mindere mate met technisch lezen (zie o.a. Ouellette, 2006). Een lezer moet de betekenis van een voldoende aantal woorden kennen om een tekst te begrijpen. Andersom vergroot een lezer zijn (lees)woordenschat door te lezen. Een lezer kan door het gebruik van leesstrategieën de betekenis van onbekende woorden uit de tekst afleiden.

Bij het interpreteren van toetsresultaten moet met deze samenhang rekening worden gehouden. Om bijvoorbeeld te bepalen wat de aard van de problemen met leesbegrip is, moeten ook de leesvaardigheid en woordenschat getoetst worden. In de huidige onderwijspraktijk worden vaardigheden meestal per afzonderlijk leerdomein getoetst, op één moment als het product van de instructie en oefening in de voorafgaande periode. Zo wordt vastgesteld met welke onderdelen een leerling problemen ervaart, maar de toetsinformatie geeft weinig richtlijnen en handvatten voor het vormgeven van instructie en interventie (Pameijer, 2006). Andere vormen van toetsen, waarbij rekening wordt gehouden met de (deel)vaardigheden van het leerproces, leerstrategieën en de leerbaarheid van de leerling bieden mogelijk meer aanknopingspunten. Voor het toetsen van specifieke doelen (tussendoelen, kerndoelen en referentieniveaus) zijn bepaalde toetsvormen mogelijk beter geschikt dan andere. Zo levert een toetsvorm, waarbij de leerling gevraagd wordt naar de (denk)stappen die hij zet of waarbij hij feedback krijgt op het antwoord, meer informatie over de leesstrategieën die hij al wel of nog niet hanteert. Vervolgens kunnen de lesactiviteiten en instructie worden afgestemd op het (taal)niveau van de leerling (Whitehead, 2007; Pellegrino 2008).

1.3 Functies en vormen van toetsen

Bij het maken van een keuze voor een toets is het van belang na te gaan met welk doel de toets wordt afgenomen. Een toets kan om verschillende redenen aan leerlingen worden voorgelegd (e.g., Johnson & Wentling 1996; Craigh 2001; Greenleaf et al. 1997; Laurier, 2004; Sanders, 2010). We kunnen de volgende zeven functies onderscheiden:

1. ***Intake en selectie.*** Op basis van het toetsresultaat wordt bepaald of leerlingen in een bepaalde opleiding kunnen instromen;
2. ***Plaatsing.*** Leerlingen worden op basis van de toets ingedeeld in niveaugroepen;
3. ***Voortgangscontrole.*** De toets wordt afgenomen om de ontwikkeling van leerlingen over de tijd te monitoren;
4. ***Bepaling leerpotentieel.*** De toets wordt afgenomen om vast te stellen wat de leerling in een vervolgtraject kan bereiken als de leerling passende instructie, hulp en begeleiding krijgt;
5. ***Diagnostiek.*** Op basis van de toetsafname worden de sterktes en zwaktes van leerlingen in kaart gebracht en geanalyseerd;
6. ***Verbetering leerproces.*** Op basis van het toetsresultaat wordt het leerproces geanalyseerd en bijgestuurd;
7. ***Certificering.*** Afhankelijk van de resultaten op de toets wordt beslist of leerlingen een onderwijsprogramma al of niet met succes afgesloten hebben.

Uit de opsomming en omschrijving van toetsfuncties blijkt dat toetsen verschillende typen feedback over het prestatieniveau van de leerling geven. Als een toets bijvoorbeeld wordt gebruikt om een leerling te plaatsen, kan na de toetsafname alleen worden vastgesteld of een leerling een bepaald tussendoel, kerndoel of referentieniveau bereikt heeft. De toets vertelt niet hoe een leerkracht of leerling om moet gaan met het gegeven dat bepaalde doelen niet bereikt zijn. In andere gevallen, zoals toetsen om het leerpotentieel vast te stellen of het leerproces te

verbeteren, geeft een toetsafname die informatie wel. Om dit onderscheid te maken worden toetsen dikwijls ingedeeld naar *summatief* en *formatief* gebruik. Er is sprake van *summatief* gebruik als met de toets wordt vastgesteld of een leerling een bepaald doel bereikt heeft. Er wordt van *formatief* gebruik gesproken als de toets wordt ingezet om het leren te bevorderen en om onderwijsbeslissingen te nemen.

Als het in discussies gaat over de effectiviteit van formatieve toetsing wordt er voornamelijk gewezen op de positieve effecten. Zo wordt er vanuit gegaan dat feedback de leerprestaties van leerlingen positief beïnvloedt (e.g., Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kluger & DeNisi, 1996; Shute, 2008). Bij summatieve toetsing gaat de discussie juist vaak over allerlei negatieve bijeffecten, zoals onzuiver handelen van scholen en verarming van het curriculum, die de beoordeling van leerlingen, leraren of scholen met zich mee kan brengen (e.g., Darling-Hammond, 2004; Ingram, Louis & Schroeder, 2004; Kornhaber, 2004; McGill-Frantzen & Allington, 2006; Mehrens, 2002; Shepard, 2003). Hoewel discussies soms nogal eenzijdig gevoerd worden, wordt er ook vaak op gewezen dat de twee vormen van toetsgebruik prima naast elkaar kunnen bestaan. Een toets kan zowel summatief als formatief worden ingezet. Met summatieve toetsing is het mogelijk om leerprestaties van leerlingen over de tijd norm- en criteriumgeoriënteerd te volgen. Als bij dezelfde toetsing gerichte feedback aan leerlingen of leraren wordt gegeven, kan deze ook voor formatieve doeleinden ingezet worden. Een eindcijfer alleen volstaat dan niet. Om van formatieve toetsing te kunnen spreken zou een leerling, al dan niet met tussenkomst van een leerkracht, minimaal moeten kunnen afleiden welke tussendoelen, kerndoelen of referentieniveaus bereikt zijn en welke nog niet. Tevens zou de leerling moeten kunnen afleiden wat er gedaan moet worden om de niet behaalde doelen te realiseren (cf. Hattie & Timperley, 2007).

In de literatuur vinden we ook andere termen terug. Er wordt, in plaats van over summatief en formatief gebruik, ook wel gesproken over *assessment of learning* en *assessment for learning*. De term *assessment of learning* is een synoniem voor summatieve toetsing. Onder *assessment for learning* vallen formatieve toetsingen (cf. Arter 2003, 264; Leahy et al. 2005; Harlen 2005). Soms worden *data-driven decision making* en *diagnostic testing* ook hiertoe gerekend (Van der Kleij, 2013). Deze twee toetsingscategorieën gaan er, net zoals de categorie *assessment for learning*, vanuit dat toetsgegevens ingezet worden om het onderwijs aan te passen en af te stemmen op de leerbehoeften van individuele leerlingen (William, 2011). Het gaat dus om formatief toetsgebruik. Aan de verschillende categorieën ligt echter een fundamenteel andere visie op leren ten grondslag. *Data-driven decision making* kan gedefinieerd worden als "… het systematisch en doelgericht werken aan het maximaliseren van leerlingprestaties" (Inspectie van het Onderwijs, 2010). Schoolteams die op deze wijze werken gaan uit van leerstandaarden en lesdoelen, verzamelen informatie over het leerproces, leggen deze vast voor nadere analyse en interpretatie, en nemen op basis hiervan beslissingen over het vervolg van het onderwijs (Parrett & Budge, 2009; Wayman et al., 2013). Bij *diagnostic testing* staat niet het meten van leerprestaties centraal, maar juist het verklaren ervan. Door informatie te verzamelen over de strategieën die leerlingen volgen bij het uitvoeren van een taak wordt geprobeerd om het leerproces en de uitkomsten te duiden (Crisp, 2012; Keeley & Tobey, 2011). Waar *data-driven decision making* zich dus richt op *wat* er geleerd wordt, richt *diagnostic testing* zich op de vraag *hoe* er geleerd wordt. De verschillende toetsingen die tot *assessment for learning* worden gerekend, hebben gemeenschappelijk dat ze zich richten op het bewaken van de kwaliteit van het leerproces (Stobart, 2008). Er wordt gezocht naar informatie, en een interpretatie ervan, die door leerlingen en hun leraren gebruikt kan worden om te beslissen waar een leerling staat in het leerproces, waar hij heen moet en hoe hij daar het beste naar toe kan werken (Broadfoot et al., 2002; Birenbaum, Kimron & Shilton, 2011).

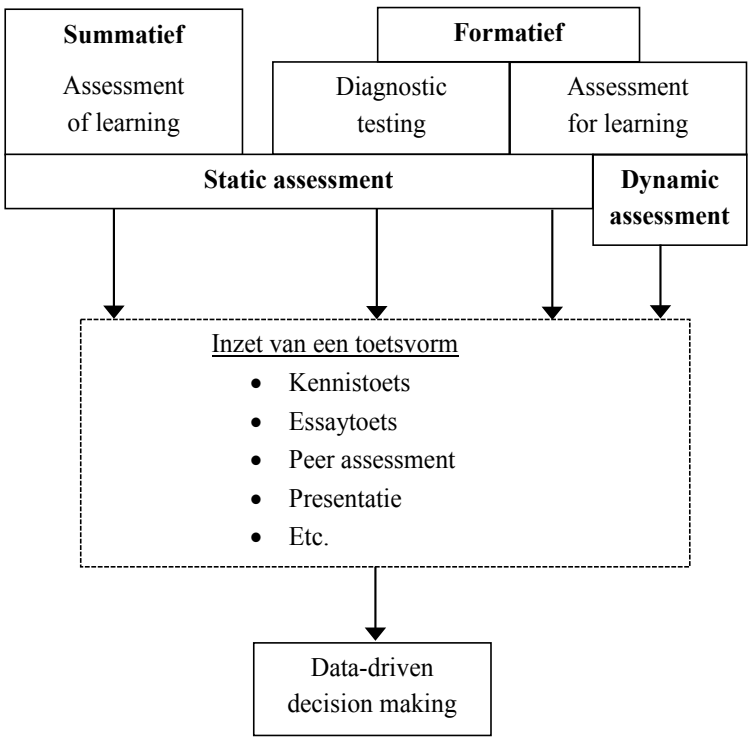
Het belangrijkste doel van *assessment for learning* is het versterken en verbeteren van de instructie en het leren (Cowie & Bell, 1999). De leerkracht heeft een belangrijke rol. Waar de leerkracht normaliter neutraal is en alleen gestandaardiseerde instructie geeft aan leerlingen, geeft de leerkracht bij een *assessment for learning* specifieke feedback naar aanleiding van (foutief) gedrag. Het doel daarvan is om het gedrag en de prestaties van leerlingen te veranderen en/of te verbeteren. Een *assessment for learning* kan op verschillende manieren vormgegeven worden. Voor de uitwerkingen worden verschillende benamingen gebruikt. In de literatuur wordt onder meer gesproken over *learning potential assessment* (e.g., Budoff, Gimon, & Corman, 1976; Budoff, Meskin, & Harrison, 1971), *mediated learning* (e.g., Feuerstein, Rand, & Hoffman, 1979), *testing the limits* (Carlson & Wiedl, 1978, 1979), *mediated assessment* (e.g., Bransford, Delclos, Vye, Burns, & Hasselbring, 1987), en *assisted learning and transfer by graduated prompts* (e.g., Campione, Brown, Ferrara, Jones, & Steinberg, 1985). Een veelgebruikte overkoepelende term is *dynamic assessment* (zie bijvoorbeeld Caffrey, Fuchs & Fuchs, 2008). Er wordt uitgegaan van het concept van de zone van naaste ontwikkeling. Vygotsky (1978) definieert deze zone als "… the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (p. 86). Het gaat om de kloof tussen de leerprestatie die een leerling zelfstandig en met behulp van anderen kan leveren. *Static assessments* zoals *diagnostic testing* geven informatie over de ondergrens van deze zone; wat kan de leerling zonder hulp van anderen? *Dynamic assessments* exploreren de bovengrens en geven hiermee een indicatie van het leerpotentieel; in welke mate is de leerling in staat om te profiteren van de instructie, hulp en

begeleiding van de leerkracht? Ofwel, wat kan de leerling in het vervolgtraject bereiken als er passende instructie, hulp en begeleiding gegeven wordt?

Anders dan bij een statische toets mag de examinator de leerlingen bij een *dynamic assessment* gerichte instructie, hulp en feedback geven. Er wordt gemeten hoeveel een leerling kan leren in plaats van hoeveel de leerling op dat moment weet of kan. *Dynamic assessment* is een vorm van formatieve assessment die informatie verschaft over het leren van de leerling, zoals over de manier waarop de leerling de taak aanpakt, de fouten die hij of zij daarbij maakt en de vaardigheid om zichzelf te corrigeren. Daarnaast verschaft een *dynamic assessment* inzicht in de veranderbaarheid of leerbaarheid van de leerling ofwel hoe gemakkelijk de leerling leert. Daarbij gaat het om a) de mate waarin de leerling profiteert van de instructie of interventie en b) de hoeveelheid instructie, hulp en begeleiding die de examinator heeft moeten geven om de leerling tot leren te brengen (Lidz & Peña, 1996). *Dynamic assessment* verschilt van statische toetsing doordat het de leerkracht direct feedback geeft om de leerling te ondersteunen.

Zoals uit het voorgaande blijkt, zijn toetsen op verschillende manieren te gebruiken en in te delen. In figuur 1.1 geven we een indeling van deze meest gangbare benamingen die in de literatuur te vinden zijn. Figuur 1.1 gaat uit van een indeling naar (a) *summatief* en *formatief* toetsgebruik, en (b) *static assessment* en *dynamic assessment*. Bij een *static assessment* met een summatief gebruiksdoel kan het bijvoorbeeld gaan om een klassikale afname van een wiskundetoets met meerkeuzeopgaven. De leerlingen maken de toets volledig zelfstandig en de leerkracht geeft uitsluitend procedurele instructie zonder feedback of hints tijdens de toetsafname. Bij een *dynamic assessment* met een formatief gebruiksdoel kan het gaan om het vaststellen en verbeteren van de vaardigheden van individuele leerlingen. De leerkracht probeert (foutief) gedrag bij te sturen door de leerling tijdens het uitvoeren van de toets inhoudelijke instructie en feedback te geven. De toetsinhoud en het onderwijs zijn dan op elkaar afgestemd. Een toetsvorm kan zijn dat leerlingen een verhaal moeten navertellen waarbij de leraar feedback geeft om de lengte en complexiteit van het verhaal te vergroten om de mondelinge taalvaardigheid (navertellen) te verbeteren (e.g. Peña et al., 2006). Veel toetsen zijn ergens tussen deze twee uitersten in te plaatsen: het onderscheid is niet altijd strikt te maken. Als leerlingen bijvoorbeeld elkaars schrijfproduct beoordelen door middel van een *peer assessment* is er sprake van een *static assessment* met een *formatief* gebruiksdoel. De term *data-driven decision making* is in Figuur 1.1 onderaan geplaatst. De reden hiervoor is dat alle toetsingen en dataverzamelingen die op school plaatsvinden hieronder vallen. Het gaat hier immers niet om een gebruiksdoel of eigenschap van één enkele toets; het is een werkwijze die beschrijft hoe scholen gericht kunnen werken aan het verhogen van leeropbrengsten door gebruik te maken van toetsinformatie.

Figuur 1.1 Overzicht verschillende categorieën van toetsgebruik



In deze literatuurstudie gezocht naar toetsvormen die vallen onder *formatief* toetsgebruik. Het kon daarbij zowel gaan om *static assessments* als *dynamic assessments*. We hanteren in deze literatuurstudie de Engelstalige termen voor dynamische toetsing (*dynamic assessment)* en diagnostische toetsing (*diagnostic assessment).* Tot dusver gingen we in op de verschillende functies van toetsing. De reden en de uitkomst verschilt per toetsing. Naast een indeling naar functie, kunnen we toetsen ook naar vorm indelen. Binnen elke categorie van toetsing wordt een groot aantal toetsvormen onderscheiden (zie ook, Hendriks & Schoonman, 2006; Schuurs & Verhoeven, 2010). Onder toetsvorm verstaan we "… het concrete pakket van regels en procedures dat voorschrijft hoe gedrag wordt uitgelokt, gescoord en geëvalueerd" (cf. Straetmans, 2006). Voorbeelden van toetsvormen zijn onder andere afstudeeropdrachten, essaytoetsen, kennis- en vaardigheidstoetsen, portfolio's, stage- of praktijkopdrachten, presentaties en peer- of self-assessments.

Een aantal *formatieve* toetsvormen wordt al vrij veel in het onderwijs gebruikt, terwijl andere toetsvormen nog nauwelijks worden toegepast, in het bijzonder als we alleen naar het taalonderwijs kijken. Het is de vraag of *formatieve* toetsvormen al gebruikt worden om de domeinen technisch lezen, begrijpend lezen, woordenschat, strategisch schrijven, en mondelinge taalvaardigheid te toetsen. Daarnaast is het zeer de vraag is of de toetsvormen en dan vooral de toetsen die op een dynamische manier worden afgenomen wel voldoen aan de belangrijkste toetstechnische eisen die we kunnen stellen (zie Wools et al., 2011). Zo zijn *portfolio's* vanwege de grote verschillen in aanpak, specificiteit en diepgang bijvoorbeeld lastig op een gestandaardiseerde en betrouwbare manier te beoordelen (e.g. Brown, 2002; Smith & Tillema 2007). *Self-assessments* blijken vaak zeer matig samen te hangen met de prestaties die leerlingen behalen op gestandaardiseerde toetsen (e.g. Ross, 1998; Alderson & Huhta 2005; Dlaska & Krekeler 2008; Matsuno 2009; Topping, 2003). Bij *peer-assessments* blijken leerlingen de feedback die zij krijgen van medeleerlingen niet altijd op prijs te stellen (Wesson, 2003). De validiteit van de meeste toetsvormen is nog nauwelijks onderzocht, al wordt er regelmatig op gewezen dat dergelijk onderzoek plaats moet vinden (Gysen & Van Avermaat 2005; Perie et al. 2009). In dit literatuuronderzoek is specifiek gezocht naar studies die ingaan op de psychometrische kwaliteit en/of effectiviteit van de toetsvorm die gebruikt wordt om een aspect van taalvaardigheid in kaart te brengen. Deze keuze heeft vermoedelijk tot gevolg dat een deel van de toetsvormen niet in dit onderzoek voorkomt, doordat de toetsvorm (a) niet past bij taal en lezen, of (b) niet aantoonbaar voldoet aan bepaalde minimale toetstechnische kwaliteitseisen. In zekere zin is dit jammer, omdat niet voor elk leerdoel een toetsvorm te beschrijven is. Ook is het aantal toetsvormen per leerdoel beperkt, hoewel het wenselijk is om te variëren in toetsvormen zodat leerlingen leren welke manier van werken hun het beste ligt en ontdekken waar ze sterk dan wel zwak in zijn (cf. Van de Mosselaer & Heylen, 2002). Aan de andere kant is het essentieel dat een toetsvorm door zowel de leerkracht als de leerling geaccepteerd wordt en informatie geeft die betrouwbaar en betekenisvol is.

1.4 Formatieve feedback

Een van de kenmerken van *formatieve* toetsing is dat feedback deel uitmaakt van de toetsprocedure. Uit de literatuurstudie van Sluijsmans, Joosten-ten Brinke en Van der Vleuten (2013) blijkt dat feedback een bewezen effectief onderdeel is van formatieve toetsingen. Volgens Shute (2007) is feedback "… intended to modify the learner's thinking or behavior for the purpose of improving learning." De feedback is een reactie op bepaald gedrag van de leerling en kan op verschillende manieren gegeven worden. De feedback kan door een persoon worden gegeven of door een computer. De feedback kan bestaan uit een hint, het meedelen of het antwoord goed of fout is, of het presenteren van een uitgewerkt praktijkvoorbeeld. Het moment waarop de feedback gegeven wordt, kan ook variëren. Soms wordt de feedback direct na het uitvoeren van een bepaalde taak gegeven, terwijl de feedback in andere gevallen pas na verloop van tijd gegeven wordt.

Een ander belangrijk onderscheid dat gemaakt kan worden, is het verschil tussen *contingente* en *non-contingente* feedback. Contingente feedback is niet of nauwelijks gestandaardiseerd en kan daardoor optimaal worden afgestemd op de behoefte van de individuele leerling. Non-contingente feedback is daarentegen sterk gestandaardiseerd en voor alle leerlingen in de klas gelijk (Caffrey, Fuchs & Fuchs, 2008). Contingente feedback zien we vooral terug in *klinische assessments* die vaak lang duren en nauwelijks gestandaardiseerd zijn. Non-contingente feedback wordt vooral in onderzoeksgerichte assessments toegepast. Deze assessments duren meestal korter en worden volgens strikte handleidingen, protocollen en scripts worden uitgevoerd. De ontwikkeling van een klinische assessment met contingente feedback kost de onderzoeker weinig voorbereiding, maar de uitvoering ervan trekt wel een zware wissel op de deskundigheid van de examinator. Gestandaardiseerde non-contingente feedback is mogelijk wat effectiever dan contingente feedback, zoals uit een meta-analyse van Caffrey, Fuchs en Fuchs (2008) blijkt.

Effectiviteit van feedback is ook afhankelijk van het type feedback dat gegeven wordt. Er valt onderscheid te maken tussen *verificatie* en *elaboratie* (Hattie & Timperley, 2007). Bij *verificatie* wordt informatie gegeven over de

juistheid van een antwoord, bijvoorbeeld goed of fout. Bij *elaboratie* geeft de leerkracht aanwijzingen. Het doel daarvan is om de leerling inzicht te geven in de strategieën die hij hanteert en nieuwe strategieën aan te leren. Elaboratieve feedback blijkt het meest effectief te zijn (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kluger & DeNisi, 1996), maar een combinatie van verificatie en elaboratie werkt vermoedelijk optimaal (cf. Kulhavy & Stock, 1989). Vormen waarbij de feedback onder verificatie valt, zijn (gebaseerd op Schute, 2007):

- ***Correct response.*** De leerling krijgt te horen wat het correcte antwoord is. Er volgt geen additionele informatie.
- ***Try-again.*** De leerling krijgt te horen dat het antwoord fout is en krijgt één of meer mogelijkheden om alsnog het correcte antwoord te geven.
- ***Error-flagging.*** De onjuistheden in de oplossing worden gemarkeerd, zonder daarbij het juiste antwoord te geven.

Onder elaboratie vallen de volgende feedbackvormen:

- ***Attribute isolation.*** De leerling krijgt informatie over een aantal kenmerkende eigenschappen van het concept dat getoetst wordt.
- ***Topic-contingent.*** Er wordt aanvullende informatie gegeven over het onderwerp dat getoetst wordt. Het kan gaan om extra instructiemateriaal.
- ***Response-contingent.*** Deze vorm richt zich op het antwoord dat de leerling geeft. Er wordt bijvoorbeeld toegelicht waarom een antwoord goed is of waarom een antwoord fout is.
- ***Hints/cues/prompts.*** Er wordt geprobeerd om de leerling via het geven van aanwijzingen in de juiste richting te sturen. Het correcte antwoord wordt niet expliciet gegeven.
- ***Bugs/misconceptions.*** De fouten en misvattingen van de leerling worden nauwgezet geanalyseerd en besproken.
- ***Informative tutoring.*** Dit is de meest uitgebreide vorm van feedback. Het antwoord wordt geverifieerd (verification), onjuistheden worden gemarkeerd (error-flagging) en er worden hints gegeven die helpen bij het vervolg. Het correcte antwoord wordt in de regel niet gegeven.

In deze literatuurstudie is de feedback die in studies op het gebied van de toetsing van taal- en leesvaardigheid toegepast wordt, geclassificeerd naar *contingent* en *niet-contingent*, en *verifiërend* en *elaboratief*. Er is bovendien vastgelegd of de feedback door een persoon (bijvoorbeeld een leraar, onderzoeker, taaltherapeut) of door de computer gegeven wordt.

1.5 Evalueren van opbrengsten bij taal en lezen

Met het oog op de vormgeving en planning van het onderwijs willen scholen, naast norm- of criteriumgeoriënteerde gegevens, ook graag beschikken over gegevens met betrekking tot het leerproces en het leerpotentieel van leerlingen. De methode-onafhankelijke toetsen en examens die scholen afnemen, geven die informatie niet. Ook methode-afhankelijke toetsen bieden in de regel nauwelijks handvatten om leerresultaten te analyseren en te duiden. Alternatieve toetsvormen zoals essaytoetsen, presentaties en kennis- en vaardigheidstoetsen die leerlingen maken met instructie en feedback van de leerkracht (i.e. dynamic assessment) kunnen scholen mogelijk wel houvast geven bij het vormgeven en evalueren van het taal- en leesonderwijs. Er is geprobeerd om op basis van een systematische literatuurstudie van wetenschappelijk onderzoek dat na 2000 verschenen is, antwoord te geven op de volgende onderzoeksvraag: “Welke toetsvormen kunnen leraren gebruiken om de vaardigheden van leerlingen op het gebied van technisch lezen, begrijpend lezen, woordenschat, strategisch schrijven, en mondelinge taalvaardigheid betekenisvol in kaart te brengen en te analyseren?” Een studie werd meegenomen als de auteur(s) aannemelijk konden maken dat de toetsvorm (a) andere informatie geeft dan de traditionele leerlingvolgsysteemtoetsen en eindexamens, (b) voldoet aan minimale toetstechnische kwaliteitseisen, en (c) een positief effect heeft op het leren van (groepen) leerlingen.

In dit rapport doen we verslag van het onderzoek en presenteren we de resultaten. In Hoofdstuk 2 gaan we in op de methoden en procedures die gevolgd zijn tijdens het onderzoek. We geven aan welk stappenplan gevolgd is, welke databases geraadpleegd zijn, en wat de zoektermen en inclusiecriteria waren. Daarnaast gaan we in op het design dat gevolgd is bij het analyseren en beoordelen van studies. In Hoofdstuk 3 vatten we de relevante studies samen en classificeren we de studies naar taaldomein, gebruiksdoel, toetsvorm, doelgroep, feedbackvorm en onderzoeksopzet. We besluiten dit rapport met een conclusie en discussie (Hoofdstuk 4) waarin we de onderzoeksresultaten samenvatten en in het perspectief van de Nederlandse onderwijspraktijk plaatsen. We doen daarnaast enkele methodologische en praktische aanbevelingen voor toekomstig ontwikkelwerk en vervolgonderzoek.

2 Methode

2.1 Procedure

Als leidraad voor de uitvoering van de kwalitatieve literatuurstudie hebben we een methodisch kader gehanteerd (Randolph, 2009). We hebben het onderzoeksproces, van literatuur zoeken en selecteren, gedocumenteerd. Bij het zoeken van relevante literatuur is er een fase van oriëntatie en selectie geweest. In de oriëntatiefase zijn de zoektermen bepaald waarmee een set van relevante literatuur kon worden geselecteerd. De uiteindelijke set met studies is door vier onderzoekers in een *balanced incomplete block design* beoordeeld. De beoordeling vond plaats op basis van een data-extractieformulier. Door gebruik van het formulier werden alle studies vergelijkbaar beoordeeld en werden overeenkomstige kenmerken vastgelegd. De belangrijkste thema's zijn geïdentificeerd en als kader gebruikt voor het kwalitatief beschrijven van de resultaten. De studies zijn schematisch in de resultatensectie weergegeven.

2.2 Databases en zoektermen

Voor het selecteren van relevante literatuur op het gebied van formatieve taaltoetsing is een 3-staps-procedure gehanteerd (zie Cooper, 1998):

1. Beschikbare studies zijn geïnventariseerd in de databases ERIC en Psycinfo op basis van relevante zoektermen;
2. Studies die niet online te downloaden waren, zijn bij auteurs opgevraagd;
3. De referenties uit de geselecteerde studies zijn bekeken om er zeker van te zijn dat alle relevante studies bij de review betrokken werden.

De definitieve zoekterm is geleidelijk aan tot stand gekomen. Eerst is naar literatuur gezocht op basis van de volgende zoektermen:

((dynamic*) OR (data driv*) OR (diagnost*) OR (formative) OR (rubric*)) AND ((test*) OR (assess*) OR (measure*) OR (evaluat*)) AND ((language) OR (reading comprehen*) OR (read*) OR (vocabulary) OR (spell*) OR (writ*)) AND ((educat*) OR (teach*) OR (curriculum)).

De zoektocht leverde een selectie van 3191 studies op. Uit deze lijst zijn *random* 100 studies geselecteerd die, in een volledig design, ter beoordeling aan vier onderzoekers zijn voorgelegd. De onderzoekers zijn bij de beoordeling van de titels en samenvattingen van de geselecteerde studies onafhankelijk van elkaar te werk gegaan. Bij de bespreking van de studies bleken zeer veel van de 100 geselecteerde studies niet bij de onderzoeksvraag te passen. Veel studies waren bijvoorbeeld niet gericht op de toetsing van taal- en leesvaardigheid in het onderwijs, maar waren afkomstig uit de medische wereld. Het ging dan bijvoorbeeld om het vaststellen van Alzheimer of schizofrenie op basis van toetsen. In overleg zijn de zoektermen aangescherpt. Bij de finale zoektocht is gesteld dat in de samenvatting tenminste één van de volgende (combinaties van) termen moeten voorkomen:

((formative assess*) OR (formative evaluat*) OR (formative test*) OR (dynamic assess*) OR (dynamic evaluat*) OR (dynamic test*) OR (diagnostic assess*) OR (diagnostic evaluat*) OR (diagnostic test*) OR (data driven assess*) OR (data driven evaluat*) OR (data driven test*)) AND ((formative assess*) OR (formative evaluat*) OR (formative test*) OR (dynamic assess*) OR (dynamic evaluat*) OR (dynamic test*) OR (diagnostic assess*) OR (diagnostic evaluat*) OR (diagnostic test*) OR (data driven assess*) OR (data driven evaluat*) OR (data driven test*)) AND ((language) OR (reading comprehen*) OR (read*) OR (vocabular*) OR (spell*) OR (writ*))

Zoektermen die werden uitgesloten en dus niet in de samenvatting mochten voorkomen waren:

((second language) OR (foreign language) OR (therap*) OR (medical) OR (psychiatric*) OR (dement*) OR (alzheimer) OR (immigrant*) OR (math*) OR (deaf*) OR (aphas*) OR (Asperger) OR (autis*) OR (alcohol*) OR (minorit*)).

