DIGITALE LEERLINGVOLGSYSTEMEN: een review van de effecten op leerprestaties

J.M. Faber & A.J. Visscher

Kennisnet

Universiteit Twente:
Vakgroep Onderzoeksmethodologie,
Meetmethoden en Data-analyse

20-02-2014

SAMENVATTING

Onderzoeksvragen

Steeds meer Nederlandse scholen gebruiken digitale leerlingvolgsystemen (Inspectie van het Onderwijs, 2013). Ook in andere landen, en werelddelen maken scholen in toenemende mate gebruik van dergelijke systemen, onder andere als een gevolg van de toename in het afnemen van toetsen (Heritage & Yeagley, 2005). Digitale leerlingvolgsystemen (DLVS-en) kunnen gedefinieerd worden als systemen waarmee leraren op basis van toetsresultaten feedback ontvangen over de resultaten van het aangeboden onderwijs.

De informatie die leraren ontvangen op basis van toetsen kan beschouwd worden als feedback aan leraren over de resultaten van zijn, of haar lesgeven (Visscher & Coe, 2002). Leraren kunnen op basis van deze feedback hun instructie aanpassen, zodat het onderwijs afgestemd wordt op de specifieke leerbehoeften van leerlingen. Wanneer dat het geval is kan het benutten van een DLVS resulteren in hogere leerresultaten. Zowel overheden als scholen investeren daarom veel in DLVS-en (Ministerie van Onderwijs, Cultuur en Wetenschap, 2007). Het is daarom van belang om te onderzoeken of het gebruiken van DLVS-en daadwerkelijk leidt tot hogere leerresultaten.

In dit onderzoeksrapport zijn de resultaten van een meta-analyse beschreven waarin dit verband onderzocht is. Voor deze meta-analyse zijn kwalitatief sterke experimentele onderzoeken naar de effecten van DLVS-en geselecteerd. De volgende vragen zijn beantwoord met de analyses:

- *Wat is het effect van digitaal leerlingvolgsysteemgebruik door leraren op de prestaties van leerlingen?*
- *Welke factoren belemmeren dan wel bevorderen het beoogde effect van digitaal leerlingvolgsysteem-gebruik op leerprestaties?*

Onderzoeksopzet

In de meta-analyse zijn 40 effecten opgenomen die afkomstig zijn uit vijftien verschillende onderzoeken. De onderzoeken zijn gevonden in één van zes databases. Aan de hand van vooraf opgestelde trefwoordenlijsten is systematisch gezocht naar relevante onderzoeken. Daarnaast zijn 126 internationale contactpersonen benaderd met de vraag of zij relevante onderzoeken van anderen kenden, dan wel zelf hebben uitgevoerd. In de databases zijn 38 onderzoeken gevonden, en 32 onderzoeken werden gevonden door het benaderen van de contactpersonen. De onderzoeken zijn gelezen en beoordeeld door twee onderzoekers die uiteindelijk bepaalden dat vijftien onderzoeken voldeden aan alle vooraf opgestelde inhoudelijke, en methodologische criteria.

De effecten in de geselecteerde onderzoeken zijn berekend met de *Cohen's d* en *Hedges's g* formules. Aan elk effect is een gewicht gekoppeld dat bepaalde hoeveel het betreffende effect meewoog in de gemiddelde effectgrootte. Deze gewichten zijn toegekend op basis van de variantie binnen de effecten. Er is gebruik gemaakt van een random effect model voor het bepalen van de gemiddelde effectgrootte. De analyses zijn uitgevoerd met het programma Comprehensive Meta-analyses.

Bevindingen

In de onderzoeken waarin de interventie gericht was op het verhogen van de leerresultaten van kleine groepen leerlingen (de interventie was bijvoorbeeld gericht op een specifiek aantal leerlingen binnen een klas) is een significante effectgrootte van 0.4 gevonden (zie Tabel 5). DLVS-gebruik resulteert in vergelijking met andere interventies die gericht zijn *op kleine groepen leerlingen* in een positief, en

relatief groot effect op leerresultaten. In de onderzoeken waar de interventie gericht was op het verhogen van de leerresultaten van alle leerlingen binnen *een gehele school*, of schoolbestuur is een significante effectgrootte van 0.06 gevonden (zie Tabel 5). DLVS-gebruik resulteert in vergelijking met andere interventies gericht op scholen in een iets lager dan gemiddeld effect op leerresultaten.