Op basis van deze criteria zijn studies geselecteerd die in de periode van 2000 tot en met 2014 gepubliceerd zijn in een *peer reviewed* (inter)nationaal tijdschrift. De zoektermen leverden een lijst met 525 potentieel relevante studies op. De referenties van deze studies zijn geëxporteerd naar Excel. Er is gecontroleerd of studies niet dubbel in de lijst voorkwamen. De studies die een tweede keer in de lijst voorkwamen zijn verwijderd wat resulteerde in een lijst van 418 studies.

2.3 Selectieproces

De titels en samenvattingen van de 418 gevonden studies zijn op relevantie beoordeeld door gebruik te maken van de volgende inclusiecriteria:

1. De studie richt zich op de (formatieve) toetsing van taal- en/of leesvaardigheid;
2. De studie is uitgevoerd in het basis-, voortgezet of hoger onderwijs;
3. De toets- en feedbackvorm wordt duidelijk omschreven en er wordt aannemelijk gemaakt dat deze voldoen aan minimale (toets)technische kwaliteitseisen;
4. De studie beoogt het effect van de toetsvorm op het onderwijs en/of het leerproces van leerlingen te kwantificeren;
5. De studie richt zich *niet uitsluitend* op leerlingen met een stoornis of beperking;
6. De studie is, bij voorkeur na *blind peer review*, gepubliceerd in een (inter)nationaal wetenschappelijk tijdschrift.

De studies zijn in een *balanced incomplete block design* beoordeeld door vier onderzoekers. Figuur 2.1 laat zien welk design precies gebruikt is. We zien dat de totale verzameling met studies opgedeeld is in 13 sets; set A bevatte 50 studies en in de overige sets zaten 30 of 31 studies. De sets werden in de regel beoordeeld door twee onderzoekers. Alleen set A is door alle onderzoekers beoordeeld. Dit was tevens de eerste set die beoordeeld is. Na de beoordeling van deze set is gecontroleerd of de onderzoekers de inclusiecriteria op dezelfde wijze interpreteerden en toepasten. Dit bleek bij verreweg de meeste studies het geval te zijn. Het vervolg van de beoordeling (sets B-M) kon daarom zonder problemen voortgezet worden in tweetallen met een wisselende samenstelling. Elke onderzoeker heeft in totaal 234 samenvattingen beoordeeld.

Figuur 2.1 Design voor het beoordelen van de potentieel relevante studies

	Set												
Onderzoeker	A	B	C	D	E	F	G	H	I	J	K	L	M
1	■	■	■					■	■	■	■		
2	■			■	■	■	■			■	■		
3	■	■	■			■	■					■	■
4	■			■	■			■	■			■	■

De studies zijn op basis van de inclusiecriteria geclassificeerd als ‘geschikt’, ‘misschien geschikt’ of ‘ongeschikt’. Studies die als ‘niet geschikt’ werden beoordeeld, werden uit de database verwijderd. Studies die als ‘misschien geschikt’ of ‘geschikt’ werden geclassificeerd, werden behouden in de database en de complete tekst werd opgezocht voor gedetailleerde beoordeling. Studies die niet online te verkrijgen waren, zijn bij de auteurs opgevraagd.

2.4 Data extractie en beoordeling

In totaal zijn 91 studies als ‘geschikt’ of ‘misschien geschikt’ aangemerkt. Deze studies zijn verdeeld over vier onderzoekers en vervolgens in detail beoordeeld aan de hand van een data-extractieformulier. Er werd gestart met het analyseren en beschrijven van de inhoud van een studie. In grote lijnen werden de volgende gegevens op systematische wijze vastgelegd:

Algemeen

- Titel, auteur, tijdschrift en jaar van publicatie
- Onderwerp, onderzoeksvragen en doelstellingen

Onderzoeksopzet

- Onderwijstype en doelgroep
- Steekproefomvang en design

Instrumentarium

- Toets- en feedbackvorm
- Afnameprocedure
- Toetstechnische kwaliteit

Resultaat

- Belangrijkste uitkomsten
- Effectiviteit en bruikbaarheid

De tweede helft van het beoordelingsformulier had betrekking op de kwaliteit van de studie. Bij het beoordelen van de kwaliteit is de methode van Petticrew en Roberts (2006) gevolgd. Dit betekent dat elke studie beoordeeld is op de volgende punten: (a) algemene oriëntatie, (b) steekproeftrekking, (c) methode, (d) data en statistische analyses, en (e) conclusie. Op basis van de uitgebreide beschrijvingen van de studies en de kwaliteitsoordelen zijn de 91 studies definitief beoordeeld op relevantie. Tussentijds vond overleg plaats om de twijfelgevallen en de voortgang te bespreken.

Van de 91 beoordeelde studies bleken slechts een aantal aan *alle* geformuleerde inclusie- en kwaliteitscriteria te voldoen. Dit waren in de regel de studies die ook al eerder, op basis van de samenvatting, het oordeel 'geschikt' toegekend hadden gekregen. Omdat het aantal studies zo beperkt was, is in een enkel geval besloten om iets ruimer met de inclusie- en kwaliteitscriteria om te springen. De studie moest dan wel relevante informatie geven met betrekking tot inclusiecriteria 1 en 2. Ook moest de toets- en feedbackvorm duidelijk omschreven zijn en in potentie meerwaarde hebben ten opzichte van de traditionele leerlingvolgsysteemtoetsen en eindexamens. Voor inclusiecriterium 5 geldt dat deze uiteindelijk minder strikt is toegepast dan vooraf beoogd werd. Zoals in paragraaf 2.3 is aangegeven, wilden we geen studies in de literatuurstudie opnemen die uitsluitend gericht waren op leerlingen met een stoornis of beperking. Dit criterium bleek onhoudbaar, omdat veel studies een formatieve (taal)toetsing ontwikkelen en evalueren met het oog op het identificeren van leerproblemen. Er is besloten om dergelijke studies in de literatuurstudie op te nemen mits er ook "reguliere" leerlingen aan de (effect)studie meededen. Ten slotte zijn ook de literatuurlijsten van de geschikte studies bestudeerd. Dit heeft geleid tot enkele toevoegingen. Soms ging het om studies die voor 2000 gepubliceerd zijn.

3 Resultaten

In dit hoofdstuk presenteren we de beschrijvende resultaten van de literatuurstudie die in totaal veertien relevante studies heeft opgeleverd. We beginnen met een algemene typering van de veertien gevonden studies (paragraaf 3.1). Vervolgens bespreken we negen studies naar *dynamic assessment* waarbij de feedback aan de leerling door een persoon (bijvoorbeeld onderzoeker, leraar, taaltherapeut) gegeven wordt (paragraaf 3.2). Tot slot bespreken we vijf studies waarbij de feedback geautomatiseerd via de computer wordt verstrekt (paragraaf 3.3).

3.1 Algemene kenmerken van de gevonden studies

De belangrijkste kenmerken van de veertien gevonden studies zijn samengevat in Tabel 3.1. Zoals in de methode is toegelicht is er een onderscheid tussen studies die voldoen aan de opgestelde zoekcriteria en studies die daar niet volledig aan voldoen en later zijn toegevoegd. In de eerste kolom van Tabel 3.1 zijn de later toegevoegde studies met een asterisk aangeduid. In alle gevallen gaat het om onderzoek zonder nameting en/of zonder controlegroep. In deze paragraaf gaan we eerst in op een aantal inhoudelijke en methodologische kenmerken van de studies.

- ***Vorm van assessment.*** Geregistreerd is welke vorm van assessment de onderzoekers gebruikt hebben. Hierbij is uitgegaan van de door de onderzoekers zelf gehanteerde termen. Van de veertien onderzochte studies gaan er volgens opgave van onderzoeker twee over de effectiviteit of bruikbaarheid van formatief toetsgebruik, tien over *dynamic assessment* en twee over *diagnostic assessment*.
- ***Taaldomein.*** Nagegaan is op welke taaldomein de toetsing in hoofdzaak betrekking heeft. Dit wil zeggen: de taalvaardigheid die gemeten wordt met de toets waarvan de onderzoekers de effectiviteit of bruikbaarheid willen vaststellen. Van alle veertien studies hebben er vijf betrekking op woordenschat, twee op mondelinge taalvaardigheid, vier op leesvaardigheid, twee op schrijfvaardigheid en één op decodeervaardigheid.
- ***Leeftijd, leerjaar en onderwijssegment.*** Nagegaan is wat de leeftijd was van de leerlingen die aan de toetsing deelnamen en in welke onderwijssector zij zich bevonden (dit wil zeggen: kleuteronderwijs, basisonderwijs, voortgezet onderwijs, middelbaar beroepsonderwijs, hoger beroepsonderwijs en universiteit). In geval de publicatie geen informatie over de leeftijd bevatte, is het leerjaar of de klas van de gevolgde opleiding gerapporteerd. Van de veertien studies zijn er drie uitgevoerd bij kleuters, zeven bij leerlingen in het basisonderwijs, drie in het voortgezet onderwijs, geen in het middelbaar beroepsonderwijs, geen in het hoger beroepsonderwijs en één in het universitair onderwijs.
- ***Type leerlingen.*** Voor zover de beschikbare informatie het toeliet, is in Tabel 3.1 een nadere omschrijving van de onderzochte groep leerlingen gegeven, zoals een indicatie van de samenstelling naar sociaal-economische of raciaal-etnische achtergrond.
- ***Taalontwikkeling.*** Met betrekking tot de taalontwikkeling van de onderzochte leerlingen is een onderscheid gemaakt in a) leerlingen met een normale taalontwikkeling en b) leerlingen met een taalachterstand of -stoornis. Aan acht van de veertien studies deden alleen zich normaal ontwikkelende leerlingen mee, en zes studies kenden zowel leerlingen met een normale taalontwikkeling als leerlingen met een taalachterstand of -stoornis.
- ***Type feedback.*** De aard van de feedback is gedefinieerd aan de hand van drie dimensies (zie paragraaf 1.4). Bij de eerste dimensie gaat het om de vraag of de feedback door een persoon (onderzoeker, leraar, taaltherapeut) dan wel door de computer gegeven wordt. Bij de tweede dimensie gaat het om het onderscheid contingent en non-contingent. De derde dimensie maakt een onderscheid in verifiërende en elaboratieve feedback. Van de veertien aangetroffen studies gebruikten er negen menselijke feedback en bij de overige vijf werd de feedback geautomatiseerd via de computer verstrekt. Begrijpelijkerwijs maakten alle vijf studies naar gecomputeriseerde feedback gebruik van non-contingente (gestandaardiseerde) feedback. Drie van de negen studies met menselijke feedback kenden contingente feedback, bij vijf studies was sprake van non-contingente feedback en in één studie werd zowel contingente als non-contingente feedback gegeven. Bij de derde dimensie stond het onderscheid tussen verifiërende en elaboratieve feedback centraal zoals gedefinieerd door Shute (2006). Bij vier studies lag het accent op verifiërende feedback: de leerlingen kregen eenvoudige informatie over het al dan niet juist zijn van het antwoord zonder aanwijzingen voor verbetering. Bij tien studies stond elaboratieve feedback centraal: de leerlingen kregen uitgebreide informatie, zoals hints, aanwijzingen, uitleg, modellen, voorbeelden of wijzigingssuggesties, om ze naar het goede antwoord te leiden.

- ***Onderzoeksontwerp.*** Een experimenteel ontwerp met een voor- en nameting bij een experimentele en controlegroep biedt meer mogelijkheden om het effect van een toetsing eenduidig vast te stellen dan een ontwerp zonder nameting en/of zonder controlegroep (Shadish, Cook & Campbell, 2002). Met betrekking tot het onderzoeksontwerp, dat gebruikt wordt om het effect van de toetsing vast te stellen, onderscheiden Sternberg & Grigorenko (2002) een sandwichmodel en een cakemodel. In het geval van het sandwichmodel wordt de toetsing uitgevoerd volgens een voormeting-instructie-nameting ontwerp. In het zogeheten cake-model is er geen sprake van een voor- en nameting. De hoeveelheid en aard van de instructie wordt op adaptieve wijze gevarieerd al naar gelang de hulpbehoefte van de leerling. Het cakemodel wordt ook wel het *train-within-test* model genoemd. Op basis van het onderzoeksontwerp dat gebruikt wordt om vast te stellen in hoeverre het beoogde doel van de toetsing bereikt wordt, is een onderscheid gemaakt in a) studies met zowel een voor- en nameting als een experimentele en controlegroep, b) studies met alleen een voor- en nameting, c) studies met alleen een experimentele en controlegroep en d) studies zonder nameting en zonder controlegroep. Van de veertien studies kenden er drie zowel een voor- en nameting als een experimentele en controlegroep, zes hadden alleen een voor- en nameting en de overige vijf kenden slechts één meetmoment zonder controlegroep of hadden twee meetmomenten zonder dat er sprake was van instrumentele overlap tussen de eerste en tweede meting (waardoor het niet mogelijk is om de leerwinst ten gevolge van de deelname aan de toetsingsprocedure vast te stellen).
- ***Type toets effectmeting.*** Er kunnen verschillende typen toetsen worden ingezet om het effect of de bruikbaarheid van de toetsingsprocedure vast te stellen. Daarbij maken we een onderscheid in a) norm-georiënteerde toetsen, b) criterium-georiënteerde toetsen, c) *dynamic assessment* toetsen en d) *dynamic assessment*-observatielijsten. Norm- en criteriumgeoriënteerde toetsen zijn vormen van statische toetsing. Binnen een voormeting-instructie-nameting ontwerp worden ze ingezet bij de voor- en nameting. Tot de categorie criterium-georiënteerd behoren ook toetsen die de onderzoeker zelf heeft gemaakt of uit bestaand toetsmateriaal heeft samengesteld (Caffrey et al., 2008). De *dynamic assessment* toetsen zijn eveneens door de onderzoeker zelf gemaakt, waarbij de inhoud en structuur in hoge mate is afgestemd op wat er tijdens de interventie is aangeboden. *Dynamic assessment* observatielijsten meten geen vak- of taalvaardigheden, maar allerlei ondersteunende vaardigheden die van belang zijn voor het vaststellen van het leerpotentieel (zoals aandacht, plannen, zelfregulatie, impulscontrole), de gevoeligheid voor de interventie, de moeite die de examinator heeft moeten doen en de transfer naar andere dan de onderwezen vaardigheden (Tzuriel, 2001). Als effectmaat werd er zes keer een genormeerde toets ingezet, vijf keer een criterium-georiënteerde toets, zes keer een *dynamic assessment* toets en vijf keer een *dynamic assessment*-observatielijst. Hierbij merken we op dat het totaal aantal geregistreerde toetsvormen hoger is dan het aantal studies omdat in één onderzoek meer toetsvormen gebruikt kunnen worden.
- ***Toetsfunctie.*** Ten aanzien van de functie van toetsing is in paragraaf 1.3 een onderscheid gemaakt in zeven, elkaar deels overlappende categorieën: 1) intake en selectie, 2) plaatsing, 3) voortgangscontrole, 4) vaststellen leerpotentieel, 5) diagnostiek, 6) verbetering leerproces en 7) certificering. In één van de veertien gevonden studies is de toetsfunctie beperkt tot het verbeteren van het leerproces, in één studie gaat het (daarnaast) om het vaststellen van het leerpotentieel (zonder de pretentie van screening en plaatsing), in zeven gevallen gaat het om het vaststellen van het leerpotentieel ten behoeve van het screenen van leerlingen met het oog op plaatsingsbeslissingen en in vijf gevallen hebben de toetsen een diagnostische functie. De overige drie functies – voortgangscontrole, intake en selectie, en certificering – komen niet voor. Bij het coderen hebben we de toetsfunctie van de geautomatiseerde feedbacksystemen in alle vijf gevallen als *diagnostic assessment* geclassificeerd (ook al benoemde de onderzoeker de studie zelf als formatief of *dynamic*). Een eerste reden is dat het onderscheid tussen *dynamic assessment* en *diagnostic assessment* hier veel minder helder is dan bij toetsprocedures met menselijke feedback het geval is. Eén van de basale onderscheidende kenmerken tussen *diagnostic* en *diagnostic assessment* is de ‘live’ interactie tussen leerling en volwassene tijdens de afname van de toets (Feuerstein, Rand & Rynders, 1988; Sternberg & Grigorenko, 2002). In het geval de digitale feedback volledig geautomatiseerd binnen een *train-within-test* ontwerp gegeven wordt, is er van ‘echte’ interactie met een volwassene geen sprake, ook al pretenderen sommigen dat een computer zich als een menselijke partner of leerkracht kan gedragen (Crook, 1991). Een tweede reden is dat er in alle vijf studies sprake is van gedetailleerde feedback over de sterke en zwakke kanten van de geleverde prestatie, wat een typisch kenmerk is van *diagnostic assessment* maar niet noodzakelijkerwijs ook kenmerkend is voor *dynamic assessment*. Vandaar dat we de toetsfunctie in de studies naar geautomatiseerde feedbacksystemen (Franzke et al., 2005; Ferster et al., 2012; Teo, 2012) als diagnostisch benoemd hebben.

Tabel 3.1 Demografische en inhoudelijke kenmerken van de aangetroffen studies

Studie	Plaats in rapport (pagina)	Vorm assessment	Taaldomein	Leeftijd of leerjaar	Type leerling	Taalontwik-keling (taalachter-stand of -stoornis)	Feedback: contingent versus non-contingent	Feedback: verifiërend versus elaboratief	Toetsfunctie	Ontwerp	Type toets effectmaat
Kester, Peña & Gillam (2001)	22	dynamisch	woordkennis (labeling skills)	3-4 jaar (N = 52)	laag-SES kinderen (m.n. Hispanics)	zonder en met	contingent en non-contingent door mens	elaboratief	vaststellen leerpotentieel	voormeting-instructie-nameting met controlegroep	norm-georiënteerd
Peña, Quinn & Iglesias (1992)	23	dynamisch	woordkennis (expressive and receptive categorisation)	3-9 jaar (N = 50)	Puerto Ricaanse en Afrikaans-Amerikaanse kinderen	zonder en met	contingent door mens	elaboratief	vaststellen leerpotentieel + screening/ plaatsing	voormeting-instructie-nameting	norm-georiënteerd + *dynamic assessment*-observatie
Ukrainetz, Harpell, Walsh & Coyle (2000)	25	dynamisch	woordkennis (expressive and receptive categorisation)	kleuters (N = 23)	Indiaanse kinderen van de Arapahoe en Shoshone stam	zonder en met	contingent door mens	elaboratief	vaststellen leerpotentieel + screening/ plaatsing	voormeting-instructie-nameting	norm-georiënteerd + *dynamic assessment*-observatie
Kapantzoglou, Adelaida Restrepo en Thompson (2010)	26	dynamisch	woordkennis (aanleren van nieuwe woorden)	4-5 jaar (N = 28)	laag-SES kinderen (m.n. Spaanstalige achtergrond)	zonder en met	non-contingent door mens	elaboratief	vaststellen leerpotentieel + screening/ plaatsing	voormeting-instructie-nameting	*dynamic assessment*-toets + observatie
Larsen & Nippold (2007)*	27	dynamisch	woordkennis (morfologische analyse t.b.v. afleiden woordbetekenissen)	10-12 jaar (N = 50)	mainstream (zich normaal ontwikkelend)	zonder	non-contingent door mens	elaboratief	verbetering leerproces	geen nameting geen controlegroep	norm-georiënteerd + *dynamic assessment*-toets
Peña, Gillam, Malek, Ruiz-Felter, Resendiz, Fiestas en Sabel (2006)	28	dynamisch	spreekvaardigheid (navertellen tekstloze beeldverhalen)	leerjaar 1 en 2 (N = 58)	kinderen van diverse raciaal-etnische afkomst	zonder en met	non-contingent door mens	elaboratief	vaststellen leerpotentieel + screening/ plaatsing	voormeting-instructie-nameting met controlegroep	*dynamic assessment*-toets + observatie
Kramer, Mallett, Schneider en Hayward (2009)	30	dynamisch	spreekvaardigheid (navertellen tekstloze beeldverhalen)	klas 3 (N = 17)	Indiaanse kinderen uit de Samson Cree Nation Reserve (Canada)	zonder en met	contingent door mens	elaboratief	vaststellen leerpotentieel + screening/ plaatsing	voormeting-instructie-nameting	*dynamic assessment*-toets + observatie
Elleman, Compton, Fuchs, Fuchs & Bouton (2011)	31	dynamisch	leesvaardigheid (inferenties)	leerjaar 2 (N = 100)	kinderen van diverse raciaal-etnische afkomst	zonder	non-contingent door mens	elaboratief	vaststellen leerpotentieel + screening	voormeting-instructie-nameting	norm-georiënteerd + *dynamic assessment*-toets

Tabel 3.1 Vervolg

Studie	Plaats in rapport (pagina)	Vorm assessment	Taaldomein	Leeftijd of leerjaar	Type leerling	Taalontwik-keling (taalachter-stand of -stoornis)	Feedback: contingent versus non-contingent	Feedback: verifiërend versus elaboratief	Toetsfunctie	Ontwerp	Type toets effectmaat
Fuchs, Compton, Fuchs, Bouton & Caffrey (2011)*	33	dynamisch	decoderen	kleuters/ leerjaar 1 (N = 318)	kinderen van diverse raciaal-etnische afkomst	zonder	non-contingent door mens	elaboratief	vaststellen leerpotentieel + screening/ plaatsing	geen nameting geen controlegroep	norm-georiënteerd + *dynamic assessment*-toets + observatie
Franzke, Kintsch, Caccamise, Johnson & Dooley (2005)	34	formatief	schrijfvaardigheid	13-14 jaar /leerjaar 8 (N = 121)	24% leerlingen uit minderheidsgroepen	zonder	non-contingent via computer	verifiërend	diagnostisch	voor- en nameting met controlegroep	criterium-georiënteerd
Teo (2012)	37	dynamisch	leesvaardigheid (inferenties)	18-19 jaar (N = 68)	eerstejaars studenten	zonder	non-contingent via computer	elaboratief	diagnostisch	voormeting-instructie-nameting	criterium-georiënteerd
Sainsbury & Benton (2011)*	38	diagnostisch	leesvaardigheid	5–7 jaar (N > 600)	± landelijk representatieve steekproef	zonder	non-contingent via computer	verifiërend	diagnostisch	geen nameting geen controlegroep	criterium-georiënteerd
Kalyuga (2006)*	39	diagnostisch	leesvaardigheid	± 13 jaar/ leerjaar 7 (N = 34)	alleen jongens	zonder	non-contingent via computer	verifiërend	diagnostisch	geen nameting geen controlegroep	criterium-georiënteerd
Ferster, Hammond, Alexander & Lyman (2012)*	36	formatief	schrijfvaardigheid (documentary writing)	leerjaar 7 midden-school (N = 87)	leerlingen van diverse raciaal-etnische afkomst	zonder	non-contingent via computer	verifiërend	diagnostisch	geen nameting geen controlegroep	criterium-georiënteerd

3.2 Studies met menselijke feedback

Van de veertien gevonden studies hebben er negen betrekking op de effectiviteit of bruikbaarheid van *dynamic assessment* waarbij de feedback door een mens gegeven wordt. In deze paragraaf bespreken we deze negen *dynamic assessment* studies. Vijf ervan hebben betrekking op woordenschat (paragraaf 3.2.1), twee op mondelinge taalvaardigheid (paragraaf 3.2.2), twee op leesvaardigheid, waarvan één op leesbegrip en één op technisch lezen (paragraaf 3.2.3). *Dynamic assessment* is een specifieke vorm van *assessment for learing* die gekenmerkt wordt door eigen rationales, onderzoeksmethoden, meetinstrumenten en jargon. Op het gebied van toetsing van intelligentie is *dynamic assessment* al veelvuldig met succes toegepast (Resing, 2006; Van der Aalsvoort, Resing & Ruijssenaars, 2002). In het domein van de taalvaardigheidstoetsing is *dynamic assessment* echter een nog relatief onbekend en weinig gebruikt fenomeen. Om de resultaten van de individuele *dynamic assessment* studies beter te kunnen interpreteren, bespreken we hierna eerst enkele kenmerken die de aangetroffen studies naar *dynamic assessment* gemeen hebben.

In de gevonden studies naar *dynamic assessment* is de aanleiding voor het onderzoek de breed gedeelde bezorgdheid over de validiteit en bruikbaarheid van de gebruikelijke 'statische' toetsen voor kinderen uit minderheidsgroepen. Statische toetsen zijn summatieve, norm-georiënteerde toetsen die volgens een strikt protocol worden afgenomen en waarbij de examinator de leerling geen instructie, hulp of feedback mag geven. De veronderstelling is dat statische toetsen de prestaties van bepaalde groepen leerlingen onderschatten. Tot die groepen behoren onder meer leerlingen met een achterstand in de instructie- en toetstaal, leerlingen die door culturele factoren minder gelegenheid hebben gehad om zich de getoetste kennis en vaardigheden eigen te maken of die minder vertrouwd zijn met de eisen die de afnamesituatie aan hen stelt en leerlingen uit het speciaal onderwijs die vanwege cognitieve, sociaal-emotionele en/of fysieke beperkingen structureel hulpbehoevend zijn. De gebruikelijke toetsen zijn voor deze groepen risicoleerlingen vaak uitzonderlijk moeilijk waardoor ze de toekomstige schoolprestaties slecht voorspellen. Daardoor komen deze leerlingen nogal eens ten onrechte in speciale onderwijsprogramma's terecht, terwijl ze met passende hulp en begeleiding gewoon in het reguliere onderwijs hadden kunnen blijven.

Door de interactie tussen leerling en examinator biedt een *dynamic assessment* een meer natuurlijke omgeving voor leren dan een statische toets, vooral als de leerling door culturele verschillen minder vertrouwd is met de toetstaken (Gutiérrez-Clellen, 1996). De interactie met de leerkracht tijdens de afname biedt de leerling meer mogelijkheden om zich de getoetste kennis en vaardigheden eigen te maken, wat ten goede kan komen aan het zelfbeeld en de motivatie. De interactie tussen leerling en examinator vermindert de zogeheten vertrouwdheidsbias en daarmee eventuele toetsangst (Carlson & Wiedl, 1992). Voor leerlingen uit minderheidsgroepen is *dynamic assessment* een minder partijdige maat voor de schoolprestaties, omdat er een minder groot beroep wordt gedaan op gemiddelde taalvaardigheid en achtergrondkennis en ervaring met taken (Caffrey et al, 2006, p. 256). De bruikbaarheid van *dynamic assessments* voor minderheidsgroepen is eerder al onderzocht op het gebied van intelligentie. Zo vonden Kaniel et al. (1991) dat Ethiopisch immigranten op een *statische* IQ-test significant lagere scores behaalden dan geboren en getogen Israëli. Dit leek niet zozeer het gevolg van een lager IQ, maar de immigranten waren minder bekend met het maken van testtaken. Zij onderwierpen de immigranten aan een kort programma van *dynamic assessment* waarbij de immigranten instructie over het uitvoeren van taken kregen. Hierna bleken immigranten niet langer onder te presteren op de IQ-test, waaruit de onderzoekers concludeerden dat *dynamic assessment* een effectief middel is om leerachterstanden te remediëren. Het grote nadeel is echter dat de minder sterke standaardisatie van de afnamecondities ten koste kan gaan van de objectiviteit en betrouwbaarheid van de meting. Een ander nadeel is dat de afname vaak bewerkelijker en tijdrovender is dan bij een statische toets het geval is.

Van de negen gevonden *dynamic assessment* studies hadden er zeven een onderzoeksontwerp volgens het sandwich model (zie paragraaf 3.1), met een voor- en nameting(waarvan twee studies naast een experimentele ook een controlegroep kenden). Met een voormeting wordt eerst de baseline score vastgesteld. Omdat de leerling de toets geheel zelfstandig maakt, is er sprake van een statische toets. Daarna volgt een (korte) interventie van onderwijs en leren, vaak aangeduid als de fase van gemedieerde leerervaring, waarbij bijvoorbeeld de leraar de mediator is. De mediatie bestaat uit instructie, met bijvoorbeeld hints en aanwijzingen, die wordt gegeven tot het leerdoel bereikt is en de leerling de taak zelfstandig kan uitvoeren (Resing, 2006). Het is daarbij van belang dat de leraar of examinator de leerling niet méér instructie geeft dan voor een zelfstandige uitvoering vereist is (een kleine hint kan al voldoende zijn om het leerpotentieel te activeren). Daarnaast is het van belang dat de hoeveelheid hulp en de mate van expliciete instructie langzaam toeneemt al naar gelang de behoefte van de leerling. De leraar registreert hoeveel en welke instructie (bijvoorbeeld hints, aanwijzingen, uitleg, modelleren) de leerling nodig heeft gehad. Daarbij wordt regelmatig gebruikgemaakt van observatielijsten (e.g., Lidz, 1987, 1991;

Peña, 1993). Na afloop van de interventie wordt de test van de voormeting of een parallelversie ervan nogmaals afgenomen om de leerwinst te bepalen. Op basis van leerresultaten van de leerling tijdens de instructie en op de nameting, trekt de leraar conclusies over het effect van de interventie, de strategieën die de leerling gebruikt bij het leren en de hoeveelheid en aard van de instructie, hulp en begeleiding die in het vervolgtraject nodig is (Feuerstein, 1979; Lidz, 1991; Peña, 1993). De hoeveelheid benodigde instructie wordt beschouwd als een omgekeerde maat voor leerpotentieel (Resing, 2006). Leerlingen die weinig instructie nodig hebben maar toch sterk vooruitgaan, worden geacht over een groot leerpotentieel te beschikken. Het grote voordeel van *dynamic assessment* is dat het procesgerichte informatie oplevert die direct gerelateerd is aan het voorafgaande onderwijs, via directe communicatie tussen leerkracht en leerling tot stand is gekomen en daardoor onmiddellijk bruikbaar is om het leerproces te verbeteren.