De volgende factoren blijken het beoogde effect van een DLVS op leerresultaten te bevorderen:

- een hoge feedbackfrequentie,
- systemen die naast feedback ook *advies geven over de instructie en verwerkingsmethoden* die passen bij de ontvangen feedback,
- een ondersteunende interventie die ten minste maandelijks plaatsvindt.

Het gemiddeld hoge effect van een DLVS in de onderzoeken met een interventie gericht op kleine groepen is niet gevonden in de onderzoeken met interventies gericht op scholen, of besturen. Het is daarom de moeite waard nader te onderzoeken hoe de succesvolle aanpakken van DLVS-gebruik voor kleine groepen leerlingen vertaald kunnen worden naar aanpakken op school-, of bestuursniveau die vergelijkbare hoge effecten sorteren.

INHOUDSOPGAVE

INLEIDING

Digitale leerlingvolgsystemen (DLVS-en) zijn systemen waarmee leraren op basis van toetsresultaten feedback ontvangen over de resultaten van het aangeboden onderwijs. Steeds meer scholen zetten zulke systemen in. In het Nederlandse basisonderwijs gebruiken de meeste scholen al zo'n volgsysteem, in het voortgezet onderwijs geldt dat voor ongeveer 25 tot 50 procent van de scholen (Inspectie van het Onderwijs, 2013). Veel gebruikte systemen in het Nederlandse basisonderwijs zijn het Cito LOVS, ParnasSys, ESIS en Dotcomschool (Faber, van Geel & Visscher, 2013). Niet alleen in Nederland gebruiken veel scholen digitale leerlingvolgsystemen; voorbeelden van Amerikaanse DLVS-en zijn het MAP (*Measures of Academic Progress*) en AM (*Accelerated Math*) (Rennie Center for Education Research & Policy, 2006). Ook in bijvoorbeeld België en Duitsland worden DLVS-en gebruikt, of zijn deze in ontwikkeling (Berkemeyer & Van Holt, 2012; Verhaeghe, Vanhoof, Valcke, & Van Petegem, 2011).

Een DLVS wordt gebruikt in combinatie met formatieve toetsen. Onder formatief toetsen vallen alle toetsactiviteiten die door leraren worden ondernomen om feedback te verzamelen waarmee zij hun instructie kunnen verbeteren (Black & Wiliam, 1998a). In verschillende meta-analyses is het effect van formatief toetsen op de leerresultaten onderzocht. Onderzoekers vinden voornamelijk positieve effecten (Black & Wiliam, 1998b; Fuchs & Fuchs, 1986), de grootte van de gevonden effecten verschilt echter wel (Kingston & Nash, 2011). Hellrung en Hartig (2013) zijn daarentegen kritisch over formatief toetsen. Voor hun onderzoek selecteerden zij studies waarin leraren externe rapportages ontvingen met daarin feedback op basis van toetsen. Uit hun review volgt dat leraren weinig gebruik maakten van dergelijke feedback. Leraren die de feedback wel gebruikten deden dat vooral op strategische wijze, ze besteden bijvoorbeeld eerder meer tijd aan het oefenen voor een toets, dan aan het verbeteren van de instructie.

Interimtoetsen of benchmarktoetsen kunnen ook formatief ingezet worden. Dit zijn toetsen die voornamelijk het voorspellen van de resultaten op verantwoordingstoetsen van overheden tot doel hebben. Interimtoetsen geven leraren daardoor ook feedback waarmee zij de instructie kunnen aanpassen, opdat de vereiste normen behaald worden. In de onderzoeken naar de effecten van interimtoetsen worden meestal geen positieve effecten gevonden (Henderson, Petrosino, Guckenburg, & Hamilton, 2007; Shaw & Wayman, 2012), of kleine, niet-significante, statistische positieve effecten (Quint, Sepanik, & Smith, 2008).

Er zijn verschillende redenen aan te voeren waarom DLVS-gebruik in combinatie met formatief toetsen in een positief effect op leerresultaten zou kunnen resulteren. Met een DLVS worden scholen bijvoorbeeld eerder zelf eigenaar van het analyseren en interpreteren van de resultaten uit toetsen, dan wanneer zij externe feedbackrapportages gebruiken waarin resultaten al voor hen geanalyseerd zijn. Formatief toetsen wordt dan wellicht eerder een vast onderdeel van het lesgeven. En juist dat laatste zien onderzoekers als een belangrijke succesfactor (Black & Wiliam, 1998b; Muralidharan & Sundararaman, 2010). Daarnaast kunnen leraren met een DLVS zelf bepalen wanneer analyses op toetsresultaten uitgevoerd worden. Leraren kunnen dus kort na een afname feedback vergaren en direct veranderingen realiseren in de onderwijspraktijk. Leraren kunnen dan zelf voorkomen dat de verstreken tijd tussen het moment van afname, en het moment van analyse te groot wordt, en de feedback dus niet meer goed aansluit bij de geldende leerbehoeften. Daarnaast kunnen leraren met een DLVS vaak op relatief simpele wijze resultaten uit toetsen analyseren, de ontwikkeling van hun

leerlingen over een langere periode grafisch in beeld brengen, en of de resultaten van leerlingen vergelijken met landelijke gemiddelden waardoor men een referentiepunt voor de eigen resultaten heeft. Dit laatste wordt benchmarking genoemd. Digitale systemen die leraren deze mogelijkheden bieden kunnen leraren zo faciliteren bij het effectiever inzetten van de feedback uit toetsen (Wayman, Stringfield, & Yakimowski, 2004).

Er is behoefte aan DLVS-en omdat scholen steeds meer, en vaker toetsen afnemen (Heritage & Yeagley, 2005; Visscher & Coe, 2003). Toetsen zijn belangrijker geworden in het onderwijs omdat ze niet meer alleen ingezet worden om leerlingen te beoordelen, maar tevens dienen voor het geven van feedback over de kwaliteit van het onderwijs, feedback die gebruikt kan worden voor de verbetering van de onderwijskwaliteit. Daarnaast moeten scholen aan de hand van toetsresultaten ook verantwoording afleggen over de onderwijskwaliteit die zij leveren. Toetsresultaten zeggen iets over de mate waarin de school erin slaagt om in het onderwijs aan te sluiten bij de leerbehoeften van leerlingen (Heritage & Yeagley, 2005).

Door de Nederlandse overheid wordt zowel het beter benutten van digitale middelen in het onderwijs, als het inzetten van feedback uit formatieve toetsen ter verbetering van de onderwijskwaliteit gestimuleerd (Ministerie van Onderwijs, Cultuur en Wetenschap, 2007). Een vergelijkbare ontwikkeling is zichtbaar in de VS. Ook daar worden scholen sinds het *No Child Left Behind* beleid gestimuleerd om meer toetsen af te nemen, en de toetsresultaten in te zetten ter verbetering van het onderwijs (Heritage & Yeagley, 2005). In navolging hiervan brengen veel organisaties DLVS-en en bijbehorende formatieve toetsen op de markt (Shepard, 2010).

Omdat overheden en scholen veel investeren in DLVS-en is het van belang om te onderzoeken of de verwachte effecten daadwekelijk gevonden worden. In de vorm van een meta-analyse is in dit onderzoek daarom onderzocht of DLVS-gebruik een positief effect heeft op leerresultaten. Om deze vraag te beantwoorden zijn de resultaten van experimentele studies en quasi-experimentele studies die aan hoge methodologische eisen voldoen opgespoord en geanalyseerd. Daarnaast is onderzocht onder welke omstandigheden het effect van DLVS-gebruik het sterkst is. In dit rapport worden daarom de volgende onderzoeksvragen beantwoord:

- *Wat is het effect van digitaal leerlingvolgsysteemgebruik door leraren op de prestaties van leerlingen?*
- *Welke factoren belemmeren dan wel bevorderen het beoogde effect van digitaal leerlingvolgsysteemgebruik op leerprestaties?*