3.2.1 Woordenschat

Van de negen studies naar de effectiviteit en bruikbaarheid van *dynamic assessment* met menselijke feedback voor het toetsen en ontwikkelen van taalvaardigheid hebben er vijf betrekking op woordenschat. In de studies van Ukrainetz, Harpell, Walsh en Coyle (2000), Peña, Quinn en Iglesias (1992) en Kapantzoglou, Adelaida Restrepo en Thompson (2010) was er wel een voor- en nameting, maar geen controlegroep. Slechts één van de vijf studies naar woordkennis kende zowel een voor- en nameting als een experimentele en controlegroep: Kester, Peña en Gillam (2001). Kester et al. (2001) onderzochten bij kleuters de effectiviteit van drie verschillende *dynamic assessment* interventies op vaardigheden in het labelen van woordbetekenissen. Peña et al. (1992) deden onderzoek naar de bruikbaarheid van een *dynamic assessment* met het categoriseren van woorden voor het identificeren van basisschoolleerlingen uit minderheidsgroepen met ernstige taalstoornissen. Ukrainetz et al. (2000) repliceerden het onderzoek van Peña, Quinn en Iglesias (1992) met een *dynamic assessment* voor productieve en receptieve categorisatievaardigheden bij kleuters van Indiaanse afkomst. Kapantzoglou et al. (2010) deden een soortgelijk onderzoek bij tweetalige kinderen en gebruikten een *dynamic assessment* waarbij nieuwe woorden aan de hand van verbale en visuele ondersteuning werden aangeleerd. Het onderzoek van Larsen en Nippold (2007) voldoet aan geen van onze twee methodologische criteria maar is toch opgenomen omdat het een nieuwe toetsprocedure betreft die relevant kan zijn voor leraren. Zij beschrijven hoe morfologische kennis als een efficiënte woordleerstrategie getoetst kan worden.

Studie 1: Kester, Peña en Gillam (2001)

- ***Achtergrond en vraagstelling.***.De kern van een *dynamic assessment* is de interventie. Cronen, Silver-Pacuilla en Condelli (2006) definiëren een interventie als "… a comprehensive program that includes curriculum development, adaptation of materials, teacher training, ongoing teacher support, and assessment" (p. 7). Op basis van de aard van de interventie onderscheidt Campione (1989) twee typen van *dynamic assessment*: klinische versus gestandaardiseerde, ook wel onderzoeksgerichte, *dynamic assessments* genoemd (Caffrey et al., 2008). De interventies in een klinisch-gerichte *dynamic assessment* nemen meestal veel tijd in beslag en de uitvoering ervan is niet of nauwelijks gestandaardiseerd. Onderzoeksgerichte *dynamic assessment* interventies duren meestal maar kort, terwijl de uitvoering ervan gestandaardiseerd is aan de hand van gedetailleerde handleidingen, protocollen of scripts. Het onderscheid tussen de beide benaderingen van *dynamic assessment* zien we terug in de aard van de feedback waarbij een onderscheid gemaakt wordt in contingente feedback, die optimaal is toegesneden op de individuele leerling, en sterk gestandaardiseerde, non-contingente feedback die voor alle leerlingen in de klas gelijk is (Caffrey et al., 2008). Klinisch gerichte *dynamic assessment*s kosten minder voorbereiding en bieden meer flexibiliteit dan onderzoeksgerichte *dynamic assessments*, maar trekken wel een zware wissel op de deskundigheid van de examinator. In de praktijk worden onderzoeksgerichte *dynamic assessments* vooral gebruikt voor screening op leerstoornissen ten behoeve van plaatsingsbeslissingen. De tijd die de leerling nodig heeft om het leerdoel te bereiken, de benodigde hoeveelheid hulp, de mate van expliciete instructie en de behaalde leerwinst vormen daarbij indicatoren voor het leerpotentieel. Er is nog maar weinig onderzoek gedaan naar de effectiviteit van verschillende manieren om de instructiecomponent binnen een *dynamic assessment* vorm te geven. Kester, Peña en Gillam (2001) deden dat wel in deze studie. Zij onderzochten het effect van drie onderzoeksgerichte *dynamic assessments* op de woordenschatontwikkeling van 52 kleuters (leeftijd 3-4) uit laag-SES gezinnen (42 Spaans, 8 Afro-Amerikaans en 2 blank). Aanleiding is het gegeven dat deze leerlingen op traditionele statische toetsen zonder instructiecomponent vaak onderpresteren. Daardoor krijgen zij vaak ten onrechte de diagnose taalstoornis en worden zij naar het speciaal onderwijs verwezen, terwijl zij met passende hulp en begeleiding gewoon in het normale onderwijs hadden kunnen blijven.

- ***Onderzoeksopzet.*** Gebruikmakend van een experimenteel voormeting-instructie-nameting ontwerp vergeleken Kester et al. (2001) het effect van drie *dynamic assessment*-methoden op de vaardigheid in labelen: 1) directe instructie, 2) gemedieerde leerervaring en 3) een hybridemethode waarin beide methoden gecombineerd werden. Daarnaast was er een controlegroep die wel meedeed aan de voor- en nameting maar geen instructie ontving. De woordenschat (labelen) werden gemeten met de *Expressive One-Word Picture Vocabulary Test-Revised* (EOWPVT-R) (Gardner, 1990) die individueel werd afgenomen.

- ***Dynamic assessment.*** In de drie *dynamic assessment*-condities werden twee instructieprocedures en twee typen materiaal met elkaar vergeleken. Eén instructieprocedure was gebaseerd op gedragstherapeutische principes. De interactie werd gestuurd door het antwoord dat op de toets gegeven werd, waarbij een goed antwoord positief bekrachtigd werd (*antecendent-response-consequence)*. De andere instructieprocedure was gebaseerd op cognitieve modificatie (*scaffolding interactions*). Er werden twee typen materiaal gebruikt: contextrijk (concrete objecten) versus contextarm (worksheets waarbij woordenschatitems werden verduidelijkt aan de hand van lijnen en vormen). De directe-instructie-interventie bestond vooral uit het drillen van specifieke woordenschatitems volgens de eerste instructieprocedure aan de hand van contextrijk materiaal (objecten). In de cognitieve modificatie conditie lag de nadruk op het aanleren van metacognitieve strategieën aan de hand van contextarm materiaal (worksheets). De hybride conditie combineerde de cognitieve modificatieprocedure met het contextrijke materiaal van de directie-instructie-methode. De combinatie van gedragsinstructie en contextarm ontbrak omdat contextarm materiaal van nature abstract is en alleen met behulp van cognitieve strategieën onderwezen kan worden.

- ***Resultaten.*** In alle drie *dynamic assessment*-condities kenden de leerlingen meer woorden dan de leerlingen in de controlegroep van wie de woordenschat niet vooruitging. De leerwinst was niet in alle drie *dynamic assessment*-condities gelijk. De leerlingen die instructie volgens de gemedieerde leerervaring en Hybride methode kregen, gingen meer vooruit (*d* respectievelijk 1.13 en .66) dan degenen die directe instructie ontvingen (*d* = .35). De leerwinst in de gemedieerde leerervaring en hybride conditie bleek toe te schrijven te zijn aan de cognitieve modificatieprocedure en niet aan het type materiaal. De gemedieerde leerervaring- en hybride condities verschilden namelijk alleen voor wat betreft het gebruik van het type materiaal en niet qua instructieprocedure die in beide gevallen uit cognitieve modificatie bestond.

- ***Conclusie.*** De onderzoekers schrijven het succes van de *dynamic assessment* toe aan de mogelijkheid om de instructie aan te passen aan verschillen tussen kinderen in voorkennis, persoonlijkheid, taalinvloed en voorafgaande onderwijservaringen. Het gebruik van cognitieve modificatie volgens de gemedieerde leerervaring en de hybride methode draagt ertoe bij dat kinderen uit laag-SES gezinnen taaltaken in overeenstemming met hun mogelijkheden kunnen uitvoeren. Het gebruik van deze *dynamic assessment* methoden zal naar verwachting leiden tot een nauwkeurigere toetsing van kinderen uit laag-SES gezinnen. Daardoor kan een beter onderscheid worden gemaakt tussen kinderen van wie de taalontwikkeling normaal verloopt en kinderen met een echte taalstoornis. Deze uitkomsten zijn praktisch van belang voor de vraag of het kind gewoon mee kan in het reguliere onderwijs (maar dan wel met meer hulp en begeleiding dan een regulier kind) of speciale taaltherapie nodig heeft. De betere diagnose zal naar verwachting leiden tot een efficiëntere verwijzing naar het special onderwijs en tot een vermindering van de kosten doordat alleen kinderen worden doorverwezen die die specialistische hulp ook inderdaad nodig hebben.

Studie 2: Peña, Quinn en Iglesias (1992)

Achtergrond en vraagstelling. *Dynamic assessment* kan worden gebruikt om de schoolprestaties van leerlingen te voorspellen. De vraag is dan in hoeverre *dynamic assessment* een aanvullende bijdrage levert gegeven het voorspellend vermogen van een intelligentietest of een statische vaardigheidstoets. Op basis van een meta-analyse rapporteert Resing (2006) een verhoging van de (gemiddelde) correlatie van .60 (alleen statisch) naar .75 (statisch plus dynamisch). Tegelijkertijd constateert zij echter dat het beeld van de predictieve validiteit vooralsnog diffuus is, omdat onderzoekers onvergelijkbare designs hanteren en de resultaten elkaar soms tegenspreken (zie ook de meta-analyses van Sternberg & Grigorenko, 2002; Swanson & Lussier, 2001; Caffrey, Fuchs & Fuchs, 2008). In een meta-analyse bestudeerden Caffrey et al. (2008) vijftien studies naar de predictieve validiteit van *dynamic assessment* in vergelijking met traditionele prestatiemetingen. Traditionele prestatiemetingen correleerden gemiddeld ongeveer even laag met leerprestaties als *dynamic assessment* (respectievelijk .41 en 49). In de *dynamic assessment* studies, waarin de leerlingen contingente feedback ontvingen, waren de correlaties met leerprestaties echter beduidend lager dan in de *dynamic assessment* studies met non-contingente feedback (gemiddeld respectievelijk .39 versus .56). Mogelijk is de standaardisatie van de interventie van belang voor het voorspellend vermogen. In een literatuurstudie vonden Caffrey et al. (2006) slechts vier bruikbare studies naar de toegevoegde waarde van

dynamic assessment voor het voorspellen van schoolprestaties op statische toetsen of tests. De resultaten waren teleurstellend. De unieke bijdrage van *dynamic assessment* bleek zeer beperkt. Opgemerkt moet worden dat taalstudies in de literatuurstudie van Caffrey et al. (2008) zwak vertegenwoordigd waren. Andere onderzoekers rapporteren meer positieve resultaten, ook op het gebied van taalvaardigheid. Met een combinatie van twee dynamische indicatoren – het aantal hints en het aantal items waarbij hulp nodig was – kon Resing (1993) bijvoorbeeld 14% extra variantie in taalprestaties voorspellen (nadat rekening gehouden was met verbaal IQ)..

Peña, Quinn en Iglesias (1992) onderzochten de bruikbaarheid van *dynamic assessment* voor het identificeren van leerlingen uit minderheidsgroepen met ernstige taalachterstand. *Dynamic assessment* wordt regelmatig gebruikt om een onderscheid te maken tussen zich normaal ontwikkelende leerlingen met een eventuele achterstand en leerlingen waarbij de leerproblemen zo ernstig zijn dat zij in speciale onderwijsprogramma's beter tot hun recht komen. De achterliggende gedachte is dat de lage, initiële leerresultaten van leerlingen niet te wijten zijn aan een leerstoornis maar aan culturele verschillen als zij op *dynamic assessment* een goede respons en hoge leerwinst laten zien (Peña, 1993). Met passende hulp en begeleiding zouden deze leerlingen op een gewone school beter op hun plaats zijn dan op een school voor speciaal onderwijs. Als een leerling echter minder positief leergedrag laat zien (bijvoorbeeld qua leerwerkhouding of metacognitieve vaardigheden) en tegelijkertijd weinig leerwinst boekt, dan is er waarschijnlijk sprake van een ernstige leerprobleem dat specialistische hulp behoeft. In vergelijking met *dynamic assessment*, zijn traditionele statische tests minder geschikt voor het plaatsen van leerlingen uit minderheidsgroepen in programma's of opleidingen voor speciaal onderwijs. Een van de redenen is dat lage prestaties op statische toetsen vaak ten onrechte aan een ernstige taalstoornis worden toegeschreven in plaats van aan een afwijkende culturele en taalkundige achtergrond. Screenings- en plaatsingstests voor jonge kinderen zijn vaak gebaseerd op zogeheten *labeling* taken waarbij de leerlingen een gepresenteerd object de bijbehorende naam moet geven. Kinderen uit minderheidsgroepen hebben echter vaak minder gelegenheid gehad om zich de vaardigheid in het labelen van objecten eigen te maken. Zo zijn er grote culturele verschillen in de manier waarop en de frequentie waarmee ouders met hun jonge kinderen praten, feitelijke vragen stellen en hen informatie laten herhalen (Heath, 1983, 1986). Ook zijn er verschillen in de mate waarin objecten met hun bijbehorende naam worden aangeduid. Kinderen uit bepaalde minderheidsgroepen zijn vooral meer geneigd om objecten aan de hand van hun functie te beschrijven in plaats van hun specifieke naam te noemen (Gutierrez-Clellen & Iglesias, 1987). Daardoor kunnen deze leerlingen onderprestaties laten zien op traditionele *labeling* taken, terwijl er geen sprake van een ernstige leerstoornis hoeft te zijn. De lage score weerspiegelt slechts een gebrek aan ervaring met de eisen van de toetstaak.

- ***Onderzoeksopzet.*** Aan het onderzoek namen vijftig drie- tot negenjarige kinderen van Puerto Ricaanse en Afrikaans-Amerikaanse komaf deel. Tijdens de voormeting werden twee statische landelijk genormeerde toetsen afgenomen: de *Expressive One-Word Picture Vocabulary Test* (EOWPVT) (Gardner, 1979) en de *Comprehension subtest* (CSSB) van de *Stanford-Binet Intelligence Scale* (Thorndike, Hagen, & Sattler, 1986). Op basis van klasobservaties, een articulatie- en hoorscreening en bij leraren en ouders werden gegevens verzameld en werden de 50 leerlingen ingedeeld in twee groepen: zonder taalachterstand en met taalachterstand of mogelijke taalstoornis. Na de interventie (*dynamic assessment)* werd de EOWPVT bij 48 leerlingen nogmaals afgenomen om hun woordenschat (*labelen*) te toetsen en aan het einde van schooljaar gebeurde dit voor de derde keer bij 27 leerlingen.

- ***Dynamic assessment.*** De onderzoekers ontwierpen twee korte *dynamic assessment* mediatiesessies van elk twintig minuten waarin het principe van het labelen centraal stond. Daarbij werd gebruik gemaakt van objecten, verhaaltjes, boeken en kaarten met afbeeldingen. De mediatie was erop gericht de leerlingen er bewust van te maken dat objecten kunnen worden aangeduid aan de hand van hun functie ('wat het doet' of 'wat je ermee doet'), hun categorie (bijvoorbeeld 'dieren' of 'voedsel') of hun label ('specifieke naam'). Het gebruik van metacognitieve strategieën - aandacht, zelfregulatie en het benutten van de volwassene als een mogelijke bron van informatie - werd geregistreerd met een aangepaste versie van de *Dynamic Assessment Recording Form* (Lidz, 1991). De modificeerbaarheid - gevoeligheid voor de instructie, de inspanning van de examinator en transfer - werd vastgesteld met de *Summary of Dynamic Assessment Results* (Lidz, 1991).

- ***Resultaten***. Beide groepen (met en zonder taalstoornis) scoorden ongeveer even laag op de EOWPVT-voormeting waarbij de leerlingen losse woordjes moesten *labelen*. Zoals verwacht bleken veel van de gemaakte fouten te wijten aan het gegeven dat leerlingen een beschrijving van de functie van de gepresenteerde objecten gaven in plaats van het specifieke label. Hieruit concluderen de onderzoekers dat deze statische toets niet geschikt is om een onderscheid te maken tussen leerlingen met en zonder een ernstige taalstoornis. De resultaten op de nameting lieten zien dat de leerresultaten van alle leerlingen vooruit

waren gegaan. De leerlingen met ernstige taalstoornis behaalden echter veel lagere scores op de nameting, bleken minder gevoelig voor de mediatie en vergden meer inspanning van de examinator dan 'normale' leerlingen. Bovendien bleek dat er met de scores op de EOWPVT-nameting een goed onderscheid gemaakt kon worden tussen beide groepen: 92% van de leerlingen werd correct geclassificeerd. Dit is opmerkelijk omdat beide groepen op de statische voormeting ongeveer even laag presteerden. Vrijwel alle leerlingen zonder taalstoornis bleken zich de vaardigheid in het benoemen of *labelen* van objecten eigen gemaakt te hebben.

- ***Conclusie.*** Peña et al. (1992) concluderen dat *dynamic assessment* een bruikbaar en beloftevol middel is om bij toetsing van woordenschat rekening te houden met het effect van taalachterstand. *Dynamic assessment* helpt te differentiëren tussen leerlingen met een taalachterstand als gevolg van een taalstoornis of door gebrek aan ervaring vanuit de (thuis)omgeving.

Studie 3: Ukrainetz, Harpell, Walsh en Coyle (2000)

- ***Achtergrond en vraagstelling.*** Ukrainetz, Harpell, Walsh en Coyle (2000) repliceerden het onderzoek van Peña et al. (1992) met een andere taak, bij een andere leeftijdsgroep en in een andere culturele context. Ook zij waren geïnteresseerd in de bruikbaarheid van *dynamic assessment* voor het vaststellen van het leerpotentieel van leerlingen uit minderheidsgroepen.

- ***Onderzoeksopzet.*** Zij onderzochten 23 Indiaanse kleuters van de Arapahoe en Shoshone stam. De leerlingen werden op basis van leerkrachtoordelen en klasobservaties ingedeeld in vijftien goede en acht zwakke taalleerders. Voor de voor- en nameting gebruikten de onderzoekers een gestandaardiseerde toets voor het meten van de receptieve en productieve woordenschat, de *expressive and receptive categorisation subtests* van de *Assessing Semantic Skills through Everyday Themes* (ASSET; Barret, Zachman & Huisingh, 1988).

- ***Dynamic assessment.*** De instructiecomponent van de *dynamic assessment* bestond uit twee mediatiesessies van ieder dertig minuten. De mediatie was gericht op het verhelderen van het idee van het groeperen van woordenschatitems (in plaats van de gebruikelijke directe instructie van specifieke woordenschatitems). Expliciete aandacht werd besteed aan het toekennen van een categorienaam aan een verzameling items (bijvoorbeeld fruit voor appels, peren en sinaasappelen). In elke sessie voerden de leerlingen twee activiteiten uit, zoals het groeperen van verschillende soorten voedsel uit een koelkast en het identificeren van dieren in een verzameling plaatjes. Om de mediatie enigszins te standaardiseren was er een script waarin het begin en einde van elke activiteit werd vastgelegd (Péna, 1993). De instructie werd gegeven volgens de twaalf mediatieprincipes van Lidz (1991). De getrouwheid van de implementatie volgens de twaalf mediatieprincipes werd gecontroleerd met de *Mediated Learning Experience Rating Scale* (Lidz, 1991) en bleek goed te zijn. Met de zogeheten *Modifiability Index* toetsten de onderzoekers de mate waarin de leerlingen positief leergedrag vertoonden en positief reageerden op de instructie. Dat gebeurde met twee observatielijsten over a) de leerstrategieën die leerlingen tijdens de mediatie toepasten en b) de gevoeligheid voor de mediatie of respons op de interventie. De observatielijst *Learning Strategies Checklist* (LSC; Peña, 1992) bevatte de aandachtspunten: aandacht, plannen, zelfregulatie, toepassing en motivatie. De observatielijst *Response to Mediation Checklist (RMC)* bevatte observatiepunten met betrekking tot a) de manier waarop de leerling op de interventie reageerde (begrijpt de leerling wat hem of haar wordt onderwezen?), b) de moeite die de examinator moest doen om de beoogde verandering te bewerkstelligen en c) een inschatting van de transfer over verschillende taken.

- ***Resultaten.*** De effectgrootte van het verschil tussen de scores van beide groepen op de *Modifiability Index* was groot en bedroeg .86. Daarbij discrimineerde de *Learning Strategies Checklist* beter tussen zwakke en sterke taalleerders dan de *Response to mediation Checklist*. Na de mediatie bleken de prestaties van alle leerlingen vooruitgegaan te zijn, maar de goede taalleerders gingen meer vooruit dan de zwakke taalleerders (1 *SD* versus .5 *SD*) en dit gold zowel voor het expressieve als receptieve gedeelte van de woordenschattoets. Na statistische correctie voor verschillen tussen beide groepen op de voormeting, bleek het verschil op de expressieve woordenschattaak echter gereduceerd tot een halve standaarddeviatie, terwijl het overeenkomstige gedeelte op het receptieve van deze test niet meer significant was.

- ***Conclusie.*** Een belangrijke conclusie betreft de gebleken geschiktheid van beknopte observatielijsten als middel voor het vaststellen van het leerpotentieel van kinderen uit minderheidsgroepen. Het verdient aanbeveling de bruikbaarheid van dit eenvoudige en tijdefficiënte middel voor andere doelgroepen, leergebieden en gebruikers (leraren) nader te onderzoeken.

Studie 4: Kapantzoglou, Adelaida Restrepo en Thompson (2010)

- ***Achtergrond en vraagstelling*.** Kapantzoglou et al. (2010) onderzochten de effectiviteit van een *dynamic assessment* bij het identificeren van tweetalige leerlingen met een primaire taalstoornis. Aanleiding was de eerder genoemde overdiagnostisering van taalstoornissen bij tweetalige leerlingen (die vaak minder gelegenheid hebben gehad om de taal te leren en over minder ervaring met testtaken beschikken). De onderzoeksvraag was in hoeverre een korte *dynamic assessment*, waarbij nieuwe woorden aan de hand van verbale en visuele ondersteuning werden aangeleerd, in staat is om een betrouwbaar onderscheid te maken tussen tweetalige leerlingen met en zonder taalstoornis.

- ***Onderzoeksopzet*.** De onderzoeksgroep bestond uit twee groepen van vooral Spaans sprekende leerlingen van vier of vijf jaar uit laag-SES gezinnen: 15 leerlingen met een normale taalontwikkeling en 13 met een taalachterstand of mogelijke taalstoornis. De indeling in beide groepen vond plaats op basis van het aantal grammaticale fouten per onderdeel op een mondelinge naverteltaak en informatie van ouders, leraren en een tweetalige logopedist. De *dynamic assessment* werd uitgevoerd volgens een voormeting-instructie-nameting ontwerp (zonder controlegroep).

- ***Dynamic assessment*.** In een korte sessie van 30 à 40 minuten onderwezen de onderzoekers drie non-woorden en drie bekende bestaande woorden. De non-woorden waren *fote*, *depa* en *kina* voor respectievelijk een onbekend dier (*an animal that could not be determined*), onbekend voedsel (*seeds*) en onbekend speelgoed (*a bubble level presented as a toy*). De doelwoorden werden onderwezen volgens de mediatieprincipes van Lidz (1991). Volgens een script werd elk woord negen keer gepresenteerd, inclusief feedback, bijvoorbeeld "Y*es, it's a depa*" bij een goed antwoord of "*It's a depa, what is it?*" bij een fout antwoord. Het script werd drie keer achter elkaar toegepast (fase 1, 2 en 3) zodat de leerling ieder woord in totaal 27 keer kreeg aangeboden. Aan het einde van elke fase werd de vaardigheid in het identificeren (receptief) en benoemen (productief) van de onbekende objecten gemeten. Hoe meer onbekende objecten de leerling correct kon identificeren en benoemen, hoe hoger de score. Daarnaast werd de modificeerbaarheid van de leerling vastgesteld met de *Learning Strategies Checklist* (LSC; Lidz, 1991; Peña, 1993) en de *Modifiability Scale* (MS; Lidz, 1987, 1991). De LSC meet de responsiviteit van de leerling (aandacht, planning, zelfregulatie en motivatie) en de MS meet de inspanning die de examinator zich moest getroosten en de hulp die de leerling tijdens de sessies nodig had.

- ***Resultaten*.** De onderzoekers voerden afzonderlijke discriminantanalyses uit voor fase 1, 2 en 3 (dit wil zeggen nadat de woorden 9, 18 en 27 keer gepresenteerd waren). Met deze analyse bepaalden ze in hoeverre het mogelijk was om op basis van de scores op woordidentificatie, woordbenoeming en LSC te voorspellen of de leerling tot de groep wel/geen taalstoornis behoorde. Vanwege de hoge correlatie met LSC werd MS uit de regressievergelijking verwijderd. Van de drie discriminantanalyses gaf alleen die voor fase 1 significantie te zien. Op basis van de scores voor woordidentificatie, woordbenoeming en modificeerbaarheid bleek het mogelijk om 79% van de leerlingen met en zonder taalstoornis correct te classificeren. De sensitiviteit (dit wil zeggen: het percentage correct geclassificeerde leerlingen met een taalstoornis) bedroeg 77% en de specificiteit (dit wil zeggen: het percentage correct geclassificeerde leerlingen zonder taalstoornis) was 80%.

- ***Conclusies*.** De gevonden percentages correcte classificaties zijn lager dan de ondergrens van 90% zoals voorgesteld door Plante en Vance (1994). Desalniettemin was dit resultaat voor de onderzoekers aanleiding om te spreken van een veelbelovend en efficiënt middel voor het screenen van taalstoornissen bij tweetalige leerlingen. De score op receptieve woordenschat (identificatie) en de score op de LSC bleken betere voorspellers van het al dan niet hebben van een taalstoornis dan de score op productieve woordenschat (benoemen). Een mogelijke verklaring is de hoge moeilijkheid van de woordbenoemingstaak. Het verdient aanbeveling na te gaan in hoeverre de accuraatheid van de voorspelling verhoogd kan worden door verbetering aan te brengen in de woordenschattaken. De LSC-observatielijst leverde een significante bijdrage aan het differentiëren van leerlingen met en zonder taalstoornis. Net als in eerder onderzoek bleek de mate waarin de leerling aandachtig, planmatig, zelf-gereguleerd en gemotiveerd leert van belang te zijn voor een accurate diagnose (e.g., Peña, 2000; Peña et al., 2001, 2006; Ukrainetz et al., 2000). Het verdient aanbeveling het gebruik van korte observatielijsten binnen *dynamic assessment* voor screeningsdoeleinden ook bij andere doelgroepen en taalvaardigheden nader te onderzoeken.

Studie 5: Larsen en Nippold (2007)*

- ***Achtergrond en vraagstelling.*** In deze studie (Larsen et al., 2007) is een *dynamic assessment* procedure onderzocht om meer zicht te krijgen op de (woordleer)strategie van morfologische analyse bij het achterhalen van woordbetekenissen. In het onderwijs komen leerlingen veel onbekende, nieuwe woorden tegen. Ze krijgen te maken met schooltaalwoorden en vakspecifieke woorden in de verschillende vakken. Het is voor een leerkracht of docent niet mogelijk om alle woorden expliciet in de klas uit te leggen. Leerlingen moeten dikwijls zelf achter de betekenis van een woord komen waarbij zij van verschillende strategieën gebruik kunnen maken, afhankelijk van de situatie. In het onderwijs wordt aan deze strategieën aandacht besteed. De lezer kan de betekenis bijvoorbeeld afleiden door gebruik te maken van de context, zoals zinnen en plaatjes. Of het woord kan worden opgezocht in een woordenboek. Een andere strategie die kan worden toegepast is morfologische analyse van woorden.

  Al vanaf de basisschool ontwikkelt zich een bewustzijn van morfemen. Leerlingen en adolescenten leren gebruik te maken van deze kennis om te bepalen wat onbekende woorden betekenen (Larsen & Nippold, 2007). Door het herkennen van het stamwoord (beest) en de voor- of achtervoegsels (-achtig) en hun betekenis, kan de leerling de betekenis van het totale woord vaststellen. De vaardigheid in het hanteren van deze woordleerstrategie breidt zich uit gedurende de schoolloopbaan van het basisonderwijs, middelbaar tot hoger onderwijs. Dit blijkt een belangrijke vaardigheid want het hangt samen met andere taalvaardigheden zoals leesbegrip.

  Manieren om morfologische analyse als strategie te toetsen zijn beperkt. Ook wordt kennis van woordbetekenissen meestal statisch getoetst, waarbij er geen zicht is op de strategieën die leerlingen toepassen. Een *dynamic assessment* kan meer inzicht geven in hoeverre een leerling gebruik kan maken van morfologische analyse om de betekenis van een woord vast te stellen (Palincsar, Brown & Campione, 1984). In tegenstelling tot een statische toets laat een *dynamic assessment* zien of de prestatie van een leerling verbetert als er feedback wordt gegeven (Swanson & Lussier, 2002). Zo kan worden bepaald over welke kennis de leerling al beschikt en waarop in het onderwijs of in een interventie moet worden aangesloten.

- ***Onderzoeksopzet.*** De onderzoeksgroep bestond uit 50 leerlingen in de leeftijd van 10 tot 12 jaar. De leerlingen hadden een normale taalontwikkeling. Er was alleen een voormeting waarbij de woordenschat en leesvaardigheid werden getoetst.

- ***Dynamic assessment.*** In hun studie beschrijven Larsen et al. (2007) een toetsprocedure om morfologische analyse dynamisch te toetsen. Deze taak *Dynamic Assessment Task of Morphological Analysis* ontwikkelden ze voor het onderzoek waarbij ze zich baseerden op een procedure zoals beschreven voor Anglins (1993). Leerlingen kregen vijftien afleidingen te zien. Deze woorden waren laag frequent maar de stam was hoogfrequent. Bijvoorbeeld het woord *puzzlement* als afgeleide van het hoogfrequente *puzzle*. Hoewel dit in het onderzoek niet werd nagegaan, werd aangenomen dat de afgeleide woordvormen niet gekend werden maar de stam van de woorden wel. Met de taak werd geprobeerd morfologische analyse bij leerlingen uit te lokken als strategie om achter de betekenis van de woorden te komen. Leerlingen kregen één-op-één de instructie om te luisteren en te kijken naar een woord om dan vervolgens te zeggen wat het woord betekende. Als het woord moeilijk was, kreeg de leerling een aanwijzing. De aanwijzingen waren zo opgesteld dat ze in toenemende mate hulp boden bij het achterhalen van de woordbetekenis (zie Tabel 3.2). Bij deze opbouw is ervan uitgegaan, dat de leerder eerst de afzonderlijke morfemen identificeert en er dan betekenis aan koppelt, daarna bepaalt hij hoe deze morfemen de betekenis van de stam veranderen en ten slotte voegt hij de informatie samen om de betekenis van het woord te bepalen. Het aantal aanwijzingen dat gegeven werd, was afhankelijk van het antwoord. Als de leerling de betekenis van het afgeleide woord goed omschreef, werd gevraagd hoe hij dit wist. Als hij daarbij verwees naar de afzonderlijke morfemen werd doorgegaan naar een volgend woord. Zo niet, werd erop gewezen dat het woord uit een paar kleinere delen bestond en of de leerling kon uitleggen wat die kleinere delen betekenden. Als de leerling de betekenis van het woord niet goed omschreef, werd er meteen op gewezen dat het woord uit een paar kleinere delen bestond en of de leerling kon uitleggen wat die kleinere delen betekenden. Als de leerling hierop geen goed antwoord kon geven werd een volgende aanwijzing gegeven. Bij elke aanwijzing gold: als het antwoord hierna goed was, werd een volgend woord genoemd; als het antwoord niet juist was, werd een volgende aanwijzing gegeven.