Opbouw van het rapport

In het eerste hoofdstuk van dit rapport wordt een definitie van digitale leerlingvolgsystemen gepresenteerd, en worden de centrale begrippen binnen deze definitie toegelicht. In het tweede hoofdstuk worden eerdere onderzoeksbevindingen over het werken met DLVS-en in het onderwijs beschreven. Dit hoofdstuk wordt afgesloten met het presenteren van de implementatiebevorderende factoren waarvan verwacht wordt dat ze het effect van een DLVS op leerresultaten zullen beïnvloeden. Vervolgens wordt de onderzoeksmethode omschreven in hoofdstuk drie. De wijze waarop de studies voor dit onderzoek zijn geselecteerd, en hoe de resultaten uit deze studies vervolgens zijn geanalyseerd wordt daar uiteengezet. In het daarop volgende vijfde hoofdstuk worden de onderzoeksresultaten gepresenteerd. In het laatste hoofdstuk worden tot slot de onderzoeksvragen beantwoord en de gevonden onderzoeksresultaten bediscussieerd.

1. EEN DEFINITIE VAN DIGITALE LEERLINGVOLGSYSTEMEN

In dit onderzoek is er naar gestreefd om alle methodologisch sterke studies naar de effecten van DLVS-en mee te nemen in de analyses. Er is daarom gekozen voor een brede definitie van DLVS-en, namelijk:

Digitale leerlingvolgsystemen zijn digitale systemen waarmee leraren op basis van toetsen feedback ontvangen over de resultaten van het door hen aangeboden onderwijs.

De DLVS-en die onder deze definitie vallen kunnen onderling sterk verschillen. Het CITO-LOVS onderscheidt zich bijvoorbeeld van andere Nederlandse systemen, omdat er analyses mee uitgevoerd kunnen worden op het categorieniveau van een toets (Faber, Van Geel, & Visscher, 2013). Dat wil zeggen dat door analyses in beeld gebracht wordt met welke specifieke toetscategorieën, (bijvoorbeeld meten, of delen en vermenigvuldigen) een leerlingen moeite hebben. Een ander systeem, zoals mCLASS kan ingezet worden om toetsen digitaal bij leerlingen af te nemen. Nadat alle toetsen afgenomen zijn geeft mCLASS leraren advies over de benodigde instructie (Ginsburg, Cannon, Eisenband, & Pappas, 2006). Een derde voorbeeld waarmee de variatie tussen DLVS-en aangegeven kan worden is het DLVS Acuity. De belangrijkste functie van dit DLVS is het voorspellen van de resultaten op *state tests*, zodat schoolleiders en leraren, indien nodig, kunnen bijsturen (*state tests* zijn de toetsen waarmee veel Amerikaanse staten bepalen of scholen hun onderwijsdoelstellingen behalen).

Uit de definitie volgt dat voor een opname in dit onderzoek er ten eerste sprake moet zijn van een *digitaal systeem*. Ten tweede moet er sprake zijn van een systeem dat mede ontworpen is voor gebruik door *leraren.* Immers, wanneer de feedback niet door leraren gebruikt wordt om het onderwijs af te stemmen op de leerbehoeften, dan zullen er logischerwijs ook geen effecten gevonden worden (McCaffrey & Hamilton, 2007).

Het derde centrale begrip in de DLVS-definitie is *feedback*. De achterliggende theorie van formatief toetsen is gebaseerd op het feedbackmechanisme. De informatie die leraren ontvangen op basis van toetsen kan beschouwd worden als feedback aan leraren over de resultaten van zijn of haar lesgeven (Visscher & Coe, 2002). Feedback kan gedefinieerd worden als van een (externe) bron afkomstige informatie over de kloof tussen een bestaand niveau en het gewenste niveau (Ramaprasad, 1983). Het positieve effect van feedback is in meerdere onderzoeken aangetoond (Hattie & Timperley, 2007; Kingston & Nash, 2011; Kluger & DeNisi, 1996). Hattie and Timperley (2007) concluderen dat feedback effectief is wanneer het tenminste informatie over de volgende drie punten geeft:

- Wat is het beginniveau van de leerling? (*feed-up*)
- Wat is de ontwikkeling van leerling gezien het beginniveau? *(feedback)*
- Hoe gaat de leerling verder en wat voor aanpakken zijn daarbij nodig?(*feed-forward*)

Feedback is effectiever wanneer ook prestatiedoelen gesteld worden (Kluger & DeNisi, 1996; Goedele Verhaeghe et al., 2010). Ontvangers van feedback die specifieke en meetbare doelen opstellen kunnen gericht feedback verzamelen en beter het resultaat van hun investeringen bepalen. Succes is zo duidelijker (nl. wanneer de gestelde doelen behaald worden) en de acties zijn gerichter dan bij abstracte doelen. Bovendien kunnen uitdagende doelen een motiverend effect hebben (Locke & Latham, 2002).