Tabel 3.2 Overzicht van aanwijzingen met opbouw.

Niveau	Aanwijzingen
0	Wat is de betekenis van het woord? En hoe weet je dit?
1	Bestaat het woord feestelijk uit kleinere delen? Welke kleinere delen?
2	Het woord feestelijk bestaat uit feest en elijk. Kun je me nu vertellen wat het woord betekent?
3	Luister naar de zin en vertel me dan wat het woord feestelijk betekent. “Anna versierde de kamer met slingers en ballonnen. Het zag er heel feestelijk uit!”
4	Welk van deze keuzes past bij de betekenis van feestelijk, waarna er 3 opties genoemd worden waaruit gekozen moet worden.

De resultaten laten zien dat er variatie tussen leerlingen bestaat in de wijze waarop ze morfologische analyse gebruiken als een strategie om de betekenis van woorden af te leiden. Zo bleek dat sommige leerlingen wel het stamwoord herkenden maar met de voor/achtervoegsels geen raad wisten. En er waren ook leerlingen die het (hoogfrequente) stamwoord niet kenden. Tot slot bleken de scores op de *dynamic assessment*-taak voor morfologische analyse positief gecorreleerd met scores op statische toetsen voor woordkennis (r = .36) en leesvaardigheid (r = .50).

- ***Conclusie.*** De onderzoekers concluderen dat de nieuwe *dynamic assessment* procedure meer zicht kan geven op de instructie die leerlingen nodig hebben bij het leren van de betekenis van weinig voorkomende woorden met behulp van morfologische analyse.

3.2.2 Mondelinge taalvaardigheid

Het literatuuronderzoek heeft twee studies naar de effectiviteit of bruikbaarheid van *dynamic assessment* voor het meten en ontwikkelen van mondelinge taalvaardigheid opgeleverd: Peña, Gillam, Malek, Ruiz-Felter, Resendiz, Fiestas en Sabel (2006) en Kramer, Mallett, Schneider en Hayward (2009). Beide studies kenden een voor- en nameting, maar alleen die van Peña et al. (2006) bevatte een experimentele en controlegroep. Beide studies richtten zich op het aspect vertelvaardigheid als onderdeel van mondelinge communicatie. Peña et al. (2006) onderzochten de effectiviteit van een dynamic assessment op het gebied van vertelvaardigheid voor het screenen van basisschoolleerlingen met een taalstoornis. Kramer et al. (2009) deden een soortgelijk onderzoek bij basisschoolleerlingen van Indiaanse afkomst met een *dynamic assessment* voor vertelvaardigheid.

Studie 1: Peña, Gillam, Malek, Ruiz-Felter, Resendiz, Fiestas en Sabel (2006)

- ***Achtergrond en vraagstelling.*** Het kunnen vertellen van een volledig en goed gestructureerd verhaal is van groot belang voor het communiceren in de thuissituatie en het leren op school. Peña et al. (2006) onderzochten de effectiviteit van *dynamic assessment* in het differentiëren tussen leerlingen uit minderheidsgroepen met een normale taalontwikkeling en een taalstoornis. Aan het onderzoek van Peña et al. (2006) namen 71 leerlingen uit de eerste en tweede klas deel. De leerlingen waren verdeeld over drie groepen: a) 27 leerlingen met een normale taalontwikkeling, b) 14 leerlingen met een verstoorde taalontwikkeling en c) een controlegroep van 30 leerlingen met een normale taalontwikkeling. De indeling in leerlingen met een normale en verstoorde taalontwikkeling vond plaats op basis van scores op gestandaardiseerde taaltoetsen, klasobservaties en oordelen van leraren, ouders en een gecertificeerde logopedisten.

- ***Onderzoeksopzet***. Aan het onderzoek namen 58 leerlingen uit het eerste en tweede leerjaar deel. De leerlingen hadden een diverse achtergrond. Het onderzoek had een voormeting-instructie-nameting ontwerp waaraan een controlegroep was toegevoegd. Op de voor- en nameting werd de vertelvaardigheid van de leerlingen beoordeeld op tien beoordelingsaspecten, onderverdeeld in de domeinen verhaalcomponenten (Setting: tijd en plaats, informatie over de personen en temporele en causale relaties), verhaalideeën en taal (complexiteit van ideeën, complexiteit van vocabulaire, kennis van dialogen, grammatica en creativiteit) en structuur (combinaties van verschillende grammaticale elementen op tekstniveau). Daarnaast werden productiviteitsmaten verzameld, zoals het aantal woorden, het aantal verschillende woorden, het aantal hoofdzinnen, de gemiddelde woordlengte van de hoofdzinnen. Bij de leerlingen die deelnamen aan de interventie werden modificatiescores verzameld met betrekking tot de gevoeligheid voor de instructie en de

moeite die de examinator moest doen en de hoeveelheid begeleiding die hij of zij moest geven (Peña, 2000; Peña et al., 2006).

- ***Dynamic assessment.*** De *dynamic assessment* bestond uit 3 fasen. In de eerste fase vertelden de leerlingen een verhaal bij een beeldverhaal zonder tekst. Dit was de voormeting. Hierna volgde fase 2 van twee individuele *dynamic assessment* sessies van elk 30 minuten gericht op het verbeteren van de vaardigheid en het gebruik van strategieën bij het navertellen van beeldverhalen (zonder tekst). De sessies hadden tot doel de lengte en complexiteit van de vertelde verhalen te verhogen. Tijdens de eerste sessie besprak de examinator eerst hoe de leerling het verhaal op de voormeting had verteld. De examinator las het verhaal van de leerling voor. Vervolgens werden besproken: de setting (tijd, plaats, informatie over de personen en de temporele en causale relaties tussen de gebeurtenissen) en de structuur van het verhaal, waarbij de examinator voorbeelden uit het verhaal van de leerling gebruikte. In de tweede sessie nam de examinator het proces van het navertellen van een woordloos beeldverhaal met de leerling door. De leerling vertelde een verhaal bij een nieuw beeldverhaal. Wederom werd daarbij aandacht besteed aan de setting van het verhaal en de structuur. Daarbij gebruikte de examinator onder meer poppen en foto's met achtergrondinformatie (o.a. bergen, bos) om te laten zien hoe je een compleet verhaal kunt vertellen. Daarnaast besteedde de examinator onder meer aandacht aan het doel van de mediatie, het belang van het kunnen vertellen van complete en gestructureerde verhalen voor thuis en op school en hoe de leerling elementen van de setting en de structuur in zijn of haar verhalen kon verwerken (zie onderstaande passage uit van het script voor de examinator):

  Stories need to tell us when and where something happened because that helps us understand the world the character lives in. So, what do we need to think about [when and where or setting]? [Use background sheet to illustrate setting, and then compare with hook]. [Refer top. 1 in *Two Friends]* How does this story start? [Pause, wait for response.] Where do you think they are? [Pause, wait for response.] What time do you think it is? [Pause, wait for response]. So, to say where and when, you could say ... [pause, let them fill in, if they don't, give example "one morning the dog and cat stood by the river"]. That tells us when and where.

- ***Resultaten.*** Na afloop van de twee mediatiesessies vertelden de leerlingen meer complete en complexere verhalen dan ervoor. Het effect van de mediatie verschilde voor leerlingen met en zonder een taalstoornis. Zich normaal ontwikkelende leerlingen behaalden een grotere leerwinst en hogere modificeerbaarheidsscores dan leerlingen met een taalstoornis. Op basis van de statische voormeting van de vaardigheid in het vertellen van verhalen bleek het niet mogelijk om de leerlingen met en zonder taalstoornis accuraat te classificeren. Gemiddeld over de verschillende beoordelingsaspecten bedroeg de sensitiviteit 26% en de specificiteit 88%. Hier gaat een hoge specificiteit dus ten koste van de sensitiviteit. De nameting van vertelvaardigheid discrimineerde beter tussen leerlingen met en zonder taalstoornis dan de voormeting: de sensitiviteit was gemiddeld 64% en de specificiteit gemiddeld 88%. Volgens de vuistregel van Plante en Vance (1994) kan dit resultaat als redelijk worden beoordeeld. De modificeerbaarheidsscores die met de observatielijst verkregen waren, gaven het beste classificatieresultaat te zien. Zich normaal ontwikkelende leerlingen behaalden hogere modificeerbaarheidsscores (M = 7.93, SD = 1.49) dan leerlingen met een taalstoornis (*M* = 3.40, *SD* = 1.24); de sensitiviteit was 93% en de specificiteit 82%. Op basis van de nameting en de modificatiescores tezamen konden alle leerlingen correct geclassificeerd worden.

- ***Conclusies.*** Aangezien op de voormeting de hoge specificiteit ten koste gaat van de sensitiviteit trekken de onderzoekers de conclusie dat een eenmalige afname van een statische toets niet geschikt is voor alle leerlingen, en dan vooral niet voor de leerlingen die een andere of in mindere mate ervaring hebben met verteltaken. De resultaten laten zien dat het mogelijk is de mondelinge vaardigheid van zwak presterende leerlingen met een normale taalontwikkeling met een korte interventie aanzienlijk te verbeteren (terwijl degenen met een leerstoornis daar veel minder van profiteren). *Dynamic assessment* maakt het mogelijk om een beter onderscheid te maken tussen leerlingen uit minderheidsgroepen met een normale taalontwikkeling en leerlingen met een taalstoornis. De onderzoekers concluderen dat de *dynamic assessment* een minder partijdige maat voor vertelvaardigheid is dan een statische toets (zoals de voormeting) omdat het informatie verschaft over de denkprocessen, de zich ontwikkelende vaardigheid en het leerpotentieel van leerlingen. *Dynamic assessment* doet meer recht aan de mogelijkheden die de leerlingen hebben.

Studie 2: Kramer, Mallett, Schneider en Hayward (2009)

- ***Achtergrond en vraagstelling.*** Voortbouwend op de studies van Peña et al. (1992) en Ukrainetz et al. (2000) onderzochten Kramer et al. (2009) de diagnostische accuratesse van een *dynamic assessment* voor vertelvaardigheid (Miller, Gillam, & Peña, 2001) in het differentiëren tussen leerlingen met een normale taalontwikkeling en leerlingen met mogelijke taalstoornis.

- ***Onderzoeksopzet.*** De onderzoeksgroep betrof 17 Indiaanse leerlingen uit leerjaar 3, woonachtig in het Samson Cree reservaat in het Canadese Alberta. Volgens opgave van schoolpersoneel kenden twaalf leerlingen een normale taalontwikkeling en hadden er vijf een taalstoornis. De vertelvaardigheid van de leerlingen werd vastgesteld met behulp van twee tekstloze prentenboeken. De *dynamic assessment* werd onderzocht met een voormeting-instructie-nameting ontwerp. De kwaliteit van de geproduceerde verhalen tijdens de voor- en nameting beoordeelden Kramer et al. (2009) met een *dynamic assessment* toets. Verschillende aspecten werden beoordeeld, zoals het aantal en de kwaliteit van verhaalcomponenten (setting, informatie over verhaalpersonen, volgorde van de gebeurtenissen en oorzaak-gevolg relaties), verhaalideeën en taalgebruik op tekstniveau (complexiteit van ideeën, complexiteit van woordenschat, grammaticale complexiteit, dialogen en creativiteit) en kenmerken van verhaalpassages, verhaalstructuur (begin van de gebeurtenis, ontwikkelingen in het verhaal, gevolgen, de afloop of het einde). Met een *dynamic assessment* observatielijst werd informatie verzameld over de moeite die de examinator moest doen en de modificeerbaarheid van de leerling tijdens de instructiecomponent.

  Dynamic assessment. Ieder kind nam deel aan twee korte mediatiesessies. De handleiding bevatte gestructureerde aanwijzingen voor de mediatie van vier verhaalcomponenten, te weten de setting, informatie over de verhaalpersonen, temporele relaties tussen gebeurtenissen en oorzaak-gevolg relaties. In de mediatiesessie kwamen twee verhaalcomponenten aan bod: een zeer slecht beheerste verhaalcomponent (score 1 of 2 op een schaal tot en met 5) en een min of meer beheerste verhaalcomponent (score 3 of 4). Tijdens een testfase werd vastgesteld welke verhaalcomponenten zeer slecht of min of meer beheerst werden. Tijdens de mediatie gaf de examinator expliciete aandacht aan het doel van de mediatie en belang van de desbetreffende verhaalcomponent voor het vertellen van het verhaal (aan de hand van voorbeelden uit een voorbeeldverhaal). Ter illustratie van mediatie waarbij informatie over de setting wordt gegeven, geven de onderzoekers het volgende voorbeeld:

  "To mediate setting information, the examiner and child would look at the storybook and collaborate on the story setting by coming up with words and phrases about when and where the animals are at the beginning of the story" (p. 123).

  In de volgende fase kreeg de leerling informatie over hoe hij of zij het geleerde kon toepassen bij het vertellen van andere, niet getoetste verhalen op een later moment. Tot slot was er een transferfase die gericht was op het samenvatten van het geleerde en het ontwikkelen van strategieën om het geleerde beter te onthouden. De tweede sessie volgde één dag na de eerste sessie en verliep op dezelfde manier als de eerste sessie met dien verstande dat nu de tweede verhaalcomponent onder de loep genomen werd. Ongeveer tien dagen na de twee mediatiesessies volgende een nameting aan de hand van een ander prentenboek zonder tekst.

- ***Resultaten.*** Op de voormeting was er geen noemenswaardig verschil in de kwaliteit van de verhalen tussen de leerlingen zonder en met een taalstoornis. Na de mediatie bleken alle leerlingen vooruit te zijn gegaan, maar voor de eerste groep was de leerwinst veel groter dan voor de tweede groep. Dit betekent dat leerlingen zonder taalleerstoornis meer van de mediatie profiteerden dan degenen met een dergelijke stoornis. Voor de beide gemedieerde verhaalcomponenten was het verschil tussen beide groepen groter dan voor de beide niet-gemedieerde componenten. Voor de beide behandelde componenten was het verschil precies één standaarddeviatie (Cohen's d = 1.00), terwijl het overeenkomstige verschil voor de beide niet-gemedieerde componenten zestiende standaarddeviatie bedroeg (Cohen's d = .60). Dit laatste verschil is opmerkelijk omdat hier volgens de onderzoekers sprake zou zijn van transfer van de behandelde naar de niet-behandelde verhaalcomponenten. Ook op de *dynamic assessment* observatielijsten over de modificeerbaarheid van de leerlingen en de moeite die de leerkracht zich moest getroosten was het verschil tussen beide groepen zeer groot (Cohen's d bedroeg respectievelijk 1.38 en 1.45). Leerlingen met een normale taalontwikkeling bleken dus gevoeliger voor mediatie en hadden minder hulp van de examinator nodig dan leerlingen met een taalstoornis.

  De accuratesse van de *dynamic assessment* in het voorspellen van wel/geen taalstoornis werd geanalyseerd met behulp van discriminantanalyse. Op basis van de vier instrumenten bleek de sensitiviteit 100% (alle vijf leerlingen met een taalstoornis werden correct geclassificeerd) en de specificiteit 92% (slechts één van de

vijftien leerlingen met een normale taalontwikkeling werd ten onrechte als een kind met een taalstoornis geclassificeerd). De onderzoekers gingen ook na in hoeverre een gereduceerd instrumentarium tot vergelijkbare resultaten zou leiden. Weglating van de beide *dynamic assessment* observatielijsten resulteerde in exact dezelfde classificatieresultaten (dit wil zeggen: sensitiviteit 100% en specificiteit 92%). Weglating van de *dynamic assessment* toetsing over de behandelde en niet-behandelde verhaalcomponenten resulteerde eveneens in 100% sensitiviteit, maar de specificiteit liep terug van 92% naar 82%. Dit laatste percentage correcte classificaties is slechts marginaal lager dan de ondergrens van 90% zoals voorgesteld door Plante en Vance (1994).

- ***Conclusie.*** De onderzoekers concluderen dat de *dynamic assessment* observatielijst een relatief efficiënt diagnostisch middel lijkt voor het screenen van taalleerstoornissen. De afname van een onderzoeksgerichte *dynamic assessment* kost weliswaar weinig tijd, maar de beoordeling van de prestaties van de leerlingen kan even tijdrovend zijn als bij een statische toets het geval is. Replicatie en kruisvalidatie van dit beschreven onderzoek met grotere steekproeven zou kunnen uitwijzen hoe groot de toegevoegde waarde is van relatief arbeidsintensieve *dynamic assessment* toetsen in vergelijking met de efficiëntere *dynamic assessment* beoordelingslijsten.

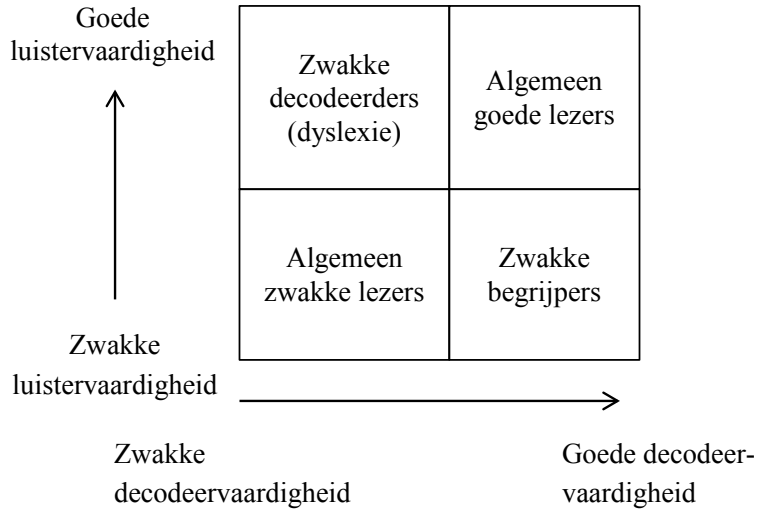
3.2.3 Leesvaardigheid

Het literatuuronderzoek heeft twee *dynamic assessment* studies op het gebied van leesvaardigheid opgeleverd, waarvan één op het gebied van leesbegrip en één op het gebied van decodeervaardigheid. Elleman, Compton, Fuchs, Fuchs en Bouton (2011) onderzochten de betrouwbaarheid en bruikbaarheid van een *dynamic assessment* op het gebied van leesbegrip voor het screenen van leerlingen die het risico lopen op ernstige leesproblemen. Fuchs, Compton, Fuchs, Bouton en Caffrey (2011) onderzochten de unieke bijdrage van een kortdurende *dynamic assessment* op het gebied van decodeervaardigheid aan voor het voorspellen van leesproblemen bij beginnende lezers. De studie van Fuchs et al. (2011) kende twee meetmomenten, maar vanwege het ontbreken van instrumentele overlap tussen de eerste en tweede meting is er strikt genomen geen sprake van een voor- en nameting. Omdat er evenmin sprake was van een controlegroep, voldoet deze studie aan geen van onze twee methodologische zoekcriteria. Vandaar dat wij de studie van een asterisk hebben voorzien. We hebben deze studie toch opgenomen om een relevant voorbeeld te kunnen geven van een *dynamic assessment* van decodeervaardigheid (of technisch lezen).

Studie 1: Elleman, Compton, Fuchs, Fuchs en Bouton (2011)

- ***Achtergrond en vraagstelling.*** Volgens de *simple view of reading* van Gough en Tunmer (1986) is leesbegrip het product van twee dimensies, namelijk decoderen en begrijpen. Figuur 3.1 geeft een schematische weergave van het model. We zien dat er drie typen zwakke lezers te onderscheiden zijn. Linksboven staan de zogeheten *zwakke decodeerders* of dyslectische leerlingen. Deze leerlingen hebben zeer veel moeite met het decoderen van woorden, maar scoren relatief goed bij begrijpend luisteren. Linksonder staan de algemeen zwakke lezers. Zij hebben problemen met zowel decoderen als begrijpend luisteren. Ten slotte staan rechtsonder de zogeheten *zwakke begrijpers*. Zij zijn goed in staat om losse woorden te decoderen, maar hebben grote moeite met begrijpend luisteren. Er wordt geschat dat 3 tot 10 procent van de schoolgaande leerlingen een zwakke begrijper is (Catts & Compton, 2012; Leach, Scarborough, & Rescorla, 2003). Omdat zwakke begripsvaardigheden in veel gevallen leiden tot leesproblemen, is een vroegtijdige signalering en behandeling van groot belang. In de praktijk worden de zwakke begrijpers echter vaak over het hoofd gezien, zeker in de eerste jaren van het basisonderwijs. Volgens Elleman et al. (2011) komt dit doordat toetsen voor aanvankelijk lezen de nadruk leggen op decodeervaardigheid. De toetsen zijn onvoldoende sensitief om zwakke begripsvaardigheden te detecteren en leiden daardoor tot een vertraagde identificatie.

Figuur 3.1 Typen lezers volgens 'The Simple View of Reading'



Bij jonge, minder vaardige, lezers wordt begrijpend luisteren vaak als proxy gebruikt voor leesbegrip (e.g., Catts, Adlof & Weismer, 2006). Hoewel een luistertoets in eerste instantie geschikt lijkt om leesbegrip onafhankelijk van decodeervaardigheid te meten, blijken luistertoetsen vaak niet adequaat te differentiëren tussen leerlingen mét en leerlingen zónder leesproblemen (e.g., Catts et al., 2006; Compton et al., 2008). Daarom stellen Elleman et al. (2011) voor om de begripsvaardigheden van leerlingen via *dynamic assessment* in kaart te brengen. Voor een accurate identificatie van de zwakke begrijpers in een klas moet een toets volgens Elleman et al. (2011) meer zijn dan een alleen een evaluatie achteraf. De toets zou ook inzicht moeten geven in het leerpotentieel van leerlingen. Elleman et al. (2011) gaan na in hoeverre leerlingen in staat zijn om impliciete informatie uit een tekst te abstraheren.

- ***Onderzoeksopzet.*** Het onderzoek had een voor- instructie-nameting ontwerp zonder een controlegroep. Er namen 100 leerlingen uit groep 4 deel aan het onderzoek. Tijdens de voor- en nameting werden verschillende statische toetsen afgenomen om de decodeervaardigheid, leesbegrip en woordenschat te toetsen.

- ***Dynamic assessment.*** Elleman et al. (2011) hebben gezocht naar een toets die (a) begripsvaardigheden kan voorspellen, (b) differentieert tussen goede en zwakke begrijpers, (c) zo min mogelijk verweven is met eerdere leerervaringen, en (d) opdrachten bevat die leerlingen na enkele keren oefenen onder de knie kunnen hebben. Ze hebben in totaal 7 teksten en 21 (7×3) inferentievragen geconstrueerd. De lengte van de verschillende teksten varieerde van 160 tot 217 woorden. De teksten waren verhalend of informerend van aard, sloten aan bij de belevingswereld van 7-jarigen, en bevatten uitsluitend impliciete informatie. De setting van het verhaal werd bijvoorbeeld nooit expliciet in de tekst vermeld. Hierdoor werden leerlingen verplicht om afleidingen (of inferenties) te maken van een tekst. Er is voor gekozen om leerlingen voornamelijk oorzakelijke verbanden tussen tekstdelen te laten leggen, ofwel *causale inferenties* te laten maken, omdat dit type inferentie (a) past bij de vaardigheden van jonge leerlingen, (b) een prominente rol speelt in verhalende teksten, en (c) gemakkelijker te maken is dan andere typen inferenties. De teksten werden hardop aan leerlingen voorgelezen. De eerste van de drie inferentievragen die per tekst gesteld werd, was steeds vrij gemakkelijk. Op deze manier was het mogelijk om ook iets te zeggen over de prestaties van zeer zwakke leerlingen. De tweede en derde inferentievraag waren moelijker. Leerlingen moesten in dat geval aanwijzingen combineren die in de tekst dichtbij elkaar, of juist ver uit elkaar, stonden (zie ook Ackerman, Jackson & Sherill, 1991; Ehrlich, Remond & Tardieu, 1999). De zeven teksten werden afgenomen volgens het schema zoals weergegeven in figuur 3.2.

Figuur 3.2. Opzet *assessment for learning* Elleman et al. (2011)

Verhaal 1	Verhaal voor training	Verhaal 2	Verhaal 3	Verhaal 4	Verhaal 5	Verhaal 6	Verhaal 7
Voormeting Geen feedback	Detective training	Dynamisch, feedback met hints	Dynamisch, feedback met hints	Dynamisch, feedback met hints	Nameting Geen feedback	Transfer, geringe samenhang Geen feedback	Transfer, informatief Geen feedback

In Figuur 3.2 is te zien dat leerlingen voorafgaand aan de *dynamic assessment* een korte training tot "reading detective" kregen. Aan leerlingen werd verteld hoe ervaren lezers de meest relevante informatie in een tekst opsporen en hoe zij die informatie omvormen tot een correcte en bondige mentale representatie van de inhoud. Na de training startte het dynamische deel van de assessment. In deze fase maken leerlingen inferenties van een tekst en krijgen zij feedback. De feedback werd gegeven in de vorm van gestandaardiseerde, steeds explicieter wordende, hints (i.e., *graduated prompts model* van Campione en Brown, 1987). De meeste hints herinnerden leerlingen eraan hoe een "reading detective" te werk gaat en leidden de leerlingen naar relevante passages in de tekst. Bijvoorbeeld: "Let's be reading detectives and use the clues to help us figure out where they are. Here the story says, "she ran down the long aisle." of "Here are some more clues. The story says *cereal aisle* and it says *cereal boxes*." De laatste hint in elke serie omvatte een samenvatting van alle eerdere hints. Het geven van hints ging door totdat de leerling een correct antwoord gaf of alle hints gegeven waren. Op basis van het aantal hints werd een schatting gemaakt van het leerpotentieel. Na het dynamische deel van de assessment volgde een nameting en werd de transfer naar andere teksten (geringere interne samenhang of informatief in plaats van verhalend) in kaart gebracht.

- ***Resultaten.*** Elleman et al. (2011) hebben onderzocht in hoeverre de ontwikkelde *dynamic assessment* in groep 4 van het basisonderwijs (a) betrouwbare toetsresultaten oplevert, (b) samenhangt met statische, reeds gevalideerde, toetsen, en (c) specifieke variantie kan verklaren bij begrijpend lezen. De resultaten waren over het algemeen positief. De betrouwbaarheid van de *dynamic assessment* was met .76 als voldoende aan te merken, de toetsresultaten van leerlingen bleken significant samen te hangen met de prestaties op een traditionele toets voor begrijpend lezen, en het aantal hints bleek, zoals verwacht, negatief samen te hangen met de begrijpend leesprestaties. De *dynamic assessment* verklaarde, naast woordenschat en decodeervaardigheid, namelijk slechts 4 procent extra unieke variantie bij begrijpend lezen (het totale percentage verklaarde variantie steeg van 74% naar 78%).

- ***Conclusie.*** De *dynamic assessment* van Elleman et al. (2011) lijkt een valide representatie te zijn van de begripsvaardigheden van leerlingen, en met de toepassing van deze procedure zijn we mogelijk beter in staat zijn om intra-individuele verschillen in de leesvaardigheid van jonge leerlingen te detecteren dan bij toepassing van een traditionele begrijpend lees- of luistertoets. De voorspellende waarde van de *dynamic assessment* viel niettemin tegen. Daarmee is het de vraag of de extra toetstijd die de afname van de *dynamic assessment* van Elleman et al. (2011) met zich meebrengt wel opweegt tegen de extra informatie die we verkrijgen in vergelijking tot de traditionele begripstoetsen.

Studie 2: Fuchs, Compton, Fuchs, Bouton & Caffrey (2011)*

- ***Achtergrond en vraagstelling.*** Aanleiding tot deze studie is de ontevredenheid over de tijdrovende en bewerkelijke testbatterijen die normaliter worden gebruikt voor het meten van het leerpotentieel van leerlingen in de Verenigde Staten. De vraag was in hoeverre een korte *dynamic assessment* voor decodeervaardigheid met deze testbatterijen kan concurreren in het voorspellen van diverse effectmaten van beginnende leesvaardigheid. Om een antwoord op deze vraag te geven onderzochten Fuchs et al. (2011) de voorspellende validiteit van een *dynamic assessment* van decodeervaardigheid in het geval van directe woordherkenning en tekstbegrip.

- ***Onderzoeksopzet.*** Aan het onderzoek namen 318 leerlingen uit het eerste leerjaar deel. De *dynamic assessment* was een onderdeel van een testbatterij met diverse voorspellers op het gebied van onder meer alfabetische kennis (letters en grafemen), benoemsnelheid, fonemisch bewustzijn, woordenschat en luistervaardigheid. Daarnaast was er een observatielijst voor het vaststellen van aandachtig, hyperactief en impulsief gedrag en een IQ-test. De te voorspellen taalvaardigheden waren directe woordherkenning en tekstbegrip die bij de nameting werden getoetst. De voorspellers werden gemeten in de herfst van het

schooljaar en de effectmaten in het voorjaar. Omdat er bij de tweede meting volledig andere toetsen werden afgenomen dan bij de eerste meting, is er strikt genomen geen sprake van een ontwerp met een voor- en nameting (waarmee leerwinst kan worden vastgesteld); ook ontbrak er een controlegroep die geen *dynamic assessment* kreeg.

- ***Dynamic assessment.*** De *dynamic assessment* procedure is voor een eerdere studie ontwikkeld (Fuchs et al., 2007). De *dynamic assessment* bestond uit één sessie waarin het decoderen op gestandaardiseerde wijze getoetst werd aan de hand van pseudowoorden van oplopende moeilijkheidsgraad. Het ging om MKM-woorden (aangeleerd als een linguïstische woordfamilie), MKMe (in Engels: woorden als *come*), en MKM(M)ing (Engels: *coming*, *tapping*, etc.). Binnen elk niveau van deze drie decodeervaardigheden waren er vijf vormen van meer en meer expliciete instructie mogelijk. Er werden steeds zes pseudowoorden na een instuctie-item getoond. Als een leerling vijf van de zes woorden goed las, werd doorgegaan met een volgend niveau (in decodeervaardigheid). Als de leerling minder dan vijf woorden goed las, kreeg hij meer expliciete instructie. Als de leerling na vijf stappen van toenemende expliciete instructie de woorden nog steeds niet goed las, werd de sessie beëindigd.

- ***Resultaten.*** De correlaties tussen enerzijds de *dynamic assessment* voor decodeervaardigheid en de drie effectmaten varieerden van .61 tot .72. In het geval van alfabetische kennis waren de correlaties met de effectmaten nog iets hoger (van .74 tot .81). De unieke bijdrage van *dynamic assessment* aan het voorspellen van directe woordherkenning en leesvaardigheid was klein en bedroeg hooguit 2% van de totale variantie.

- ***Conclusie.*** Hoewel de onderzoekers spreken van een veelbelovend resultaat, toonden zij zich niet tevreden. Naar hun mening gedroeg de *dynamic assessment* zich te veel als een traditionele statische toets. Net als veel statische toetsen bleek de *dynamic assessment* een bodemeffect te vertonen (waarbij te weinig laag presterende leerlingen zich de decodeervaardigheid eigen hadden gemaakt). De auteurs houden een pleidooi voor het gebruik van observatielijsten als het erom gaat het leerpotentieel van de leerling vast te stellen. Zij beschouwen de gebruikte observatielijst voor het meten van aandacht, hyperactiviteit en impulsiviteit zelfs als een volwaardige concurrent van de door hen gebruikte *dynamic assessment* toets voor decodeervaardigheid.

3.3 Studies met geautomatiseerde feedback

Het literatuuronderzoek heeft vijf studies naar de effectiviteit of bruikbaarheid van taalvaardigheidstoetsing met geautomatiseerde feedback opgeleverd. Van de vijf genoemde studies voldeed alleen het onderzoek van Franzke et al. (2005) volledig aan onze twee methodologische zoekcriteria (zowel een voor- en nameting als een experimentele en controlegroep). Vandaar dat we de beschrijving van de vijf studies met Franzke et al. (2005) beginnen. Het onderzoek van Ferster et al. (2012) kende wel een voor- en nameting maar geen controlegroep. De overige drie studies werden bij één groep op één moment uitgevoerd en bezitten dus de minste bewijskracht.

3.3.1 Schrijfvaardigheid

Franzke et al. (2005) onderzochten de effectiviteit van gecomputeriseerde feedback op het gebied van schrijfvaardigheid. Eveneens binnen het domein van schrijfvaardigheid stelden Ferster, Hammond, Alexander en Lyman (2012) de vraag in hoeverre computer-gegenereerde feedback vergelijkbaar is met menselijke feedback en in hoeverre die feedback bruikbaar is voor leraren en leerlingen.

Studie 1: Franzke, Kintsch, Caccamise, Johnson en Dooley (2005)

- ***Achtergrond en vraagstelling.*** Bij traditionele toetsen voor lees- en schrijfvaardigheid spelen drie problemen (e.g. Landauer, Lochbaum & Dooley, 2009). Ten eerste worden in de traditionele toetsen over het algemeen andere dingen van leerlingen gevraagd dan in het dagelijks leven (gebrekkige *face validity*). Ten tweede vormen de opgaven in de toetsen slechts een zeer beperkte steekproef van wat leerlingen allemaal moeten weten. Er is een risico dat leraren alleen de leerstof behandelen die in de toets aan de orde komt (*teaching to the test*). Ten derde dragen de traditionele toetsen nauwelijks bij aan het leren van nieuwe (vervolg)vaardigheden. Ze vertellen de leerkracht hoeveel een leerling weet en welke leerstofgebieden extra aandacht behoeven, maar leerlingen leren door het maken van de toets in principe geen dingen die ze nog niet kennen. Juist vaardigheden als lezen en schrijven vragen volgens Landauer et al. (2009) om een andere manier van toetsen. Feedback speelt bij die manier van toetsen een centrale rol. Leerlingen hebben over het algemeen namelijk weinig mogelijkheden om lezen en schrijven te oefenen met feedback (Black & Wiliam,

1998), terwijl oefening met lezen en schrijven in combinatie met feedback op maat (Graham & Harris, 2005) een zeer positieve uitwerking heeft op de prestaties van leerlingen.

Om aan deze tekortkomingen tegemoet te treden ontwikkelden Landauer et al. (2009) een *web-based* toets, genaamd WriteToLearn, waarmee de lees- en schrijfprestaties van leerlingen gemeten, gestimuleerd en bijgestuurd worden terwijl de leerling de toets maakt. WriteToLearn bestaat uit twee onderdelen. Bij afname van het eerste onderdeel (Summary Street) maken leerlingen samenvattingen van korte artikelen of boekhoofdstukken. Bij afname van het tweede onderdeel (Intelligent Essay Assessor) schrijven leerlingen een essay over een bepaald onderwerp. Beide onderdelen zijn te bereiken via een webbrowser. Leerlingen loggen in, geven aan welk artikel ze willen samenvatten of over welk onderwerp ze een essay willen schrijven, maken vervolgens hun samenvatting of essay en dienen het ten slotte in via de webbrowser. Na het indienen van het essay of de samenvatting wordt het stuk geëvalueerd en volgt feedback.

Er wordt algemeen aangenomen dat het schrijven van samenvattingen en essays een effectieve manier is om de vaardigheden van leerlingen op het gebied van lezen en taal te versterken (cf. Shanahan, 2005). Er bestaat dan ook weinig twijfel over de betrouwbaarheid en validiteit van de toetsscores die WriteToLearn levert (Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005; Wade-Stein & Kintsch, 2004). Niettemin is er nog weinig gericht onderzoek gedaan naar de effectiviteit van de *web-based* toets WriteToLearn. De studies die wel zijn verricht laten positieve resultaten zien. Ook Franzke et al. (2005) steden de vraag wat het effect is van het onderdeel Summary Street op de lees- en schrijfvaardigheid.

- ***Onderzoeksopzet.*** Er namen 121 leerlingen uit vier klassen in grade 8 (leeftijd 13-14 jaar) aan het onderzoek deel. Het onderzoek was opgezet als een *randomized controlled trial* met random toewijzing van leerlingen aan de interventie- en controlegroep. Tijdens de voor- en nameting werden onder andere leesbegrip en woordenschat getoetst en de kwaliteit van de samenvattingen werd beoordeeld. Na de voormeting kreeg de ene helft van de klas geautomatiseerde feedback op hun schrijfprestaties en de andere helft niet. De controlegroep maakte samenvattingen van dezelfde teksten met een tekstverwerkingsprogramma en kregen evenveel revisietijd als de experimentele groep.

- ***Diagnostic assessment.*** De leerlingen in de interventiegroep werkten gedurende vier weken met Summary Street. De leerlingen maken een samenvatting en krijgen geautomatiseerd feedback over de: (a) *inhoud* – sluit de samenvatting voldoende aan bij de tekst?, (b) *lengte* – is de oorspronkelijke tekst adequaat ingekort?, (c) mate van *kopiëren* – is er niet teveel rechtstreeks gekopieerd uit de oorspronkelijke tekst?, (d) *spelling* – zijn de woorden correct gespeld?, (e) mate van *redundantie* – zit er niet teveel herhaling in de samenvatting?, en (f) *relevantie* – staan er zinnen in die slecht samenhangen met de rest en ook weggelaten hadden kunnen worden? De cyclus wordt in de regel drie tot acht keer herhaald. Op deze manier kunnen leerlingen gericht naar een bepaald leerdoel toewerken en is de toetsafname niet alleen een evaluatie achteraf, maar staat de toetsafname ook in het teken van het aanleren en verbeteren van de vaardigheid in het samenvatten van teksten.

- ***Resultaten.*** Na afloop van de vier weken boekten de beide groepen evenveel leerwinst op het gebied van leesbegrip en woordenschat. De leerlingen die Summary Street gebruikten, deden het echter significant beter dan de leerlingen in de controlegroep op de leesvaardigheidsitems waarbij de leerling een samenvatting moest geven (Cohen's d = .42). Daarnaast was de inhoudelijke kwaliteit van de geschreven samenvattingen in de groep die Summary Street gebruikte, bijna één standaarddeviatie hoger dan in de controlegroep (Cohen's d = .90). Dit betrof niet alleen de algemene kwaliteit en de inhoud van de samenvattingen, maar ook de organisatorische en stilistische kwaliteit van de teksten. Kennelijk hebben de leerlingen die Summary Street gebruikten baat gehad bij de inhoudelijke feedback. De verbetering van de stilistische kwaliteit van de teksten is opmerkelijk omdat de feedback hier niet op gericht was. Er werden geen verschillen tussen beide groepen gevonden bij spelling, interpunctie en zinsbouw maar daarover bood het programma dan ook geen feedback. Niet alle leerlingen profiteerden in gelijke mate van de gecomputeriseerde feedback: de laag tot gemiddeld presterende leerlingen hadden daar meer baat bij dan de bovengemiddeld presterende leerlingen.

- ***Conclusie.*** Dit onderzoek toont de mogelijke meerwaarde aan van een web-based toets voor het meten en verbeteren van de schrijfvaardigheid. De resultaten zijn vergelijkbaar met de uitkomsten van ander onderzoek. In een onderzoek van Wade-Stein en Kintsch (2004) boekten de leerlingen die Summary Street gebruikten significant meer vooruitgang in begrijpend lezen en schrijfvaardigheid dan de leerlingen in de controlegroep. De samenvattingen van de leerlingen in de experimentele groep bleken bovendien ook in de periode ná afronding van de studie van hogere kwaliteit te zijn dan de samenvattingen van de controlegroep. Daarnaast besteedden de leerlingen in de feedbackconditie veel meer tijd aan het verbeteren van hun samenvattingen dan degenen in de controlegroep. In onderzoek van Caccamise et al. (2007) bevatten de samenvattingen van

de leerlingen die WriteToLearn gebruikten vijftig procent meer relevante inhoud dan de samenvattingen van de controlegroep. Ook bleek WriteToLearn een positieve uitwerking te hebben op de prestaties van leerlingen op een traditionele begrijpend leestoets. Nader onderzoek moet uitwijzen of een web-based toets als WriteToLearn inderdaad houvast geeft bij het meten, stimuleren en bijsturen van de lees- en schrijfontwikkeling van leerlingen. In dat onderzoek behoeft het geautomatiseerde evaluatie- en feedbacksysteem dat ten grondslag ligt aan WriteToLearn speciale aandacht. Het is onduidelijk of het systeem gemakkelijk over is te zetten naar een taal als het Nederlands en/of er bestaande systemen zijn die op een dergelijke wijze, na aanpassingen, kunnen worden ingezet.

Studie 2: Ferster, Hammond, Alexander en Lyman (2012)*

- ***Achtergrond en vraagstelling.*** In de hectiek van de hedendaagse klas is het vaak niet mogelijk om leerlingen frequente, tijdige en kwalitatief hoogwaardige feedback op hun schrijfproducten te geven. Een mogelijke oplossing voor dit probleem is de feedback geautomatiseerd, met behulp van de computer, aan te bieden. Tegenwoordig zijn er diverse systemen voor *Automated Essay Scoring* (AES) beschikbaar die leerlingen gedetailleerde feedback geven over hoe zij hun schrijfproducten kunnen verbeteren. Ferster et al. (2012) vroegen zich af in hoeverre computer-gegenereerde feedback vergelijkbaar is met menselijke feedback en in hoeverre die feedback bruikbaar is. De feedback werd gegeven op het schrijven van essays en scripts voor digitale documentaires over geschiedkundige onderwerpen.

- ***Onderzoeksopzet.*** De onderzoeksgroep bestond uit 87 leerlingen uit grade 7 van een Amerikaanse middenschool. Er was een gebalanceerd onderzoeksontwerp met een voor- en nameting. De leerlingen schreven een traditioneel essay of een script voor een digitale documentaire (dit wil zeggen: een kort digitaal filmpje bestaande uit een montage van beelden, tekst en grafieken met een door de student ingesproken verhalende tekst).

- ***Diagnostic assessment.*** De ongeveer 70 essays en 70 digitale documentaires werden beoordeeld door twee beoordelaars en met de *Criterion*^{TM} *online essay evaluation service* van *Educational Testing Service's* (ETS). De beoordelaars beoordeelden de essays en scripts op zes aspecten, te weten inhoud, organisatie, verhaaltoon, woordkeuze en conventies; daarnaast gaven zij een holistisch oordeel over de algehele kwaliteit. *De Automated Essay Scoring* (AES) van ETS bood de studenten een holistische score voor de algehele kwaliteit en uitgebreide feedback op het gebied van grammatica, taalgebruik, conventies, stijl en organisatie.

- ***Resultaten.*** De correlatie tussen de oordelen van de beoordelaars en de computer bedroeg .79 voor de essays en .73 voor het script bij de digitale documentaires. De gevonden hoge correlaties zijn in overeenstemming met eerder onderzoek waarin overeenstemmingspercentages tussen de 80 en 90 procent werden gevonden (Attali & Burstein, 2006; Burstein, 2003; Hearst, 2000). De onderzoekers keken ook naar de kwaliteit van de geautomatiseerde feedback door na te gaan in hoeverre de opmerkingen bij de scripts voor de leerlingen begrijpelijk waren. De meeste opmerkingen bleken hout te snijden, al waren ze wel in algemene termen geformuleerd en boden ze daardoor weinig hulp bij de revisie van de tekst. Een voorbeeld was de opmerking: "*the essay provides a clear sequence of information; provides pieces of information that are generally related to each other*." Binnen de 144 met de computer beoordeelde schrijfproducten vonden de onderzoekers minder dan tien voorbeelden van foutieve beoordelingen, alle van grammaticale aard. Een voorbeeld was de zin die begon met "*In the early 1900's"* waarbij de computer het gebruik van het lidwoord '*the*' ten onrechte als foutief markeerde. De correlatie tussen de lengte van de tekst en de holistische score voor AES was beduidend hoger dan voor het menselijk oordeel (*r* was respectievelijk .81 versus .67).

- ***Conclusie.*** De bemoedigende resultaten doen vermoeden dat geautomatiseerde feedback kan uitgroeien tot een bruikbaar middel aan de hand waarvan leerlingen hun schrijfproducten kunnen verbeteren. Zolang de computer echter geen feedback op de vakinhoudelijke kwaliteit van de geschreven teksten kan geven, zal AES menselijke beoordeling nooit volledig kunnen vervangen. Aangezien er een hoge correlatie was tussen het cijfer en de lengte van de tekst, biedt AES studenten meer mogelijkheden om hun cijfer kunstmatig te verhogen door het programma met het schrijven van langere teksten om te tuin te leiden. Voor het Nederlandse taalgebied zijn nog geen *Automated Essay Scoring* systemen beschikbaar (Feenstra, 2014) die leerlingen gedetailleerde feedback geven over hoe zij hun schrijfproducten kunnen verbeteren. Het verdient aanbeveling een dergelijk systeem voor het Nederlandse taalgebied te ontwikkelen.

3.3.2 Leesvaardigheid

Teo (2012) voerde een actieonderzoek uit naar de mogelijkheid van een digitale dynamic assessment voor het verbeteren van het leesbegrip. Sainsbury and Benton (2011) doen verslag van de ontwikkeling van een diagnostisch feedbacksysteem voor het toetsen van beginnende leesvaardigheid. Kalyuga (2007) deed een kleinschalig onderzoek naar de betrouwbaarheid, validiteit en bruikbaarheid van een diagnostisch feedbacksysteem voor leesvaardigheid.

Studie 1: Teo (2012)

- ***Achtergrond en vraagstelling.*** Teo (2012) beschrijft een actieonderzoek waarin de mogelijkheid van een digitale dynamische toets bij het verbeteren van het leesbegrip werd onderzocht. De aanleiding is dat het Taiwanese beleid ten aanzien van Engels onderwijs en het toetsen ervan is veranderd. Van leraren wordt verwacht dat zij niet alleen van statische toetsen gebruikmaken maar van diverse procedures om het leerproces en de leervorderingen van studenten in kaart te brengen (Chan, 2006). Dit met het oog op de extra informatie die andere toetsprocedures kunnen leveren aan leraren om het onderwijs op de studenten af te stemmen. *Dynamic assessment* met hulp (mediatie) en feedback helpt om het leerpotentieel van leerlingen zichtbaar te maken. Gedurende een dynamische toetsprocedure zou een vorm van leren moeten plaatsvinden, iets wat tijdens de afname van een statische toetsen niet gebeurt (Feuerstein, Feuerstein & Falik, 2010). Bij *dynamic assessment* moet de student vragen beantwoorden maar kan op basis van feedback het antwoord verbeteren. Het wordt duidelijk of de student in staat is om geleerde of aangereikte principes zelf in nieuwe situaties toe te passen. Hierin speelt de examinator of docent een belangrijke rol. Maar één-op-één begeleiding kan in klassen met veel leerlingen of studenten een probleem voor leraren vormen. Teo (2012) redeneerde dat de inzet van ICT hierin uitkomst kan bieden.

- ***Onderzoeksopzet.*** Aan het onderzoek namen 68 universitair studenten van 18 en 19 jaar oud deel. Er werd gekeken met een statische leesbegripstoets bij voor- en nameting of de studenten na een periode van tien wekelijkse dynamische testsessies van een lesuur vooruitgingen in het gebruik van metacognitieve leesstrategieën.

- ***Diagnostic assessment****.* Het doel was om ICT in te zetten bij het begrijpend lezen van Engelse teksten waarbij het Engels een vreemde taal voor de studenten is. Er werd een gecomputeriseerde dynamische procedure ontwikkeld waarin het geven van feedback en aanwijzingen gecombineerd werd met toetsing. Volgens Teo en Jen (2012) speelt de leraar in de nieuwe procedure een relatief bescheiden rol als begeleider en facilitator die alleen intervenieert als het echt nodig is. De geautomatiseerde feedback en aanwijzingen waren gericht op het gebruik van metacognitieve strategieën bij het lezen van teksten. Het gebruik van dergelijke strategieën voor, tijdens en na het lezen draagt bij aan een goed leesbegrip (o.a. Pressley & Afflerback, 1995). In de leesinstructie zou dan ook aan deze strategieën aandacht moeten worden besteed. Het computerprogramma werd zo ontworpen dat de student hulp kreeg bij het beantwoorden van een vraag over een tekst. De vraag werd gesteld met vijf meerkeuze-antwoorden. Na elk onjuist antwoord, volgde een aanwijzing. Er werden vier niveaus onderscheiden van impliciet naar meer expliciet (zie Tabel 3.3).

Tabel 3.3 Vier niveaus van feedback, geordend van impliciet naar expliciet.

Niveau	Type feedback
1	Er wordt uitleg gegeven over het maken van inferenties bij het lezen. Er wordt uitleg gegeven over het belang van tussen de regels lezen. Het antwoord kan niet worden gegeven door naar de tekst alleen te kijken. Er wordt gevraagd om de hoofdgedachte van de passage te vinden. De betekenis van de kernwoorden wordt gegeven en er worden tips gegeven over waar de hoofdgedachte meestal staat in een passage.
2	Er wordt een aanwijzing gegeven waarbij er verwezen wordt naar een bepaalde paragraaf of zin. Ook wordt er uitleg gegeven over de overkoepelende betekenis van de specifieke context, bijvoorbeeld: de auteur probeert iets duidelijk te maken over een bepaalde periode in de maatschappij, wat is dat?
3	Er wordt een aanwijzing gegeven waarbij er verwezen wordt naar één zin, frase/zinsdeel of woord. Dit is context-specifiek in plaats van uitleg over de overkoepelende betekenis.
4	Het antwoord wordt gegeven met stapsgewijze uitleg over hoe je als lezer tot het juiste antwoord kunt komen en welke informatie uit de tekst je daarvoor nodig hebt.

- ***Resultaten.*** Er was een significante vooruitgang (*t*(67)=-2.70, *p=.009*) op de leesbegripstoets. Uit de portfolio's die de studenten in hun eigen taal (Chinees) bijhielden bleek dat zij, reflecterend op de taak, meer gebruik van metacognitieve strategieën rapporteerden. Studenten die op niveau 3 aanwijzingen nodig hadden, bleken meer baat te hebben bij één-op-één hulp van de docent.

- ***Conclusie****.* Hoewel er een positief effect op de nameting was, is niet met zekerheid vast te stellen dat de vooruitgang geboekt werd door (deels) oefening en niet specifiek door *dynamic assessment*. Er was immers geen controlegroep. Ook was de geautomatiseerde feedback niet voor alle studenten voldoende. De zwakste lezers hadden behoefte aan hulp van de docent. Met de informatie van *dynamic assessment* was de docent wel beter in staat te bepalen welke hulp en instructie voor bepaalde studenten nodig was. Het programma hielp op de studenten te identificeren die individuele hulp van de docent nodig hadden.

Studie 2: Sainsbury and Benton (2011)*

- ***Achtergrond en vraagstelling.*** Dit onderzoek (Sainsbury & Benton, 2011) was gericht op het ontwikkelen van een digitaal instrument dat bruikbaar zou zijn voor het diagnostisch toetsen van '*early reading skills*' door leraren in het basisonderwijs in de UK. Snelle ontwikkelingen op technisch gebied maken inmiddels het computer-gebaseerd toetsen ('e-assessment') mogelijk, wat goed gebruikt zou kunnen worden in de schoolpraktijk van zowel formatief als summatief toetsen. E-assessment is uitermate geschikt om onmiddellijke analyses en zorgvuldig geformuleerde rapportages op te leveren over de bij de leerling aanwezige kennis en vaardigheden, die nodig is voor het geven van toegespitste feedback aan de leerling. Om formatief effectief te zijn moet feedback vooral beschrijvend van aard zijn, in plaats van numeriek (met gebruik van scores in getallen). Sainsbury en Benton (2011) doen verslag van de ontwikkeling van een digitaal toetsinstrument, gericht op het geven van beschrijvende feedback op '*early reading skills'* van leerlingen, die bruikbaar zou zijn voor zowel leraren als leerlingen.

  Het doel van de studie was om statistische analyses te vinden waarmee informatie wordt opgeleverd over patronen in het scoregedrag van leerlingen en waarmee deze patronen te interpreteren en zo weer te geven zijn dat ze direct bruikbaar zijn voor leraren in de lessituatie. Er wordt onderzocht hoe e-assessment gecombineerd kan worden met een specifieke statistische analyse gericht op het opsporen van latente klassen ofwel ontwikkelingstypen van leerlingen gebaseerd op hun toetsresultaten, om zo onderbouwde, formatieve informatie te leveren voor gebruik in les- en leersituaties. Hiermee worden in deze studie de theoretische inzichten en empirische bevindingen uit onderzoek naar formatief toetsen (o.a. Black & Wiliam, 1998a, 1998b) – inclusief de cruciale rol van feed back hierbij (o.a. Wiliam, 2004) – gecombineerd met inzichten en bevindingen uit onderzoek naar cognitief-diagnostisch toetsen (o.a. Leighton & Gierl, 2007; Huff & Goodman, 2007).

- ***Onderzoeksopzet.*** Digitale toetsen gericht op '*early reading skills'* (vaardigheden: fonologische segmentatie, rijmen en geschreven woordherkenning) werden ontwikkeld en afgenomen bij ruim 600 leerlingen van 5-7 jaar in 26 scholen in Engeland. Om patronen in scoregedrag van leerlingen te vinden is gebruikgemaakt van 'Latent Class Analysis' (LCA, Hagenaars & McCutcheon, 2002), een statistische methode waarmee onderliggende typen van individuen ('latente klassen') gevonden kunnen worden op basis van hun antwoorden op tests en vragenlijsten. Het latente klasse model relateert een set van geobserveerde responsen aan een onderliggende latente variabele of factor. Deze latente variabele wordt niet direct geobserveerd maar wordt afgeleid uit de geobserveerde data. Latent Class Analysis (LCA) werd gebruikt om de leerlingen op basis van hun responsen in te delen in verschillende categorieën lezers met verschillende leer- en ontwikkelingsprofielen (bijvoorbeeld 'sight reader', 'sound reader', 'developing reader', 'balanced reader'). Deze profielen werden kort beschreven. Een 'sight reader' wordt bijvoorbeeld omschreven als een lezer die woorden direct herkent en op deze leesstrategie vertrouwt bij het lezen, terwijl vaardigheden in het onderscheiden van fonemen minder goed ontwikkeld is.

- ***Diagnostic assessment.*** Zodra de gedragsprofielen bepaald waren op basis van de analyse van de scores van de leerlingen op de digitale toetsen, werden ze verwerkt in een speciaal geprogrammeerd automatisch rapportagesysteem, waaruit formatieve rapporten kwamen, die direct bruikbaar waren, zowel voor de leerkracht als voor de leerling. Bij elk profiel werd een beschrijving gegeven van de belangrijkste kenmerken van de sterke en zwakke punten in de scores van de leerlingen in die groep. Daarbij werd ook een advies gegeven voor vervolgstappen voor de leraar en een advies in leeftijdsadequate bewoordingen voor de leerling. In het advies werd het niveau van de leerling beschreven (de score) en de kenmerken van het leesgedrag (in staat om onregelmatige woorden direct te herkennen, in staat om fonemen te onderscheiden).

In het rapport stonden ook adviezen aan de leraar voor het onderwijs (blijf werken aan het snel en goed lezen van woorden, besteed aandacht aan auditieve synthese en analyse).

- ***Resultaten.*** Het rapportagesysteem is gebruikt door de leraren van de deelnemende leerlingen. De leraren stuurden de digitale scores van hun leerlingen op en kregen onmiddellijk online toegang tot verschillende kwalitatieve en kwantitatieve rapporten, waaronder de profielen van hun leerlingen. Een deel van de leraren heeft tijdens een bijeenkomst en een beperkt aantal schoolbezoeken feedback aan de onderzoekers gegeven over het gebruik van dit systeem. De opgeleverde profielen leken valide en in overeenstemming met het beeld dat de leraren zelf hadden van de sterke en zwakke punten van hun eigen leerlingen. De beschreven gedetailleerde feedback en suggesties voor instructie in de volgende stappen leken ook bruikbaar voor de leraren.

- ***Conclusie.*** De resultaten van het proefdraaien met dit systeem levert een globaal beeld op, dat nog nader onderzoek behoeft. Zo moet nog worden onderzocht in hoeverre de leerlingprofielen accuraat zijn. Ook is er meer empirische evidentie nodig voor de mate en wijze waarop de leraren de profielbeschrijvingen in de praktijk kunnen gebruiken om hun instructie aan te passen en bovendien voor het meten van hoe succesvol dit is in het vergroten van de opbrengsten bij de leerling. Het daadwerkelijk aan de slag gaan met dit systeem in een setting waarin formatieve toetsing met feedback nuttig lijkt, is dus nog niet beschreven in deze studie en moet nog verder uitgeprobeerd en onderzocht worden. Dit vormt een duidelijke aanbeveling voor follow-up onderzoek op dit vlak, zowel voor het vaststellen van de accuraatheid van de automatisch gegenereerde profielen als het in kaart brengen van het daadwerkelijke gebruik van de geleverde feedback voor instructie door leraren en het effect hiervan op de ontwikkeling van specifieke vaardigheden bij leerlingen in deze leeftijdsgroep (jonge, beginnende lezers).

Studie 3: Kalyuga (2006)*

- ***Achtergrond en vraagstelling.*** Instructiemethoden die effectief zijn voor beginnende leerlingen kunnen later in de schoolloopbaan, als leerlingen vaardiger zijn, ineffectief worden (Kalyuga, Ayres, Chandler, & Sweller, 2003; Kalyuga, Chandler, & Sweller, 2001). Om deze reden is het belangrijk om het vaardigheidsniveau van leerlingen nauwgezet te monitoren, zodat leraren op verschillende momenten in de schoolloopbaan kunnen bepalen welke leerstof en didactische benadering het meest geschikt is om leerlingen te begeleiden. Traditionele toetsen geven vaak niet de diagnostische informatie die de leerkracht nodig heeft. De observatie dat een leerling een serie wiskundige vergelijkingen (bijvoorbeeld $4x = -2$) in een meerkeuzetoets correct kan oplossen, vertelt bijvoorbeeld niet *hoe* de wiskundige vergelijkingen zijn opgelost. De leerling kan als beginner te werk zijn gegaan en door *trial-and-error* tot de oplossing gekomen zijn, maar kan het vraagstuk ook als expert aangepakt hebben en door toepassing van een geautomatiseerde procedure direct te weten zijn gekomen dat –½ het juiste antwoord is. Ook de traditionele begrijpend leestoetsen geven doorgaans weinig tot geen informatie over het cognitieve proces en de methode die ten grondslag ligt aan het antwoord van een leerling (cf. Magliano & Millis, 2003). Een leerling kan bijvoorbeeld eerst de vragen gelezen hebben en vervolgens in de tekst zijn gaan zoeken naar de juiste antwoorden, maar kan de tekst ook van het begin tot het einde gelezen hebben, daarbij een coherente mentale representatie van de tekst geconstrueerd hebben, en daarna gestart zijn met het beantwoorden van de vragen.

  Om zicht te krijgen op de mate waarin leerlingen hun kennis in bruikbare cognitieve structuren bijeengebracht hebben, moeten er toetsen zijn die de complexiteit van kennis en kunde weten te vangen (Marshall, 1995). Het is de vraag hoe de kennisstructuur van leerlingen op een bepaald leerstofgebied efficiënt in kaart gebracht kan worden. Vaak wordt aan leerlingen gevraagd om hardop te denken als zij een vraagstuk oplossen of een tekst lezen (Ericsson & Simon, 1993; Magliano & Millis, 2003). Deze manier van toetsen is echter tijdrovend en lastig te automatiseren (Kalyuga, 2007). Een alternatief is om elk vraagstuk in de toets kort aan leerlingen te presenteren en te vragen welke stap zij als eerste zouden zetten om het vraagstuk op te lossen. Vaardige leerlingen zullen het vraagstuk herkennen en direct in staat zijn om een doeltreffende oplossingsstrategie op te halen uit het lange-termijn geheugen (Kalyuga, 2003; Kalyuga & Sweller, 2004). Zwakkere leerlingen zullen over het algemeen een minder probate oplossingsstrategie kiezen. Deze zogeheten *rapid schema-based approach* is reeds toegepast en beproefd bij rekenen en wiskunde (Kalyuga, 2004; Kalyuga & Sweller, 2005). Enkele jaren geleden heeft Kalyuga (2007) geprobeerd om de *rapid schema-based approach* toe te passen bij begrijpend lezen. Voor zover bekend is de *rapid schema-based approach* zoals die door Kalyuga (2007) is toegepast bij begrijpend lezen nog niet geïmplementeerd in toetsen die in de onderwijspraktijk gebruikt worden. Er is ook nog maar beperkt onderzoek gedaan naar de bruikbaarheid en effectiviteit van de benadering bij het meten van taalvaardigheden.

Het is niet eenvoudig om de strategieën die leerlingen toepassen bij het lezen van teksten vast te stellen. Het leerstofgebied begrijpend lezen is immers weinig gestructureerd en er zijn vele manieren om het doel (tekstbegrip) te bereiken, zeker in vergelijking met leerstofgebieden zoals rekenen en wiskunde. Kalyuga (2007) stelt voor om tekstbegrip te meten door een serie zinnen aan te bieden die variëren in structuur en grammaticale complexiteit. Het idee is dat zwakkere leerlingen vanwege hun beperkte kennis alleen de eenvoudige zinnen kunnen begrijpen. Van ervaren leerlingen wordt verwacht dat zij, afhankelijk van hun kennis van grammatica en structuur, ook complexere zinnen kunnen begrijpen.

- ***Onderzoeksopzet.*** Kalyuga (2007) heeft een kleinschalige studie uitgevoerd bij 34 dertienjarige leerlingen. De *rapid test* werd digitaal afgenomen en bevatte 18 zinnen die opliepen in moeilijkheidsgraad. Elk zin werd kort aangeboden. De tijdspanne werd bepaald door het aantal woorden in de zin. Per woord werd ongeveer 1 seconde gerekend. Proefonderzoeken lieten zien dat 1 seconde per woord voldoende zou moeten zijn om de zin eenmaal te lezen. Na elke zin werden, na elkaar, vier uitspraken getoond. Bij elke uitspraak moest de leerling aangeven of deze correct of incorrect was. De responsen werden dichotoom gescoord.

- ***Diagnostic assessment.*** Om zwakkere leerlingen niet onnodig te belasten met zinnen die ver boven hun niveau liggen, gebruikt Kalyuga (2007) in zijn onderzoek een toets die oploopt in moeilijkheidsgraad. Bij het construeren van zinnen en controle-uitspraken is aandacht besteed aan het niet overschrijden van de maximale capaciteit van het werkgeheugen. Volgens Gibson (1998) laat het werkgeheugen maximaal vier wisselingen van onderwerp in een zin toe. Het aantal afhankelijkheidsrelaties in een zin of het aantal voorwaardelijkheden beperkt zich bij voorkeur tot twee per keer (cf. Kimball, 1973). De toets begint met enkele eenvoudige zinnen, wordt dan doorgegaan met enkele moeilijkere samengestelde zinnen, en de toets eindigt met zeer complexe zinnen waarin wisselingen van onderwerp, afhankelijkheidsrelaties en/of voorwaardelijkheden voorkomen. Bij elke zin heeft Kalyuga (2007) vier uitspraken geconstrueerd die al dan niet correct zijn. Bij de zin: “De artiest, die optrad voor het publiek dat was komen opdagen om van de show te genieten, vertrok” zouden de uitspraken om te controleren of de leerling de zin begrepen heeft bijvoorbeeld kunnen zijn: (a) “de artiest vertrok”, (b) “de artiest genoot van de show”, (c) “het publiek kwam opdagen voor de show”, en (d) “het publiek vertrok”. In dit geval zijn uitspraken één en drie correct.

  Kalyuga (2007) onderscheidt vier niveaus van leesvaardigheid en classificeert leerlingen in één van deze niveaus op basis van de moeilijkheidsgraad *D* van een zin. De moeilijkheidsgraad hangt af van het aantal wisselingen in onderwerp en het aantal afhankelijkheidsrelaties en/of het aantal voorwaardelijkheden in een zin. Zinnen met een *D*-waarde van 1, 2 of 3 horen respectievelijk bij leesniveaus 1, 2 en 3. Zinnen met een *D*-waarde van 4 of hoger sluiten aan bij leesniveau 4. Bij leesniveau 1 kan bijvoorbeeld de volgende zin horen: “Het lawaai van de drukke stad werd verruild door de stilte van het platteland” (D = max [0, 1] = 1). Bij leesniveau 2 kan er een voorwaardelijkheid in de zin voorkomen, zoals: “De student, die de cursus leuk vond, heeft het project afgerond” (*D* = [2, 2] = 2). Een zin met een wisseling in onderwerp en >1 voorwaardelijkheden, zoals: “De student, die gekozen was door de leerkracht, rapporteerde aan de trainer die het schoolteam samenstelde” (*D* = [2, 3] = 3), is een voorbeeld van zin die bij leesniveau 3 hoort. Het hoogste leesniveau (4) omvat zinnen zoals: “ Het gegeven dat de student, die door de leerkracht geprezen werd, zakte voor de cursus, baarde de directeur zorgen” (*D* = [5, 3] = 5). De meest ingewikkelde zinnen (ook leesniveau 4) bevatten een groot aantal voorwaardelijkheden en onderwerpwisselingen, bijvoorbeeld: “Dit is de cursus waarvoor de student, die van de school gestuurd is die door de media bekritiseerd werd, zakte” (*D* = [7, 4] = 7). Als dergelijke zinnen kort worden aangeboden, kunnen alleen zeer vaardige leerlingen de zin begrijpen. Voor zwakkere lezers is de zin te complex.

- ***Resultaten.*** De scores van de leerlingen bleken goed samen te hangen met de scores die dezelfde leerlingen behaalden op een traditionele begrijpend leestoets (.66). De betrouwbaarheid bleek met .73 voldoende te zijn voor het nemen van minder belangrijke beslissingen op het individuele niveau. In vergelijking met de traditionele begrijpend leestoets was de afnametijd bij de *rapid test* met een factor 3.8 korter.

- ***Conclusie.*** De resultaten zijn hoopgevend. Nader onderzoek zou moeten uitwijzen of de *rapid schema-based approach* inderdaad de diagnostische informatie geeft die leraren nodig hebben om te bepalen welke leerstof en didactische aanpak aansluit bij de begrijpende leesvaardigheid van leerlingen. Ook zou nagegaan moeten worden in hoeverre de aanpak toe te passen is bij vaardigheden als woordenschat of schrijfvaardigheid.

4 Conclusies en discussie

4.1 Onderzoeksresultaten

In deze literatuurstudie stelden we de vraag of formatieve toetsen het onderwijs houvast kunnen geven bij het vormgeven en evalueren van het taalonderwijs in verschillende onderwijssectoren. Op basis van een aantal criteria selecteerden we veertien studies die we in detail hebben beschreven. Van elke studie beschreven we de achtergrond en vraagstelling, de onderzoeksopzet, de toetsingsvorm, resultaten en conclusie. Daarbij maakten we onderscheid in toetsvormen met feedback gegeven door mensen of via de computer. Het ging om studies naar dynamische toetsing en diagnostische toetsing waarbij we de Engelstalige termen *dynamic assessment* en *diagnostic assessment* hanteren. Het doel van toetsing kon zijn om leerlingen te identificeren of te onderscheiden (screening en plaatsing), het leerproces te verbeteren, het leerpotentieel vast te stellen of de zwakke en sterke punten van een leerling vast te stellen (diagnostisch). Zoals hieronder uiteengezet, kan uit de studies voorzichtig de conclusie worden getrokken dat de resultaten overwegend positief zijn.

In de gevonden studies werd een vorm van *dynamic assessment* een aantal keer gehanteerd als middel om nauwkeuriger onderscheid te maken tussen leerlingen met een taalachterstand en leerlingen met een taalstoornis (Peña, et al., 1992, 2006; Kester, et al., 2001; Kapantzoglou et al., 2010; Ukrainetz et al., 2000). Er wordt gesteld dat assessments die op dynamische wijze worden afgenomen een meer betrouwbare en bruikbare schatting geven van de vaardigheid van leerlingen dan inhoudelijk identieke assessments die op traditionele wijze worden afgenomen (zie Dillon, 1997). *Dynamic assessments* zouden niet alleen inzicht geven in het huidige vaardigheidsniveau van leerlingen, maar ook laten zien hoe het vaardigheidsniveau door middel van interventies te beïnvloeden is. *Static assessments* worden daarentegen op zo'n manier afgenomen dat leren niet zal plaatsvinden (Feuerstein, Feuerstein & Falik, 2010). Leerlingen uit minderheidsgroepen en/of uit een omgeving met lage SES presteren dikwijls zwak op de statische standaardtoetsen en er wordt onterecht een taalstoornis gediagnosticeerd. Hun onderprestatie heeft soms te maken met de taakeisen van de toets, waarmee de leerlingen niet vertrouwd zijn. In de studies uit deze review naar *dynamic assessment* was er sprake van een vorm van mediatie: de leerlingen namen deel aan één of meer sessies waarin ze taalopdrachten kregen waarop feedback werd gegeven. In enkele studies werd gevonden dat *static assessments* geen betrouwbaar onderscheid maakten tussen leerlingen met en zonder taalstoornis, terwijl dat op basis van een *dynamic assessment* wel kon. Positieve resultaten zijn gevonden voor het toetsen van woordenschat (Kester et al., 2001; Peña et al., 1992; Ukrainetz et al., 2000; Kapantzoglou et al., 2010) en mondelinge taalvaardigheid (Peña et al., 2006). De betere bruikbaarheid voor screenings- en plaatsingsdoelen bleek onder meer uit de resultaten op een gestandaardiseerde toets op de nameting (Kester et al., 2001; Peña et al., 1992) en/of uit een observatielijst die tijdens de sessie werd gehanteerd (Ukrainetz et al., 2000; Kapantzoglou et al., 2010). Kester et al. (2001) concludeerden dat *dynamic assessment* het beter mogelijk maakt om de instructie aan te passen aan verschillen tussen kinderen in voorkennis, persoonlijkheid, taalinvloed en voorafgaande onderwijservaringen.

In een aantal studies werd een vorm van *dynamic assessment* ingezet om meer zicht te krijgen op het leerproces van de leerling. Het kan een benadering zijn om inzicht te krijgen in het huidige kennis- en vaardigheidsniveau en hoe dit kan worden beïnvloed (Dörfler, Golke & Artelt, 2009). Larsen en Nippold (2007) onderzochten een toetsvorm waarbij gekeken werd in hoeverre een leerling in staat was om een morfologisch analytische strategie toe te passen om de betekenis van nieuwe, onbekende woorden te bepalen. De toets gaf nauwkeurig zicht op welke kennis leerlingen beheersten en gaf op deze manier meer zicht op de instructie die leerlingen nodig hebben. Elleman et al. (2011) concludeerden eveneens dat DA meer zicht gaf op intra-individuele verschillen. Zij onderzochten een toets voor leesbegrip waarbij leerlingen vragen beantwoordden en feedback met hints kregen. Leerlingen die meer hints nodig hadden presteerden zwakker op een standaardtoets voor leesbegrip. *Dynamic assessement* gaf meer inzicht in de begripsvaardigheden die de leerlingen beheersten.

Hoewel de positieve effecten van *dynamic assessments* alom benadrukt worden, is er slechts sporadisch onderzoek gedaan naar de betrouwbaarheid, validiteit en effectiviteit van alternatieve toetsvormen voor taal en lezen die op dynamische wijze worden afgenomen. Veel studies richten zich op de vormgeving en implementatie van een training of interventie als onderdeel van een toetsing en niet op het betrouwbaar en valide meten van taal- en leesvaardigheid (cf. Dörfler, Golke & Artelt, 2009; Grigorenko & Sternberg, 1998). De studie van Elleman et al. (2011) is een uitzondering. Zij gingen in op de (a) interne consistentie van de toetsvorm – betrouwbaarheid, (b) de

correlatie met een gevalideerde traditionele toets – validiteit, en (c) voorspellende waarde van de toetsvorm bij het identificeren van zwakke leerlingen – effectiviteit. Dergelijk onderzoek is niet alleen belangrijk om zicht te krijgen op de bruikbaarheid van alternatieve toetsvormen, maar ook op het nut ervan. Vanwege de noodzaak van standaardisatie is de uitvoering van de interventiecomponent van een (onderzoeksgerichte) *dynamic assessment* vaak complex. In hun studie wijzen Elleman et al. (2011) op de voordelen van dynamische toetsvormen, maar de inspanning die de (individuele) afname van de test vraagt, moet ook worden overwogen. Evenals de effectiviteit van de verschillende componenten waaruit *dynamic assessment* bestaat. Een aantal van de door ons aangetroffen dynamische assessments namen echter weinig tijd in beslag (o.a. Ukrainetz et. al, 2000; Peña et al., 1992; Kapantzoglou et al., 2010). Kramer et al. (2009) gebruikten een onderzoeksgericht *dynamic assessement* die weinig tijd in beslag nam, maar concludeerden dat de beoordeling van de leerlingprestaties even tijdrovend kan zijn als bij statische toetsen. Ukrainetz et al. (2000) suggereerden dat observatielijsten beknopt, eenvoudig en tijdefficiënt kunnen worden ingezet om leerpotentieel vast te stellen. Fuchs et al. (2011) stelden eveneens dat een observatielijst geschikt is om leerpotentieel vast te stellen. Kapantzoglou et al. (2010) concludeerden dat observatielijsten kunnen worden gebruikt om taalstoornissen te screenen en onderscheid te maken tussen taalachterstand en -stoornis.

Gezien de tijdsinvestering die *dynamic assessments* vragen, wordt soms een oplossing gezocht in het gebruik van computers. Niet de leraar geeft dan feedback op de prestatie van de leerling maar de computer genereert feedback. Doordat de feedback niet in echte interactie plaatsvindt, en de uitkomsten vooral zicht geven op de sterktes en zwaktes van een leerling, beargumenteerden we dat hier sprake is van *diagnostic assessment* en niet zozeer van *dynamic assessment.* In de door ons gevonden studies werd de computer ingezet om de schrijfvaardigheid (Ferster et al., 2012; Franzke et al., 2005) of leesvaardigheid (Teo, 2012; Siansbury & Benton, 2011; Kalyuga, 2006) te toetsen. De inzet van een digitaal systeem dat feedback geeft op samenvattingen (Franzke et al., 2005) of op schrijfproducten (Ferster et al., 2012) van leerlingen had veelbelovende resultaten. De feedback werd gegeven op onder andere de lengte, de woordkeuze, spelling en schrijfregels of conventies, maar de techniek is nog niet zo ver dat de computer ook feedback op de vakinhoud kan geven. Dit beperkt de mogelijkheden van geautomatiseerde feedback op schrijfopdrachten als deze worden ingezet om de vakinhoud te toetsen. Bovendien bleek het uiteindelijke eindoordeel over een tekst, gegeven door de computer, niet altijd betrouwbaar. Zo woog bijvoorbeeld de lengte van een tekst mee, waardoor leerlingen die een lange, maar kwalitatief zwakke tekst schreven, toch een positief oordeel kregen. De leerlingen konden het programma als het ware om de tuin leiden door een lange tekst te schrijven. Ferster et al. (2012) wijzen erop dat de computer de docent niet volledig kan vervangen. Een soortgelijke conclusie trekt Teo (2012) na een studie naar het toetsen van leesbegrip met de computer. De computer genereerde feedback na het fout beantwoorden van meerkeuzevragen en was gericht op het gebruik van metacognitieve strategieën bij het lezen van teksten. Studenten met zwak leesbegrip leken meer baat te hebben bij één-op-één hulp en instructie van de docent. Het was wel mogelijk om met gecomputeriseerde *diagnostic assessment* deze leerlingen te selecteren.

Hoewel de conclusies in de door ons gevonden studies veelbelovend zijn en we voorzichtig positief zijn, was de literatuurstudie van Caffrey, Fuchs en Fuchs (2008) kritischer ten aanzien van de effectiviteit van *dynamic assessment*s. Traditionele toetsen bleken de latere schoolprestaties van leerlingen even goed, en op eenzelfde manier, te voorspellen *als dynamic assessments*. Daarbij merken we op dat taalstudies in de literatuurstudie van Caffrey et al. (2008) zwak vertegenwoordigd waren en dat de conclusie over de beperkte unieke bijdrage van *dynamic assessment* op slechts vier studies gebaseerd was. Een aantal studies op het gebied van taalvaardigheid uit onze literatuurstudie (Fuchs et al, 2008; Elleman et al, 2011) lijken de voorzichtige conclusie van de literatuurstudie van Caffrey et al. (2008) te bevestigen. De resultaten zijn echter moeilijk te vergelijken vanwege verschillen in het aantal en de aard van de voorspellers en criteria, de toegepaste onderzoeksdesigns en de gebruikte analysetechnieken om de incrementele predictieve validiteit vast te stellen. Er is ook onderzoek met positieve resultaten van *dynamic assessment.* In diverse studies verklaart *dynamic assessment* veel hogere percentages unieke variantie, tussen de 9% en 21% (Resing, 1993; Byrne et al, 2000; Meijer, 1993; Swanson, 1994). Nader onderzoek naar de relatieve bijdrage van *dynamic assessment* voor het voorspellen van taalvaardigheid in vergelijking tot statische toetsen zou uitgevoerd moeten worden. Als al onderzoek gedaan wordt naar de effectiviteit van feedback wordt dit vaak gedaan door leerlingen naar hun (waarde)oordeel over de ontvangen feedback te vragen (Carless, 2006; Handley & Williams, 2009; Walker, 2009). Hoewel dergelijke studies laten zien hoe leerlingen de feedback ervaren, komen we niet te weten op welke wijze, en hoe effectief, leerlingen de feedback toepassen bij het uitvoeren van de opdracht (Shrestha & Coffin, 2012).

Een ander punt waarop vooral nadruk heeft gelegen in studies naar formatieve toetsvormen, is de productontwikkeling, hoewel de ontwikkeling binnen een pedagogisch kader zou moeten plaatsvinden met meer aandacht voor de processen van leren en toetsing. Gesuggereerd wordt dat formatieve toetsen voordelen kunnen hebben op de diepere leerprocessen van leerlingen, het behoud van motivatie en de bevordering van zelfwaarde en zelfregulerend leren (Koh, 2008). Er is echter nog weinig bekend over de complexe relatie tussen vormen van

toetsing, leerprocessen en leerstrategieën. Bovendien wordt aangenomen dat de effectiviteit van toetsen beïnvloed wordt door zowel contextuele en persoonlijke kenmerken (Al-Kadri et al., 2012). Naar de relatie tussen toetsvormen, leerprocessen en leerstrategieën en de rol van contextuele als persoonlijke kenmerken is nog onderzoek nodig. Hierbij aansluitend blijkt dat leraren formatieve toetsvormen vooral inzetten als middel om het lesgeven te ondersteunen en minder om het leerproces te beïnvloeden (e.g. Schuwirth & van der Vleuten, 2004). Zij hanteren het als een manier van instructie geven en niet als manier om het onderwijs af te stemmen op het leerproces van de leerlingen. In onderzoek zou meer aandacht kunnen worden besteed aan de inzet van formatieve toetsvormen bij de afstemming tussen het onderwijsaanbod en de leerontwikkeling van leerlingen.

4.2 Praktische implicaties

Met de invoer van de exameneisen in voortgezet onderwijs en middelbaar beroepsonderwijs neemt de noodzaak toe de mogelijkheden van leerlingen goed in kaart te brengen zodat er extra ondersteuning en/of hulpmiddelen kunnen worden ingezet. Van leraren wordt meer en meer verwacht dat zij de leeropbrengsten van de leerlingen systematisch volgen, hun instructie gedifferentieerd aanbieden en hun onderwijs vastleggen in leerplannen. Leraren dienen uit te gaan van leerstandaarden en lesdoelen, informatie te verzamelen tijdens het leerproces, deze vast te leggen voor nadere analyse en interpretatie, en op basis hiervan beslissingen te nemen over het vervolg van het onderwijs (Parrett & Budge, 2009). Door het gebruik van vormen van toetsing en de interpretatie van toetsingsgegevens kunnen aanpassingen in het onderwijsaanbod worden gedaan. In de huidige onderwijspraktijk wordt er voornamelijk gebruikgemaakt van statische toetsvormen. Op deze wijze wordt vastgesteld met welke onderdelen een leerling problemen ervaart, maar het geeft weinig richtlijnen en handvatten voor het vormgeven van instructie en interventie. Met de statische standaardtoetsen is het dan wel mogelijk te bepalen hoe een leerling presteert ten opzichte van zijn leeftijdgenoten, maar leraren hechten meer belang aan informatie over het leerproces en de leerbaarheid voor het plannen van hun onderwijs (Bosma, Hessels en Resing, 2012).

Uit de studies die in deze literatuurstudie beschreven zijn, blijkt dat formatieve toetsvormen zoals *dynamic assessment* en *diagnostic assessment,* mogelijk een bijdrage kunnen leveren aan het taal- en leesonderwijs, hoewel meer onderzoek en bewijs van effectiviteit wenselijk is. Toetsvormen waarbij een vorm van mediatie wordt toegepast (door het geven van feedback en hints, door mensen of geautomatiseerd), kunnen meer zicht geven op het leerproces van de leerling. Deze vormen van toetsing brengen andere vaardigheden die van belang zijn bij bijvoorbeeld begrijpend lezen in kaart dan een statische toets (Elleman, et al., 2011). Studies noemen dit voordeel van formatief toetsgebruik ook bij andere taaldomeinen, zoals onder andere woordenschat (e.g. Larsen & Nippold, 2007), technische leesvaardigheid (e.g. Fuchs et al., 2007), mondelinge taalvaardigheid (e.g. Kramer et al., 2009) en schrijfvaardigheid (e.g. Franzke et al., 2005). In het onderwijs maken leraren al dikwijls gebruik in hun instructie van componenten die deel uitmaken van formatieve toetsing zoals *scaffolding*, feedback, aanwijzingen en hints en aangepaste instructie. Zo passen ze het onderwijsaanbod en toetsen in het onderwijs wat een leerling nodig heeft. Dit vindt echter niet systematisch en gestandaardiseerd plaats om het leerproces en het leerpotentieel van leerlingen te toetsen. Formatieve toetsen waarvan de effectiviteit is aangetoond, zijn (nog) niet beschikbaar.

Uit een aantal door ons gevonden studies blijkt dat observatielijsten eenvoudig, maar efficiënt kunnen zijn om leerpotentieel van leerlingen vast te stellen en op die manier ook taalstoornissen van taalachterstanden te onderscheiden (Ukrainetz et al., 2000; Kapantzoglou et al., 2010). Het ging om vragenlijsten, waarmee werd vastgelegd welke leerstrategieën een leerling toepaste en hoe de leerling reageerde op de instructie, hoeveel instructie er gegeven moest worden en in hoeverre er transfer naar nieuwe taken plaatsvond. Uit de resultaten van deze studies bleek dat observatielijsten leraren van bruikbare informatie kunnen voorzien. Er zijn al verschillende observatielijsten beschikbaar voor het onderwijs, uit de methode of om bijvoorbeeld aspecten van mondelinge taalvaardigheid van leerlingen te beoordelen (e.g. Gijsel & Van Druenen, 2011). Deze vragenlijsten richten zich doorgaans echter niet op het vaststellen van het leerpotentieel van de leerlingen. Ukrainetz et al (2000) en Kapantzoglou et al (2010) suggereerden dat de bruikbaarheid van observatielijsten nader onderzocht moet worden bij andere doelgroepen, leergebieden, taalgebieden en gebruikers zoals leraren.

Dynamic assessment bleek in verschillende studies een betrouwbare procedure te zijn om onderscheid te maken tussen leerlingen met en zonder taalstoornis, terwijl dat op basis van een statische toets niet mogelijk was (Peña, et al., 1992, 2006; Kester, et al., 2001; Kapantzoglou et al., 2010; Ukrainetz et al., 2000). Het gaat hier om leerlingen die door hun thuisomgeving (lage SES, andere cultuur) of thuistaal (tweedetaalverwerver) dikwijls zwakke scores behalen op een statische toets doordat zij onbekend zijn met de taakeisen. Een voorbeeld daarvan is dat leerlingen vanuit de thuissituatie gewend zijn de woorden naar hun functie te omschrijven (‘je kunt het eten’) in plaats van te benoemen (‘een appel’), wat hun score op een standaard woordenschattaak negatief beïnvloedt. Ook in de Nederlandse onderwijspraktijk hebben leraren dikwijls te maken met anderstalige leerlingen of

leerlingen uit laag sociaal milieu. Inzet van formatieve toetsen zou ertoe kunnen bijdragen dat de leermogelijkheden van deze leerlingen nauwkeuriger in kaart worden gebracht. Het onderscheid tussen leerlingen die wel/geen speciale onderwijsvoorzieningen nodig hebben, is dan beter te maken. De verwachting is dat minder leerlingen zullen worden doorverwezen naar speciaal (basis)onderwijs of schakelklassen, doordat het reguliere onderwijsaanbod beter op de leerlingen kan worden aangepast (Kester et al., 2001).

Hoewel het merendeel van de studies in het basisonderwijs werd uitgevoerd, denken we dat er een vertaalslag te maken is naar andere onderwijssectoren. In de studie van Teo (2012) wordt een voorbeeld gegeven van een digitale leesbegripstoets in het hoger onderwijs, waarbij studenten feedback kregen en zo hun metacognitieve leesstrategieën konden verbeteren. Maar ook andere domeinen zoals mondelinge taalvaardigheid en schrijfvaardigheid komen gedurende de hele schoolloopbaan aan bod, waarbij de doelen met betrekking tot kennis en vaardigheden en daarmee ook de instructiebehoeften veranderen. Door de inzet van formatieve toetsing kunnen leraren beter zicht krijgen op de leer- en instructiebehoeften van leerlingen en studenten op verschillende momenten in de schoolloopbaan. Om deze toetsen betrouwbaar, valide en effectief vorm te geven is nader onderzoek nodig.

4.3 Onderzoeksmethoden, -technieken en -instrumenten

Er is gezocht naar literatuur over effectieve en bruikbare toetsvormen die in het teken staan van het onderwijzen, leren en toetsen van taal- en leesvaardigheid. Vanwege de logistieke en financiële implicaties kunnen potentieel bruikbare toetsvormen feitelijk pas op grote schaal worden geïmplementeerd in het taal- en leesonderwijs als de effectiviteit ervan is vastgesteld. Ons onderzoek beidt nog onvoldoende houvast. Bij de uitkomsten van ons onderzoek zijn namelijk enkele methodologische kanttekeningen te plaatsen die de generaliseerbaarheid van de resultaten en conclusies beperken. De kanttekeningen hebben betrekking op het aantal gevonden studies, de omvang van de steekproeven, de onderzoeksdesigns, de opzet van de interventies, de toetstechnische kwaliteit van de onderzoeksinstrumenten, en de effectmaten die gebruikt worden.

- ***Aantal studies.*** De resultaten van dit onderzoek moeten geïnterpreteerd worden in het licht van de schaarste aan beschikbare studies. We hebben slechts veertien studies gevonden die voldeden aan onze inhoudelijke en methodologische zoekcriteria. Daarbij merken we op dat we onze in- en exclusiecriteria tussentijds versoepeld hebben, omdat we anders nog minder studies overgehouden zouden hebben (zie paragraaf 2.3). Hoewel het onderzoek een aantal potentieel bruikbare toetsvormen heeft opgeleverd, is het aantal gevonden studies te klein om de veronderstelde interacties tussen enerzijds effectiviteit en bruikbaarheid, en anderzijds taalvaardigheid en doelgroep, te kunnen analyseren. Geen van de gevonden studies had als expliciet doel om de differentiële effectiviteit van een nieuwe toetsvorm voor verschillende taaldomeinen of onderwijssegmenten te onderzoeken. Het is met andere woorden problematisch om conclusies te trekken over de relatieve effectiviteit of bruikbaarheid van de gevonden toetsvormen voor het onderwijzen, toetsen en ontwikkelen van taal- en leesvaardigheden bij verschillende groepen leerlingen.
- ***Omvang van de steekproeven.*** Volgens de bekende vuistregel van Whitehurst (2003) vereist het kunnen aantonen van een middelmatig effectieve interventie een steekproef van minimaal 300 individuen. De individuen moeten gelijk verdeeld zijn over de interventie- en controlegroep. Is de klas of school de eenheid, dan zijn ten minste vijftig tot zestig klassen of scholen noodzakelijk. Raudenbush (2003) komt tot vergelijkbare aantallen. De steekproeven in de gevonden studies waren over het algemeen als zeer klein ($N < 60$) aan te merken. Het verdient aanbeveling om vervolgonderzoek te verrichten dat minder kleinschalig is. Pas daarna kunnen onderbouwd suggesties gedaan worden ter verbetering van de toetspraktijk in het taal- en leesonderwijs.
- ***Onderzoeksdesign.*** Een gerandomiseerd en gecontroleerd veldexperiment is het meest geëigende middel om de effectiviteit van een interventie – in ons geval een nieuwe toetsvorm – te onderzoeken (Shadish, Cook & Campbell, 2002; Coalition for Evidence-based policy, 2005; Towne & Hilton, 2004; Togerson et al., 2004; Onderwijsraad, 2006). Een onderzoeksontwerp is gerandomiseerd als de eenheden (de leerlingen, leraren, klassen of scholen) volgens een aselect trekkingsmechanisme aan een interventie- of controlegroep worden toegewezen. Het grote voordeel van *random* toewijzing is dat de groepen bij de start van de interventie in beginsel volledig identiek zijn op zowel bekende en onbekende factoren als gemeten en niet-gemeten factoren. Verschillen tussen beide groepen op de eindmeting kunnen daardoor relatief eenduidig aan de nieuwe toetsvorm worden toegeschreven. Een belangrijk argument voor een gerandomiseerd en gecontroleerd veldexperiment is de hardheid van de conclusies over het succes van de interventie en de geloofwaardigheid en bruikbaarheid van de onderzoeksresultaten voor onder meer beleidsmakers. Van alle veertien gevonden studies is alleen het onderzoek over schrijfvaardigheid van Franzke et al. (2005) opgezet als een gerandomiseerd en gecontroleerd veldexperiment. De overige dertien studies zijn eerder als ontwerponderzoek te

karakteriseren dan als experimenteel effectonderzoek. Leighton (2011) trekt een vergelijkbare conclusie. In dat onderzoek werden slechts 13 van de 300 studies van voldoende kwaliteit geacht met betrekking tot onder andere het onderzoeksdesign.

- ***Interventie en feedback.*** Voor de gevonden toetsvormen geldt dat de afname- en beoordelingscondities minder gestandaardiseerd zijn dan bij de gebruikelijke statische toetsvormen het geval is. Eerder maakten we onderscheid tussen klinische interventies waarin meestal contingente feedback gegeven wordt en onderzoeksgerichte interventies waarin meestal non-contingente feedback gegeven wordt. De bestudeerde studies bevatten een aantal summiere beschrijvingen van potentieel effectieve interventies waarin onderwijzen, leren en toetsen geïntegreerd zijn. Slechts in enkele van de gevonden studies bleken duidelijke richtlijnen voor de implementatie van de interventie beschikbaar te zijn in de vorm van gepubliceerde afnamehandleidingen, protocollen of scripts. In vrijwel alle studies werd de interventie onder “gunstige” omstandigheden uitgevoerd door de onderzoekers zelf of door intensief getraind personeel. Er was zelden sprake van een volledig zelfstandige implementatie door leraren in een reguliere onderwijssetting. Dit is problematisch, omdat leraren ertoe neigen om interventies anders uit te voeren dan de ontwikkelaars bedoeld hebben. Dat is bijvoorbeeld het geval als leraren vasthouden aan oude routines, omdat zij tevreden zijn met hun eigen manier van onderwijzen en toetsen. Onderzoekers dienen daarom te checken of de interventie getrouw uitgevoerd wordt. Slechts in enkele van de gevonden studies hebben de onderzoekers een dergelijke check uitgevoerd. Dit is een ernstige lacune, omdat het nauwkeurig observeren en registreren van de manier waarop leraren de interventie implementeren van essentieel belang is om te begrijpen waarom de interventie al dan niet succes heeft.
- ***Kwaliteit onderzoeksinstrumenten.*** Betrouwbaarheid en validiteit van de onderzoeksinstrumenten is een belangrijk kenmerk van goed uitgevoerd onderzoek. Helaas werden in slechts enkele van de aangetroffen studies gegevens gerapporteerd over de betrouwbaarheid en validiteit van de gebruikte onderzoeksinstrumenten. Zo verstrekten vier studies helemaal geen gegevens over de betrouwbaarheid. Gezien de ogenschijnlijk breed gedeelde bezorgdheid over de validiteit en bruikbaarheid van de gebruikelijke statische toetsen voor bijvoorbeeld leerlingen uit minderheidsgroepen, is dit wellicht begrijpelijk. Met het oog op de interpretatie van onderzoeksresultaten is het echter noodzakelijk om onderzoek te doen naar de betrouwbaarheid en validiteit van de onderzoeksinstrumenten en daarover te rapporteren.
- ***Gebruikte effectmaten.*** In de gevonden studies werd een breed scala aan instrumenten gebruikt om het effect van een toetsvorm aan te tonen. Vaak was niet duidelijk waarom de onderzoekers een bepaald type instrument gebruikten. Ten aanzien van de gebruikte effectmaten maakte we eerder gewag van landelijk genormeerde toetsen, zelfgemaakte criteriumgeoriënteerde toetsen, *dynamic assessment* toetsen en *dynamic assessment* observatielijsten (zie paragraaf 3.1). De effectmaten verschillen in de hardheid van het bewijs dat geleverd wordt. *Dynamic assessment* toetsen hebben de zwakste bewijskracht, omdat ze nauw aansluiten bij de leerstof die in de interventie wordt aangeboden en hiermee hooguit “nabije” transfer kunnen aantonen. Landelijk genormeerde toetsen leveren het sterkste bewijs, omdat het een interventie-onafhankelijke effectmaat betreft die indicatief is voor “verre” transfer. In slechts één studie werd zowel gebruikgemaakt van landelijk genormeerde toetsen als van *dynamic assessment toetsen* en observatielijsten (Fuchs et al., 2011). Het verdient aanbeveling om het effect van veelbelovende nieuwe toetsvormen in vervolgonderzoek met meerdere typen effectmaten te onderzoeken, Hierbij denken wij aan een combinatie van landelijk genormeerde toetsen (verre transfer), *dynamic assessment* toetsen (nabije transfer) en *dynamic assessment* observatielijsten (non-cognitieve vaardigheden). Pas dan kan daadwerkelijk vastgesteld worden in hoeverre nieuwe toetsvormen een positieve bijdrage leveren aan het leerproces van leerlingen en het beslisproces van leraren.

4.4 Aanbevelingen voor vervolgonderzoek

In de literatuur wordt gewezen op de positieve rol die formatieve toetsvormen kunnen spelen bij opbrengstgericht werken. Er zijn tal van studies te noemen waarin een nieuwe toetsvorm met succes lijkt te zijn ingezet. Het aantal studies dat zich specifiek richt op het taal- en leesonderwijs en daarbinnen op de domeinen technisch lezen, begrijpend lezen, woordenschat, strategisch schrijven, en mondelinge taalvaardigheid is echter zeer beperkt. Bovendien bevindt het onderzoek naar de effectiviteit en bruikbaarheid van de alternatieve toetsvormen ter verbetering van het taalleren zich nog in een pril stadium. Met uitzondering van de studie van Franzke et al. (2005) zijn de aangetroffen studies op het gebied van het taal- en leesonderwijs eerder als oriënterend ontwerponderzoek te karakteriseren dan als (quasi-) experimenteel effectonderzoek. Er wordt bijvoorbeeld nauwelijks aandacht besteed aan de contextuele randvoorwaarden waarbinnen formatieve toetsingen succesvol functioneren (Johnson & Burdett, 2010). Aangezien de meeste studies in een gecontroleerde omgeving uitgevoerd worden door speciaal getrainde toetsleiders bestaat het risico dat de positieve effecten van formatieve toetsingen te sterk worden aangezet (Pryor & Torrance, 1998). Mogelijk zijn de effecten minder positief als het onderzoek plaatsvindt in de

drukte van de alledaagse onderwijspraktijk. Het is essentieel dat uitgezocht wordt onder welke omstandigheden, en bij welke doelgroep(en), een toets- en feedbackvorm effectief is.

Vervolgonderzoek wordt bij voorkeur uiteengelegd in vier fasen. In de eerste fase kunnen veelbelovende toetsvormen in samenwerking met onder meer leraren, leerplanontwikkelaars en toetsdeskundigen in kleinschalige pilots op uitvoerbaarheid en effectiviteit getest worden (Van den Akker, 1999; Van den Akker, Gravemeijer, McKenney & Nieveen, 2008). Op het moment dat de nieuwe toetsvorm voldoende uitgekristalliseerd en doorontwikkeld is, dient de tweede fase zich aan. In deze fase dienen de uitvoerbaarheid en effecten in minder kleinschalige settings onder “gunstige omstandigheden” te worden vastgesteld. Het kan bijvoorbeeld gaan om een situatie waarin de onderzoekers de implementatie zorgvuldig regisseren en controleren, er gewerkt wordt met hoog opgeleide en enthousiaste leraren, en er veel tijd en geld beschikbaar is voor materiaalontwikkeling en scholing van leraren (zie bijvoorbeeld Raudenbush, 2003). Nadat de uitvoerbaarheid en effectiviteit in kleinschalige settings zijn vastgesteld, kan het onderzoek worden opgeschaald. In de derde fase testen de onderzoekers de nieuwe toetsvorm dan ook uit bij andere leraren, bij andere groepen leerlingen en eventueel bij andere leerstofgebieden. Pas na afronding van deze fase is “hard” onderzoek naar de relatieve effectiviteit van de nieuwe toetsvorm in vergelijking met andere (traditionele) toetsvormen op zijn plaats. De studies die in dit onderzoek bestudeerd zijn, bevonden zich in de eerste of de tweede onderzoeksfase.

Doordat het ontwerp- en effectonderzoek op het gebied van (formatieve) taaltoetsing zich nog in de beginfase bevindt, kan één van de belangrijkste onderzoeksvragen momenteel niet beantwoord worden. Het is onbekend in hoeverre de onder gunstige omstandigheden vastgestelde (positieve) effecten stand houden zonder dat aan het karakter van de toets- en feedbackvorm afbreuk wordt gedaan. Bij het “exporteren” van de toetsvorm naar de onderwijspraktijk kunnen zich allerlei problemen voordoen. Zo heeft de onderzoeker in de nieuwe situatie meestal minder controle over de implementatie van de toetsvorm, zijn de leraren vaak minder enthousiast, zijn er minder faciliteiten beschikbaar voor scholing en begeleiding en moeten leraren het materiaal dikwijls aan hun specifieke omstandigheden aanpassen. Dit kan de effecten van de innovatie nadelig beïnvloeden. Aangezien er op dit moment geen overtuigend bewijs is voor de effectiviteit van nieuwe vormen van taaltoetsing, en ook niet is aangetoond dat de toetsvormen bruikbaar zijn in een brede range van toepassingssituaties, is de tijd nog niet rijp om de nieuwe toetsvormen op grote schaal te verspreiden in het taal- en leesonderwijs. Vanuit internationaal perspectief bezien lijkt het taal- en leesonderwijs het meeste baat te hebben bij onderzoeken die zich richten op fase drie en vier. Het gaat dan om grootschaligere onderzoeken die de bruikbaarheid en effectiviteit van veelbelovende toetsvormen analyseren in verschillende praktijksituaties en leerstofgebieden. Het is zeer de vraag of dit onderzoek ook al in Nederland kan plaatsvinden. Internationale studies zijn vermoedelijk niet één-op-één te vertalen naar de Nederlandse situatie. Daarom lijkt het verstandig om eerst op basis van relevante internationale literatuur (zie Hoofdstuk 3) enkele relatief kleinschalige, praktijkgerichte ontwerponderzoeken in het Nederlandse onderwijs uit te zetten. Voor de veelbelovende toets- en feedbackvormen kan vervolgens in een gerandomiseerd en gecontroleerd veldexperiment op zoek gegaan worden naar de gewenste “harde” effectiviteitsbewijzen.

De generaliseerbaarheid van uitkomsten over doelgroepen zou zowel in ontwerp- als effectonderzoeken aandacht moeten krijgen. Over het algemeen worden studies uitgevoerd bij een specifieke leeftijdsgroep. Dat was ook het geval bij de studies die we, met het oog op de beantwoording van de onderzoeksvraag, bestudeerd hebben. Vooral de leerlingen die aan het begin van de schoolloopbaan staan, worden opvallend vaak bij een onderzoek betrokken. Het gaat dan onder meer om onderzoek naar *dynamic assessment* bij jonge leerlingen in het basisonderwijs (e.g. Elleman, Compton, Fuchs, Fuchs & Bouton, 2011; Fuchs et al., 2011; Sainsbury & Benton, 2011; Burton & Watson, 2007). Het hoger onderwijs is ondervertegenwoordigd. Er wordt weliswaar gewezen op de waarde die formatieve toetsingen kunnen hebben in het hoger onderwijs (e.g., Carless, 2006; Walker, 2009; Weaver, 2006), maar er wordt weinig onderzoek verricht in dit onderwijstype. De studies die beschikbaar zijn, hebben in veel gevallen betrekking op het meten van spreek- en luistervaardigheden in één-op-één context (e.g., Antón, 2009; Oskoz, 2005; Poehner, 2005; Ableeva & Lantolf, 2011). Als het om het meten van schrijfvaardigheid gaat, richten onderzoekers zich in de regel op de constructie en analyse van beoordelingsschalen, en niet op de relatie tussen de schrijftoets en de ontwikkeling van schrijfvaardigheid (Huot, 2002). Teo (2012) is één van de weinigen die in het hoger onderwijs onderzoek heeft gedaan naar de effectiviteit van een *dynamic assessment* voor begrijpend lezen. Het is belangrijk om bij verschillende doelgroepen onderzoek te doen naar de effectiviteit van een toets- en feedbackvorm. Toets- en feedbackvormen die effectief zijn voor beginnende leerlingen kunnen later in de schoolloopbaan, als leerlingen vaardiger zijn, immers ineffectief worden (Kalyuga, Ayres, Chandler, & Sweller, 2003; Kalyuga, Chandler, & Sweller, 2001). Momenteel is niet te zeggen in hoeverre de uitkomsten van de gevonden studies te generaliseren zijn over leeftijdsgroepen. Het is wenselijk om de toets- en feedbackvormen in toekomstig onderzoek te relateren aan beschikbare leerlijnen voor het taal- en leesonderwijs. Mogelijk kunnen de dynamische toetsingen die nu ingezet worden bij jonge leerlingen met taal- en leesproblemen ook in andere onderwijssectoren helpen bij het bepalen welke leerlingen extra ondersteuning nodig hebben om de einddoelen te behalen. Dit wordt bij voorkeur eerst onderzocht in enkele kleinschalige (Nederlandse) pilotstudies en daarna in een zo dringend gewenst gerandomiseerd veldexperiment met ten minste 300 deelnemers.

Literatuur

Aarnoutse, C., Verhoeven, L., Zandt, R. van het, & Biemond, H. (2003). *Tussendoelen gevorderde geletterdheid, leerlijnen voor groep 4 tot en met 8*. Nijmegen: Expertisecentrum Nederlands.

Ackerman, B. P., Jackson, M., & Sherill, L. (1991). Inference modification by children and adults. *Journal of Experimental Child Psychology*, *52*, 166–196.

Akker, J. van den (1999). Principles and methods of development research. In: J. van den Akker, R., Branch, K., Gustafson, N., Nieveen & Tj. Plomp (Eds.), *Design approaches and tools in education and training* (pp. 1–14). Dordrecht: Kluwer.

Akker, J. van den, Gravemeijer, K., McKenney. S., & Nieveen, N. (2008). Introduction to educational design research. In: Van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (Eds.), *Educational design research* (pp. 3-7). London: Routledge.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the ‘ Common European Framework. *Language Testing, 22*, 301–320.

Al-Kadri, H.M., Al-Moamary, M.S., Roberts, C., & van der Vleuten, C.P.M. (2012). Exploring assessment factors contributing to students' study strategies: Literature review. *Medical Teacher, 34,* 42–50.

Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development, 58*, 10 (Serial No. 238).

Arter, J.A. (2003). *Assessment for learning: Classroom assessment to improve student achievement and well-being*. ERIC Documents (2002 ERIC Document Reproduction Service No. ED 480 068).

Attali, Y. & Burstein, J. (2006). *Automated Essay Scoring With e-rater® V.2. Journal of Technology, Learning, and Assessment, 4*, 3. Available from http://www.jtla.org

Baron, J. (2007). *Checklist for reviewing a randomized controlled trial of a social program or project, to assess whether it produced valid evidence*. Online beschikbaar via: http://coalition4evidence.org/wp-content/uploads/uploads-dupes-safety/Checklist-For-Reviewing-a-RCT-Jan10.pdf.

Barret, M., Zachman, L., & Huisingh, R. (1988). *Assessing semantic skills through everyday themes*. Moline: IA: Lingui-systems.

Birenbaum, M., Kimron, H. & Shilton, H. (2011). Nested contexts that shape assessment for learning: School-based professional learning community and classroom culture. *Studies in Educational Evaluation, 37*, 35-8.

Black, P. & Wiliam, D. (1998a). *Inside the black box: raising standards through classroom assessment*. London: School of Education, King’s College.

Black, P. & Wiliam, D. (1998b). Assessment and classroom learning. *Assessment in Education*, *5*, 1, 7–74.

Bosma, T., Hessels, M.G.P. & Resing, W.C.M. (2012). Teachers’ preference for educational planning: Dynamic testing, teaching’ experience and teachers’ sense of efficacy. *Teaching and Teacher Education, 28,* 560-567.

Bransford, J.D., Delclos, V., Vye, N., Burns, S. & Hasselbring, T. (1987). Approaches to dynamic assessment: Issues, data and future directions. In C. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potentials* (pp. 479-495). New York: Guilford Press.

Broadfoot, P. M., Daugherty, R., Gardner, J., Harlen, W., James, M., & Stobart, G. (2002). *Assessment for learning: 10 principles*. Cambridge, UK: University of Cambridge School of Education.

Brown, C. A. (2002). *Portfolio Assessment: How Far Have We Come?* ERIC Documents (2002 ERIC Document Reproduction Service No. ED 477941).

Budoff, M., Gimon, A., & Corman, L. (1976). Learning potential measurement with Spanish-speaking youth as an alternative to IQ tests: A first report. *Interamerican Journal of Psychology, 8,* 233–246.

Budoff, M., Meskin, J., & Harrison, R. H. (1971). Educational test of the learning-potential hypothesis. *American Journal of Mental Deficiency, 76,* 159–169.

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113-121). Mahwah,NJ: Lawerence Erlbaum Associates.

Byrne, B., Fielding-Barnsley, R., & Ashley, L. (2000). Effects of preschool phoneme identity training after six years: Outcome level distinguished from rate of response. *Journal of Educational Psychology, 92,* 659–667.

Caffrey, E., Fuchs, D., & Fuchs, L.S. (2008). The predictive validity of dynamic assessment: A review. *Journal of Special Education, 41*, 254-270.

Caccamise, D., Franzke, M., Eckhoff, A., Kintsch, E., & Kintsch, W. (2007). Guided practice in technology-based summary writing. In D. S. Mc-Namara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 375–396). Mahwah, NJ: Lawrence Erlbaum Associates.

Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), *Dynamic testing* (pp. 82–115). New York, NY: Guilford.

Campione, J. C., Brown, A. L., Ferrara, R. A., Jones, R. S., & Steinberg, E. (1985). Breakdowns in flexible use of

information: Intelligence-related differences in transfer following equivalent learning performance. *Intelligence, 9*, 297–315.
Campione, J. C. (1989). Assisted assessment: A taxonomy of approaches and an outline of strengths and weaknesses. *Journal of Learning Disabilities, 22,* 151-165.
Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education, 31*, 219–233.
Carlson, J.S., & Wiedl, K.H. (1978). Use of testing-the-limits procedures in the assessment of intellectual capabilities in children with learning difficulties. *American Journal of Mental Deficiencies, 82*, 6, 559-564.
Carlson, J.S., & Wiedl, K.H. (1979). Toward a differential testing approach: Testing-the-limits employing the Raven matrices. *Intelligence, 3*, 4, 323–344.
Carlson, J. S., & Wiedl, K. H. (1992). The dynamic assessment of intelligence. In H. C. Hol lywood & D. Tzuriel (Eds.), *Interactive assessment* (pp. 167–186). Berlin, Germany: Springer-Verlag.
Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language processing deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech*, *Language*, *and Hearing Research*, *49*, 278–293.
Catts, H. W., Compton, D., Tomblin, J. B., & Bridges, M. (2012). Prevalence and nature of late emerging reading disabilities. *Journal of Educational Psychology*, *104*, 166-181.
Chan, Y. C. (2006). Elementary school EFL teachers' beliefs and practices of multiple assessment. *Reflections on English Language Teaching, 7*, 37–62.
Coalition for evidence-based policy (2005). *Key items to get right when conducting a randomised controlled trial in education*. Online beschikbaar via: http://coalition4evidence.org/wp-content/uploads/2012/05/Guide-Key-items-to-Get-Right-RCT.pdf
Compton, D. L., Fuchs, D., Fuchs, L. S., Elleman, A. M., & Gilbert, J. K. (2008). Tracking children who fly below the radar: Latent transition modeling of students with late-emerging reading disability. *Learning and Individual Differences*, *18*, 329–337.
Cooper, H.M. (1998). *Research synthesis and meta-analysis: A step-by-step approach*. Thousands Oaks, CA: Sage.
Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education, 6*, 101-116.
Graham, S., & Harris, K. R. (2005). Improving the writing performance of young struggling writers: Theoretical and programmatic research from the Center on Accelerating Student Learning. *Journal of Special Education, 39*, 19–33.
Craig, D.V. (2001). *Alternative, dynamic assessment for second language learners*. ERIC Documents (2001 ERIC Document Reproduction Service No. 453 691).
Crisp, G. T. (2012). Integrative assessment: Reframing assessment practice for current and future learning. *Assessment & Evaluation in Higher Education, 37*, 33–43.
Cronen, S., Silver-Pucuilla, H., & Condelli, L. (2006). *Conducting large-scale research in adult ELS: Challenges and approaches for the explicit literacy impact study*. Washington, DC: American Institutes for Research.
Crook, C. (1991). Computers in the zone of proximal development: Implications for evaluation. *Computers and Education 17,* 1, 81–91.
Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record, 106*, 1047-1085.
Dillon, R. F. (1997). Dynamic testing. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 164– 186). Westport: Greenwood Press.
Dlaska, A., & Krekeler, C. (2008). Self-assessment of pronunciation. *System, 36*, 506–516.
Dörfler, T., Golke, S., & Artelt, C. (2009). Dynamic assessment and its potential for the assessment of reading competence. *Studies in Educational Evaluation, 35,* 77–82.
Ehrlich, M., Remond, M., & Tardieu, H. (1999). Processing of anaphoric devices in young skilled and less skilled comprehenders: Differences in metacognitive monitoring. *Reading and Writing: An Interdisciplinary Journal*, *11*, 29–63.
Elleman, A.M., Fuchs, D.L., Fuchs, D., Fuchs, L.S., & Bouton, B. (2011). Exploring dynamic assessment as a means of identifying children at risk of developing comprehension difficulties. *Journal of Learning Disabilities, 44, 4,* 348-357.
Elliot, J. (2003). Dynamic assessment in educational settings: Releasing potential. *Educational Review, 55, 1,* 15-32.
Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.
Feenstra, H. (2014). *Assessing writing ability in primary education* (proefschrift). Enschede: Ipskamp Drukkers.
Ferster, B., Hammond, T.C., Alexander, R.C., & Lyman, H. (2012). Automated formative assessment as a tool to scaffold student documentary writing. *Journal of Interactive Learning Research, 23,* 1, 81-99.
Feuerstein, R. (1979). *The dynamic assessment of retarded performers*. Baltimore: University Park Press.
Feuerstein, R., Rand, Y., & Hoffman, M. B. (1979). *The dynamic assessment of retarded per formers: The Learning Potential Assessment Device*. Baltimore: University Park Press.
Feuerstein, R., Rand, Y., & Rynders, J. E. (1988) *Don't accept me as I am. Helping retarded performers excel*.

New York: Plenum.
Feuerstein, R., Feuerstein, R. S., & Falik, L. H. (2010). *Beyond smarter: Mediated learning and the brain's capacity for change*. New York, NY: Teachers College Press.
Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary Street: Computer support for comprehension and writing. *Journal of Educational Computing Research*, 33, 53–80.
Fuchs, D. Compton, D.L., Fuchs, L.S., Bouton, B. & Caffrey, E. (2011). The construct and predictive validity of a dynamic assessment of young children learning to read: Implications for RTI frameworks. *Journal of Learning Disabilities, 44(4),* 339-347.
Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68,* 1–76.
Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*, 6–10.
Graham, S., & Harris, K. R. (2005). Improving the writing performance of young struggling writers: Theoretical and programmatic research from the Center on Accelerating Student Learning. *Journal of Special Education, 39*, 19–33.
Greenleaf, C., Gee, M. K., & Ballinger, R. (1997). *Authentic Assessment: Getting Started*. ERIC Documents (1997 ERIC Document Reproduction Service 411 474).
Grigorenko, E. (2009). Dynamic Assessment and response to intervention: Two sides of one coin. *Journal of Learning Disabilities, 42(2)*, 111-132.
Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin, 124,* 75–111.
Gutiérrez-Clellen, V. F. (1996). Language diversity: Implications for assessment. In K. N. Cole, P. S. Dale, & D. J. Thal (Eds.), *Assessment of communication and language* (Vol. 6; pp. 29–56). Baltimore, MD: Brookes.
Gutierrez Clellen, V., & Iglesias, A. (1987, November). Expressive vocabulary of kindergarten and first grade Hispanic students. Paper presented at the American Speech-Language-Hearing Association national convention, New Orleans.
Gysen, S., Van Avermaat (2005). Issues in functional language performance assessment: the case of the certificate Dutch as a foreign language. *Language Assessment Quarterly, 2*, 51–68.
Hagenaars, J. & McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.
Handley, K. & Williams, L. (2009) From copying to learning? Using exemplars to engage students with assessment criteria and feedback, *Assessment and Evaluation in Higher Education, 34*, 95–108,
Harlen, W. (2005). Teachers' summative practices and assessment for fearning: Tensions and synergies. *Curriculum Journal, 16*, 207-223.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*,81–112.
Hearst, M. (2000). The debate on automated essay grading. *IEEE Intelligent Systems, 15*, 22-37.
Heath, S.B. (1983). *Ways with words: Language, life, and work in communities and classrooms.* Cambridge, England: Cambridge University Press.
Heath, S.B. (1986). Sociocultural contexts of language development. In California State Department of Education (1986), *Beyond language: Social and cultural factors in schooling language minority children* (pp. 143-186). Los Angeles: Evaluation Dissemination and Assessment Center, California State University.
Hendriks, P., & Schoonman, W. (Eds.). (2006). *Handboek assessment deel 1: Gedragsproeven*. Assen: Van Gorcum.
Huff, K. & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: theory and applications* (pp. 19–60). New York: Cambridge University Press.
Johnson, S. D., & Wentling, T. L. (1996). An alternative vison for assessment in vocational teacher education. In N. K. Hartley & T. L. Wentling (Eds.), *Beyond tradition: Preparing the teachers of tomorrow's workforce* (pp. 147-166). Columbia: University of Missouri.
Ingram, D.,Louis, K. S. & Schroeder, R.G. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. Teachers College Record, 106, 1258–1287.
Kaniel, S., Tzuriel, D., Feuerstein, R., Ben-Shacher, N., & Eitan, T. (1991). Dynamic assessment: Learning and transfer of Ethiopina immigrants to Isreal. In R. Feuerstein, P. Klein, & A. Tannenbaum (Eds.), *Mediated learning experience* (pp. 197-209). London, England: Freund.
Kalyuga, S. (2003). Rapid assessment of learners' knowledge in adaptive learning environments. In U Hoppe, F. Verdejo, & J. Kay (Eds.), *Artificial intelligence in education: Shaping the future of learning through intelligent technologies* (pp. 167–174). Amsterdam: IOS Press.
Kalyuga, S. (2004, April). *Rapid dynamic assessment of expertise to manage cognitive load during instruction.* Paper presented at the 2004 Annual Meeting of the American Educational Research Association (AERA), San Diego, CA.
Kalyuga, S. (2007). Rapid assessment of Learners' Proficiency: A cognitive load approach.*Educational Psychology, 26, 6, 735–749.*
Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). Expertise reversal effect. *Educational Psychologist, 38*, 23–31.

Kalyuga, S., Chandler, P., & Sweller, J. (2001). Learner experience and efficiency of instructtional guidance. *Educational Psychology, 21,* 5–23.
Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology, 96,* 558–568.
Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology, Research and Development, 53,*83–93.
Kaniel, S., Tzuriel, D., Feuerstein, R., Ben-Shacher, N., & Eitan, T. (1991). Dynamic assessment: Learning and transfer of Ethiopian immigrants to Israel. In R. Feuerstein, P. Klein, & A. Tannenbaum (eds.), *Mediated learning experience* (pp. 179-209). London, England: Freund.
Kapantzoglou, M., Restrepo, M.A., & Thompson, M.S. (2010). Dynamic assessment of word learning skills: Identifying language impairment in bilingual children. *Language, Speech, and Hearing Services in Schools, 43, 1*, 81-96.
Keeley, P., & Tobey, C. R. (2011). *Mathematics formative assessment*. Thousand Oaks, CA: Corwin.
Kester, E.S. Peña, E.D., & Gillam, R.B. (2001). Outcomes of dynamic assessment with culturally and linguistically diverse students: A comparison of three teaching methods within a test-teach-retest framework. *Journal of Cognitive Education and Psychology, 2,* 37-54.
Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition, 2,* 15–47.
Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119,* 2, 254–284.
Koh, L.C. (2008). Refocusing formative feedback to enhance learning in pre-registration nurse education. *Nurse Education in Practice, 8,* 223-230.
Kornhaber, M. L. (2004). Appropriate and inappropriate forms of testing, assessment, and accountability. *Educational Policy, 18*, 45-70.
Kramer, K., Mallett, P., Schneider, Ph., & Hayward, D. (2009). Dynamic assessment of narratives with Grade 3 children in a first nations community. Revue canadienne d'orthophonie et d'audiologie, 33, 3, 119-128.
Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*, 279-308.
Landauer, T.K. Lochbaum, K.E. & Dooley, S. (2009). A New Formative Assessment Technology for Reading and Writing. *Theory Into Practice, 48,* 44-52
Larsen, J.A., & Nippold, M.A. (2007). Morphological analysis in school-age children: Dynamic assessment of a word learning strategy. *Language, Speech and Hearing Services in Schools, 38,* 201-2012.
Laurier, M. (2004). Evaluation and multimedia in second-language learning. *ReCALL, 16,* 475-487
Leach, J., Scarborough, H., & Rescorla, L. (2003). Late-emerging reading disabilities. *Journal of Educational Psychology*, *95*, 211–224.
Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: Minute by minute, day by day. *Educational Leadership, 63*, 18-24.
Leighton, J. P. & Gierl, M. J. (2007).Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: theory and applications* (pp. 3–18). New York: Cambridge University Press.
Lidz, C.S. (1987). *Dynamic assessment: An interactional approach to evaluating learning potential*. New York, NY: Guilford Press.
Lidz, C.S. (1991). *Practitioner's guide to dynamic assessment*. New York, NY: Guilford Press.
Lidz, C. S., & Peña, E. D. (1996). Dynamic assessment: The model, its relevance as a non-biased approach and its application to Latino American preschool children. *Language, Speech, and Hearing Services in Schools*, 27, 367-372.
Losardo, A., & Notari-Syverson, A. (1995). What children can say. How do we find out? *American Speech-Language-Hearing Association Special Interest Divisions: Language Learning and Education, 2*, 6-12.
Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction, 13,* 251–283.
Marshall, S. (1995). Some suggestions for alternative assessments. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 431–453). Hillsdale, NJ: Lawrence Erlbaum Associates.
Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*, 75-100.
McGill-Frantzen, A., & Allington, R. (2006). Contamination of current accountability systems. *Phi Delta Kappan, 87*, 762-766.
Meestringa, T., Ravesloot, C. & de Vries, H. (2010). *Concretisering referentieniveaus schrijven en lezen in het voortgezet onderwijs.* Enschede: SLO.
Meijer, J. (1993). Learning potential, personality characteristics, and test performance. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological, and practical issues* (pp. 341–362). Lisse, Netherlands: Swets & Zeitlinger B.V.

Mehrens, W. A. (2002). Consequences of assessment: What is the evidence? In G. Tindal (Ed.), *Large-scale assessment programs for all students: validity, technical adequacy, and implementation* (pp. 149-177). Mahwah, N.J.: Erlbaum.
Miller, L., Gillam, R., & Peña, E. (2001). *Dynamic assessment and intervention: Improving children's narrative abilities.* Austin, TX: PRO-ED.
Muter, V., Hulme, C., Snowling, M.J., Stevenson, J. (2004). Phonemes, rimes,vocabulary and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Developmental Psychology, 40 (5)*, 665-681.
Onderwijsraad (2006). *Naar meer evidence-based onderwijs*. Den Haag: Drukkerij Artoos.
Onderwijsraad (2011). *Advies om de kwaliteit van het beroepsonderwijs*. Den Haag: Onderwijsraad
Ouellette, G.P. (2006). What's meaning got to do with it: the role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology, 98 (3),* 554-566.
Palincsar, A. S., Brown, A. L., & Campione, J. C. (1994). Models and practices of dynamic assessment. In G. P.
Wallach & K. G. Butler (Eds.), *Language learning disabilities in school-age children and adolescents: Some principles and applications* (pp. 132–144). New York: MacMillan
Parrett, W., & Budge, K. (2009). Tough questions in assessment. *Educational Leadership,* 67, 2, 22-27.
Pellegrino, J. (2008). Assessment for learning: Using assessment formatively in classroom instruction. *International Journal of Psychology, 43*(3-4), 381-381.
Peña, E. (1993). *Dynamic assessments: a nonbiased approach for assessing the language of young children*. Unpublished doctoral dissertation. Austin, TX: Temple University.
Peña, E.D. (2000). Measurement modifiability in children from culturally and linguistically diverse backgrounds. *Communication Disorders Quarterly, 21*, 2, 87-97.
Peña, E.D., Gillam, R.B., Malek, M., Ruiz-Felter, R., Resendiz, M., Fiestas, C., & Sabel, T. (2006). Dynamic assessment of school-age children's narrative ability: An experimental investigation of classification accuracy. *Journal of Speech, Language, and Hearing Research, 49*, 1037–1057.
Peña, E.D., Iglesias, A., & Lidz, C.S. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology, 10,* 138-154.
Peña, E., Quinn, R., & Iglesias, A. (1992). The application of dynamic methods to language assessment: A non-biased procedure. *The Journal of Special Education, 26*, 269-280.
Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues & Practice, 28*, 5-13.
Plante, E.D., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25*, 15–24.
Pressley, M., & Afflerback, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Erlbaum.
Punt, L. & de Krosse, H. (2012). Interactief lees- en schrijfonderwijs. *Werken met tussendoelen in de onderbouw van het vo*. Nijmegen: Expertisecentrum Nederlands.
Randolph, J. J. (2009). A guide to writing the dissertation literature review. *Practical Assessment, Research & Evaluation, 14,* 13. Online available from http://pareonline.net/pdf/v14n13.pdf
Raudenbush, S.W. (2003). Designing field trials of educational educations. Paper prepared for the national invitational conference 'Conceptualizing scale-up: Multidisciplinary perspectives'. Online available from http://www.ssicentral.com/hlm/techdocs/DRDC.pdf
Resing, W.C.M. (1993). Measuring inductive reasoning skills: The construction of a learning potential test. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological, and practical issues* (pp.219–242). Lisse, Netherlands: Swets & Zeitlinger B.V.
Resing, W.C.M. (2006). *Zicht op potentieel. Over dynamisch testen, variabiliteit in oplossingsgedrag en leerpotentieel van kinderen (*Oratie). Universiteit Leiden: Leiden.
Resing, W.C.M. (1997). Leerpotentieel onderzoek: wat is de meerwaarde? In T. Engelen-Snaterse & R. Kohnstamm (Eds.), *Kinder- en jeugdpsychologie. Trends* (pp. 283-303). Lisse: Swets & Zeitlinger.
Resing, W.C.M., Ruijssenaars, A.J.J.M., & Bosma, T. (2002). Dynamic Assessment: using measures for learning potential in the diagnostic process. In G. M. Van der Aalsvoort, W. C. M. Resing, & A. J. J. M. Ruijssenaars (Eds.), *Advances in cognition and educational practice: Vol. 7. Learning potential assessment and cognitive training: Actual research and perspectives on theory building and methodology* (pp. 29-64). New York: Elsevier.
Ross, S. (1998). Self-assessment in second language testing : A meta-analysis and analysis of experiential factors. *Language testing 15*, 1, 1-20.
Rushton, A. (2005). Formative assessment: A key to deep learning? *Medical Teaching, 27*, 509-513.
Sainsbury, M. & Benton, T. (2011). Designing a formative e-assessment: Latent class analysis of early reading skills. *British Journal of Educational Technology, 42,* 500-514.
Sanders, P. (red.). *Toetsen op school*. Arnhem: Cito.
Schuurs, U., & Verhoeven. L. (2010). *Meten van leerprestaties in het (v)mbo: assessment for learning en assessment of learning*. Nijmegen: Expertisecentrum Nederlands.

Schuwirth, L.W. & van der Vleuten, C. (2004). Merging views on assessment. *Medical Education, 38*, 1208-1210.
Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
Shanahan, T. (2005). *The national reading panel report: Practical advice for teachers*. Naper ville, IL: Learning Point Associates/North Central Regional Educational Laboratory (NCREL). (ERIC Document Reproduction Service No. ED 489 535).
Shepard, L. A. (2002). The hazards of high-stakes testing. *Issues in Science and Technology, 19*, 53-58.
Shrestha, P, & Coffin, C. (2012). Dynamic assessment, tutor mediation and academic writing development. *Australasian Journal of Special Education, 35,* 137-172.
Schute, V.J. (2007). *Focus on formative feedback*. Princeton, NJ: Educational Testing Service.
Sluijsmans, D., Joosten-ten Brinke, D., & Vleuten, C van der (2013). *Toetsen met leerwaarde. Een reviewstudie naar de effectieve kenmerken van formatief toetsen*. Maastricht: Universiteit Maastricht.
Smith, K., & Tillema, H. (2007). Use of Criteria in Assessing Teaching Portfolios: *Judgemental practices in summative evaluation. Scandinavian Journal of Educational Research, 51*, 103-117.
Sternberg, R. J., & Grigorenko, E .L. (2001). All testing is dynamic testing. *Issues in Education, 7*, 137-170.
Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge: Cambridge University Press.
Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. London: Routledge.
Straetmans, G. J. J. M. (2006). *Bekwaam beoordelen en beslissen. Beoordelen in competentiegerichte beroepsopleidingen* (lectorale rede). Enschede: Saxion Hogescholen.
Swanson, H. L. (1994). The role of working memory and dynamic assessment in the classification of children with learning disabilities. *Learning Disabilities Research and Practice, 9,* 190–202.
Swanson, H. L. (1995). Effects of dynamic testing on the classification of learning disabilities: The predictive and discriminant validity of the Swanson-Cognitive Processing Test (S-CPT). *Journal of Psychoeducational Assessment, 13,* 204–229.
Swanson, H. L., & Lussier, C. M. (2001). A selective synthesis of the experimental literature on dynamic assessment. *Review of Educational Research, 71*, 321-363.
Teo, A. , & Jen, F. (2012). Promoting EFL students' inferential reading skills through computerized dynamic assessment. *Language Learning & Technology, 16, 3,* 10-20.
Torgerson, C., Brooks, G., Porthouse, J., Burton, M., Robinson, A., Wright, K. & Watt, I. (2004). *Adult literacy and numeracy interventions and outcomes: A review of controlled trials*. London: Institute of Education.
Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising new modes of assessments: In search of qualities and standards* (pp. 55-87). Dordrecht: Kluwer Academic Publishers.
Towne, L., & Hilton, M. (2004). *Implementing randomised field trials in education: Report of a workshop* (Washington, DC, September 24, 2003). Washington, DC: The National Academies Press.
Tzuriel, D. (2001). *Dynamic assessment of young children*. New York, NY: Kluwer Academic/Plenum.
Ukrainetz, T.A. Harpell, S. Walsh, C. & Coyle, C. (2000). A preliminary investigation of dynamic assessment with native American kindergartners. *Language, Speech, and Hearing Services in Schools, 31, 2,* 142-154.
Van der Aalsvoort, G.M., Resing, W.C.M., & Ruijssenaars, J.J.M. (Eds.), *Advances in cognition and educational practice: Vol. 7. Learning potential assessment and cognitive training: Actual research and perspectives on theory building and methodology* (pp. 29-64). New York: Elsevier.
Van der Kleij, F.M. (2013). *Computer-based feedback in formative assessment*. Enschede: Universiteit Twente.
Van de Mosselaer, H. & Heylen, L. (2002). Toets- en assessmentbeleid op opleidingsniveau en praktische richtlijnen. In: F. Dochy, L. & H. van de Mosselaer (Eds.), Assessment in onderwijs. Nieuwe toetsvormen en examinering in studentgericht onderwijs en competentiegericht onderwijs (pp. 225-248). Utrecht: Lemma.
Verhoeven, L. & Aaarnoutse, C. (1999). *Tussendoelen beginnende geletterdheid, een leerlijn voor groep 1 tot en met 3.* Nijmegen: Expertisecentrum Nederlands.
Verhoeven, L., Biemond, H. & Litjens, P. (2007). *Tussendoelen mondelinge communicatie, een leerlijn voor groep 1 tot en met 8.* Nijmegen: Expertisecentrum Nederlands.
Verhoeven, L., & van Leeuwe, J. (2011). Vocabulary growth and reading development across the elementary school years. *Scientific studies of reading, 15 (1)*, 8-25.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*, 333–362.
Walker, M. (2009). An investigation into written comments on assignments: Do students find them usable? *Assessment and Evaluation in Higher Education, 34*, 67–78.
Wayman, J. C., Spikes, D. D., & Volonnino, M. (2013). Implementation of a data initiative in the NCLB era. In K.

Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in education: Challenges and opportunities* (pp. 135–153).
Wesson, C.J. (2013). Introducing patchwork assessment to a social psychology module: The utility of feedback. *Psychology Teaching Review, 19, 2,* 97-105.
Whitehurst, W.J. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: U.S. Department of Education Institute of Education Sciences.
Whitehead, D. (2007). Literacy Assessment Practices: Moving from Standardised to Ecologically Assessments in Secondary Schools. *Language and Education, 21*(5),434-452.
Wiliam, D. (2004). *Keeping learning on track: integrating assessment with instruction*. Invited address to the 30th annual conference of the International Association for Educational Assessment, Philadelphia, PA.
Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*, 3–14.
Wools, S., Sanders, P. F., Eggen, T. J. H. M., Baartman, L. K. J., & Roelofs, E. (2011). Evaluatie van een beoordelingssysteem voor de kwaliteit van competentieassessments. *Pedagogische studiën, 88*, 23-40.

Bijlage

Tabel 1 Technische leesvaardigheid

Niveau	Omschrijving
Tussendoelen beginnende geletterdheid groep 1-3	***Start.*** Kent de meeste letters; kan de letters fonetisch benoemen; kan klankzuivere woorden (km, mk, mkm) ontsleutelen zonder eerst de afzonderlijke letters te verklanken. ***Vervolg.*** Kan klankzuivere woorden lezen (mmkm, mkmm en mmkmm); leest woorden met afwijkende spellingpatronen en meerlettergrepige woorden; maakt gebruik van groot scala aan woordidentificatietechnieken; herkent veel woorden automatisch.
Tussendoel gevorderde geletterdheid groep 4-5 en groep 6-8	Herkent lettercombinaties en letterpatronen (o.a. sch, au, ui, oei); herkent lettergrepen in geschreven woorden (o.a. horen, lelijk); herkent het letterpatroon van leenwoorden; maakt gebruik van de betekenis van een woord (feest – feestelijk); maakt gebruik van de context bij een woord; gebruikt leestekens op de juiste wijze; leest groepen van woorden als een geheel; leest een tekst met het juiste dynamische en melodisch accent; leest een tekst in het juiste tempo en zonder spellinguitspraak; houdt bij het voorlezen rekening met het leesdoel en met het publiek.
Referentieniveau 1F	Kan teksten zodanig vloeiend lezen dat woordherkenning tekstbegrip niet in de weg staat.
Referentieniveau 2F	---
Referentieniveau 3F	---
Referentieniveau 4F	---

Tabel 2 Begrijpend lezen

Niveau	Omschrijving
Tussendoelen beginnende geletterdheid groep 1-3	Kan voorspellingen doen over het verdere verloop van een verhaal; kan een voorgelezen verhaal naspelen en navertellen; begrijpt eenvoudige verhalende en informatieve teksten.
Tussendoel gevorderde geletterdheid groep 4-5	Activeert kennis over thema in de tekst; koppelt verwijswoorden aan antecedenten; lost het probleem van een moeilijke zin op; voorspelt de voortgang in een tekst; leidt informatie af uit de tekst; onderscheidt verschillende soorten teksten; herkent de structuur van verhalende teksten.
Tussendoel gevorderde geletterdheid groep 6-8	Zoekt, selecteert en verwerkt op een doelbewuste en efficiënte manier informatie uit verschillende bronnen; leidt betekenisrelaties tussen zinnen en alinea's af; herkent inconsistenties; stelt zelf vragen tijdens het lezen; bepaalt de hoofdgedachte van een tekst; maakt een samenvatting; herkent de structuur van teksten; plant, stuurt, bewaakt en controleert leesgedrag; beoordeelt teksten op hun waarde.
Referentieniveau 1F	Zakelijke teksten ***Algemeen.*** Kan eenvoudige teksten lezen over alledaagse onderwerpen en over onderwerpen die aansluiten bij de leefwereld. ***Taakuitvoering.*** Herkent specifieke informatie, wanneer naar één expliciet genoemde informatie-eenheid gevraagd wordt (letterlijk begrip); kan (in het kader van het leesdoel) belangrijke informatie uit de tekst halen en kan de manier van lezen daarop afstemmen; kan informatie en meningen interpreteren die dicht bij de leerling staan; kan een oordeel over een (tekst)deel verwoorden; kan informatie opzoeken in duidelijk geordende naslagwerken; kan schematische informatie lezen en relaties met de tekst expliciteren. Fictieve teksten ***Algemeen.*** Kan jeugdliteratuur belevend lezen. ***Taakuitvoering.*** Herkent basale structuurelementen; kan meelezen met een personage; kan uitleggen hoe een personage zich voelt; kan gedichten en verhaalfragmenten parafraseren of samenvatten; kan relaties leggen tussen tekst en werkelijkheid; kan spannende/humoristische/dramatische passages in de tekst aanwijzen; herkent verschillende emoties in de tekst; evalueert de tekst met emotieve argumenten; kan met medeleerlingen leeservaringen uitwisselen; kan interesse in bepaalde fictievormen aangeven.

Tabel 2 (Vervolg)

Niveau	Omschrijving
Referentieniveau 2F	Zakelijke teksten ***Algemeen****.* Kan teksten lezen over alledaagse onderwerpen, onderwerpen die aansluiten bij de leefwereld van de leerling en over onderwerpen die verder van de leerling af staan. ***Taakuitvoering****.* Kan de hoofdgedachte van de tekst weergeven; maakt onderscheid tussen hoofd- en bijzaken; legt relaties tussen tekstdelen (inleiding, kern, slot) en teksten; ordent informatie voor een beter begrip; herkent beeldspraak; legt relaties tussen tekstuele informatie en meer algemene kennis; kan de bedoeling van tekstgedeeltes/ specifieke formuleringen duiden; kan de bedoeling van de schrijver verwoorden; kan relaties tussen en binnen teksten evalueren en beoordelen; kan eenvoudige teksten beknopt samenvatten; kan systematisch informatie zoeken. Fictieve teksten ***Algemeen.*** Kan eenvoudige adolescentenliteratuur herkennend lezen. ***Taakuitvoering****.* Herkent het genre; herkent letterlijk en figuurlijk taalgebruik; kan situaties en verwikkelingen in de tekst beschrijven; kan het denken, voelen en handelen van personages beschrijven; kan de ontwikkeling van de hoofdpersoon beschrijven; kan de geschiedenis chronologisch navertellen; kan bepalen in welke mate de personages en gebeurtenissen herkenbaar en realistisch zijn; kan personages typeren; kan het onderwerp van de tekst benoemen; evalueert de tekst met realistische argumenten en kan persoonlijke reacties toelichten met voorbeelden uit de tekst; kan met medeleerlingen leeservaringen uitwisselen; kan de interesse in bepaalde genres of onderwerpen motiveren.
Referentieniveau 3F	Zakelijke teksten ***Algemeen****.* Kan een grote variatie aan teksten over onderwerpen uit de (beroeps)opleiding en van maatschappelijke aard zelfstandig lezen; leest met begrip voor geheel en details. ***Taakuitvoering.*** Kan tekstsoorten benoemen; kan de hoofgedachte in eigen woorden weergeven; begrijpt en herkent relaties als oorzaak-gevolg, middel-doel, opsomming e.d.; maakt onderscheid tussen hoofd- en bijzaken, meningen en feiten, standpunt/drogreden en argument; trekt conclusies naar aanleiding van de tekst en over de opvattingen van de auteur; kan doel van de schrijver en talige middelen die gebruikt zijn om het doel te bereiken aangeven; kan de tekst opdelen in betekenisvolle eenheden en kan de functie van deze eenheden benoemen; kan de argumenten in een betogende tekst op aanvaardbaarheid beoordelen; kan informatie in een tekst beoordelen op waarde voor zichzelf en anderen; kan een tekst beknopt samenvatten voor anderen; kan de betrouwbaarheid van bronnen beoordelen; vermeldt bronnen; kan snel informatie vinden in langere rapporten/ingewikkelde schema's. Fictieve teksten ***Algemeen.*** Kan adolescentenliteratuur en eenvoudige volwassenliteratuur kritisch en reflecterend lezen. ***Taakuitvoering.*** Herkent vertel- en dichttechnische procedés en kan de werking toelichten; herkent veelvoorkomende stijlfiguren; kan causale verbanden leggen op het niveau van de handelingen van personages en de gebeurtenissen; kan expliciete doelen en motieven van personages opmerken; kan expliciete en impliciete doelen en motieven opmerken en benoemen; kan betekenis geven aan symbolen; kan aangeven welke vraagstukken centraal staan en de hoofdgedachte of boodschap van de tekst weergeven; evalueert de tekst ook met morele en cognitieve argumenten; kan uiteenzetten tot welke inzichten de tekst heeft geleid; kan met leeftijdgenoten discussiëren over de interpretatie en kwaliteit van teksten en over de kwesties die door de tekst worden aangesneden; kan interesses in bepaalde vraagstukken motiveren; kan de persoonlijke literaire smaak en ontwikkeling beschrijven.

Tabel 2 (Vervolg)

Niveau	Omschrijving
Referentieniveau 4F	Zakelijke teksten ***Algemeen.*** Kan een grote variatie aan teksten lezen over tal van onderwerpen uit de (beroeps)opleiding en van maatschappelijke aard en kan die in detail begrijpen. ***Taakuitvoering.*** Maakt onderscheid tussen uiteenzettende, beschouwende of betogende teksten; maakt onderscheid tussen objectieve en subjectieve argumenten; onderscheidt drogreden van argument; herkent argumentatieschema's; herkent ironisch taalgebruik; kan een vergelijking maken met andere teksten en tussen tekstdelen; kan impliciete relaties tussen tekstdelen aangeven; herkent en interpreteert persoonlijke waardeoordelen; kan argumentatie analyseren en beoordelen; kan een tekst beoordelen op consistentie; kan taalgebruik beoordelen; kan van een tekst een goed geformuleerde samenvatting maken die los van de uitgangstekst te begrijpen valt. Fictieve teksten ***Algemeen.*** Kan volwassenliteratuur interpreterend en esthetisch lezen. ***Taakuitvoering***. Herkent ironie; kan verschillende betekenislagen onderscheiden; kan stilistische, inhoudelijke en structurele bijzonderheden opmerken; kan zich empatisch identificeren met verschillende personages; kan het algemene thema formuleren; kan teksten in cultuur-historisch perspectief plaatsen; evalueert de tekst ook met structurele en esthetische argumenten; kan teksten naar inhoud en vorm vergelijken; kan interpretaties en waardeoordelen van leeftijdgenoten en literaire critici beoordelen; kan interesse in bepaalde schrijvers motiveren.

Tabel 3 Woordenschat

Niveau	Omschrijving
Tussendoelen mondelinge communicatie groep 1-3	Beschikt over een basiswoordenschat; breidt gericht de (basis)woordenschat uit; leidt nieuwe woordbetekenissen af uit verhalen; is erop gericht woorden productief te gebruiken; maakt onderscheid tussen betekenisaspecten van woorden.
Tussendoelen mondelinge communicatie groep 4-5	Verbreedt en verdiept de woordkennis; hanteert strategieën voor het afleiden van woordbetekenissen en onthouden van woorden; kent betekenisrelaties tussen woorden (onderschikking/bovenschikking); begrijpt figuurlijk taalgebruik.
Tussendoelen mondelinge communicatie groep 6-8	Kan woordenschat zelf verbreden en verdiepen; kan strategieën verwoorden voor het afleiden en onthouden van woordbetekenissen; kan woorden buiten de context definiëren; legt zelf betekenisrelaties tussen woorden; past figuurlijk taalgebruik toe.
Referentieniveau 1F	Domein Mondelinge taalvaardigheid Beschikt over voldoende woorden om te praten over vertrouwde situaties en onderwerpen, maar zoekt nog regelmatig naar woorden en varieert niet veel in woordgebruik. Domein Leesvaardigheid Kent de meest alledaagse woorden, of kan de betekenis van een enkel onbekend woord uit de context afleiden. Domein Schrijfvaardigheid Gebruikt voornamelijk frequent voorkomende woorden.
Referentieniveau 2F	Domein Mondelinge taalvaardigheid Beschikt over voldoende woorden om zich te kunnen uiten, het kan soms nog nodig zijn een omschrijving te geven van een onbekend woord. Domein Leesvaardigheid De woordenschat van de leerling is voldoende om teksten te lezen en wanneer nodig kan de betekenis van onbekende woorden uit de vorm, de samenstelling of de context afgeleid worden. Domein Schrijfvaardigheid Varieert het woordgebruik, fouten met idiomatische uitdrukkingen komen nog voor.
Referentieniveau 3F	Domein Mondelinge taalvaardigheid Beschikt over een goede woordenschat; kan variëren in de formulering; trefzekerheid in de woordkeuze is over het algemeen hoog, al komen enige verwarring en onjuist woordgebruik wel. Domein Leesvaardigheid --- Domein Schrijfvaardigeid Brengt variatie in woordgebruik aan om herhaling te voorkomen; woordkeuze is meestal adequaat, er wordt slechts een enkele fout gemaakt.
Referentieniveau 4F	Domein Mondelinge taalvaardigheid Beschikt over een breed repertoire aan woorden en idiomatische uitdrukkingen en uitdrukkingen uit de spreektaal. Domein Leesvaardigheid --- Domein Schrijfvaardigheid Er zijn geen merkbare beperkingen in het woordgebruik; het woordgebruik is rijk en zeer gevarieerd.

Tabel 4 Schrijfvaardigheid (Strategisch schrijven)*

Niveau	Omschrijving
Tussendoelen beginnende geletterdheid groep 1-3	Schrijft functionele teksten, zoals lijstjes, briefjes, opschriften en verhaaltjes.
Tussendoel gevorderde geletterdheid groep 4-5	Schrijft korte teksten (antwoorden op vragen, berichten, afspraken) en lange teksten; kent kenmerken van tekstsoorten; stelt onderwerp vast; is zich bewust van het schrijfdoel en lezerspubliek; verzamelt informatie uit enkele bronnen; ordent informatie; kiest geschikte woorden; formuleert gedachten en gevoelens in eenvoudige zinnen; schrijft korte teksten met de juiste spelling en interpunctie; leest eigen tekst na en reviseert die met hulp van anderen; maakt opmerkingen bij de eigen tekst.
Tussendoel gevorderde geletterdheid groep 6-8	Schrijft allerlei soorten teksten; herkent en gebruikt enkele kenmerken van tekstsoorten; stelt het schrijfdoel en het lezerspubliek van tevoren vast; verzamelt informatie uit verschillende soorten bronnen; ordent vooraf de gevonden informatie; kiest de juiste woorden en formuleert gedachten en gevoelens in enkelvoudige en samengestelde zinnen; schrijft langere teksten met de juiste spelling en interpunctie; besteedt aandacht aan de vormgeving en de lay-out; leest de geschreven tekst na en reviseert die zelfstandig; reflecteert op het schrijfproduct en –proces.
Referentieniveau 1F	***Algemeen.*** Kan korte, eenvoudige teksten schrijven over alledaagse onderwerpen of over onderwerpen uit de leefwereld. ***Taakuitvoering.*** De informatie is zodanig geordend, dat de lezer de gedachtegang gemakkelijk kan volgen en het schrijfdoel bereikt wordt; de meest bekende voegwoorden zijn correct gebruikt, met andere voegwoorden komen nog fouten voor; fouten met verwijswoorden komen voor; samenhang in de tekst en binnen samengestelde zinnen is niet altijd duidelijk; gebruikt basisconventies bij een formele brief; kan formeel en informeel taalgebruik hanteren; maakt redelijk accuraat gebruik van eenvoudige zinsconstructies; kan een titel gebruiken; voorziet een brief op de gebruikelijke plaats van datering, adressering, aanhef en ondertekening; besteedt aandacht aan de opmaakt van de tekst.
Referentieniveau 2F	***Algemeen****.* Kan samenhangende teksten schrijven met een eenvoudige lineaire opbouw, over uiteenlopende vertrouwde onderwerpen uit de (beroeps)opleiding en van maatschappelijke aard. ***Taakuitvoering.*** Gebruikt veelvoorkomende verbindingswoorden correct; de tekst bevat een volgorde in inleiding, kern, slot; kan alinea's maken en inhoudelijke verbanden expliciet aangeven; maakt soms nog onduidelijke verwijzingen en fouten in de structuur van de tekst; kan in teksten met een eenvoudige lineaire structuur trouw blijven aan het doel van het schrijfproduct; past het woordgebruik en toon aan het publiek aan; vertoont een redelijke grammaticale beheersing; gebruikt titel en tekstkopjes; heeft bij langere teksten (meer dan 2 A4) ondersteuning nodig bij aanbrengen van de lay-out.

Tabel 4 (Vervolg)

Niveau	*Omschrijving*
Referentieniveau 3F	***Algemeen.*** Kan gedetailleerde teksten schrijven over onderwerpen uit de (beroeps)opleiding en van maatschappelijke aard, waarin informatie en argumenten uit verschillende bronnen bijeengevoegd en beoordeeld worden. ***Taakuitvoering.*** De gedachtelijn is in grote lijnen logisch en consequent met hier en daar een niet hinderlijk zijspoor; relaties als oorzaak en gevolg, voor- en nadelen, overeenkomst en vergelijking, zijn duidelijk aangegeven; verbanden tussen zinnen en zinsdelen in samengestelde zinnen is over het algemeen goed aangegeven door het gebruik van juiste verwijs- en verbindingswoorden; alinea's zijn verbonden tot een coherent betoog; kan verschillende schrijfdoelen hanteren en in een tekst combineren; kan de opbouw van de tekst aan het doel van de tekst aanpassen; kan schrijven voor zowel publiek uit eigen omgeving als voor een algemeen lezerspubliek; past register consequent toe; toont een betrekkelijk grote beheersing van de grammatica; incidentele vergissingen, niet-stelselmatige fouten en kleine onvolkomenheden in de zinsstructuur kunnen nog voorkomen; geeft een heldere structuur aan de tekst; geeft in een langere tekst een indeling in paragrafen; stemt de lay-out af op doel en publiek.
Referentieniveau 4F	***Algemeen.*** Kan goed gestructureerde teksten schrijven over allerlei onderwerpen uit de (beroeps)opleiding en van maatschappelijke aard; kan relevante kwesties benadrukken, standpunten uitgebreid uitwerken en ondersteunen met redenen en voorbeelden ***Taakuitvoering.*** Geeft een complexe gedachtegang goed en helder weer; geeft duidelijk aan wat de hoofdzaken zijn en wat ondersteunend is in het betoog; geeft relevante argumenten voor het betoog inzichtelijk weer; verwijzingen in de tekst zijn correct; lange, meervoudig samengestelde zinnen zijn goed te begrijpen; kan schrijven voor zowel publiek uit eigen omgeving als voor een algemeen lezerspubliek; kan verschillende registers hanteren en heeft geen moeite om het register aan te passen aan de situatie en het publiek; kan schrijven in een persoonlijke stijl die past bij een beoogde lezer; handhaaft consequent een hoge mate van grammaticale correctheid, fouten zijn zeldzaam; lay-out en paragraafindeling zijn bewust en consequent toegepast om het begrip bij de lezer te ondersteunen.

* De beschrijving is beperkt tot strategisch schrijven. Spelling (technisch schrijven) is hier buiten beschouwing gelaten. In het referentiekader is spelling terug te vinden in de niveaubeschrijving van begrippenlijst en taalverzorging.