Toetsen betreft het vierde centrale begrip in de definitie. Het type toets is bepalend voor de inhoud van de feedback. Perie, Marion and Gong (2009) onderscheiden toetsen op basis van twee kenmerken: de *frequentie van afname,* en de *reikwijdte van de toets*. Toetsen die frequent afgenomen worden bieden

de mogelijkheid om vlak na het behandelen van een bepaald deel uit de lesmethode gegevens te verzamelen over het effect daarvan. De frequentie van een toets zal een rol spelen bij het effect van een DLVS, omdat leraren feedback uit toetsresultaten beter kunnen koppelen aan het lesgeven wanneer er weinig tijd verstreken is tussen de toetsafname en de feedback (Hellrung & Hartig, 2013).

Ook de *reikwijdte* van de toets kan van invloed zijn op het effect van een DLVS op leerresulaten, omdat het mede de inhoud van een toets zal bepalen. Analyses op toetsen met een specifieke inhoud zullen eerder de mogelijkheid bieden om het presteren op specifieke leercategorieën in beeld te brengen, dan bredere toetsen (bijvoorbeeld toetsen waarin meerdere vakgebieden aan de orde komen). Brede toetsen worden vaak door een groot aantal scholen afgenomen waardoor analyses op deze toetsen de mogelijkheid bieden om het functioneren van leerlingen te vergelijken met het functioneren van gemiddelde leerlingen in dezelfde leeftijdscategorie.

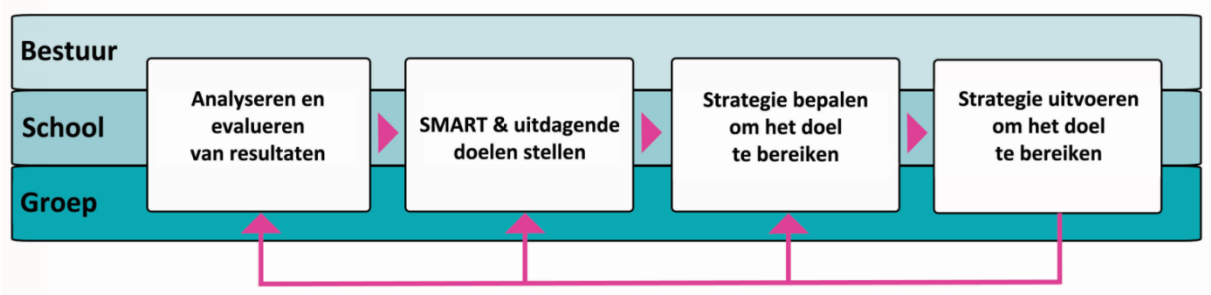
Een derde belangrijk kenmerk van toetsen is of de toets een onderliggende *schaal* bevat. Wanneer dit het geval is kunnen leraren de resultaten van afnames op verschillende momenten met elkaar vergelijken en zo de groei, de ontwikkeling van leerlingen in beeld brengen. Daarnaast kan de ontwikkeling van leerlingen ook beter vergeleken worden met andere groepen leerlingen (bijvoorbeeld met leerlingen uit het voorgaande leerjaar), omdat bekend is wat de *gemiddelde* ontwikkeling is.

Naast de *toetsfrequentie*, *de reikwijdte* en het wel, of niet kunnen werken met een onderliggende *schaal* is het *vakgebied* dat getoetst wordt ook een factor die van invloed is, omdat deze bepalend is voor de inhoud van de feedback. Binnen het vakgebied rekenen zijn bijvoorbeeld veel specifieke leerstofinhoudelijke categorieën te onderscheiden, dit in tegenstelling tot het vakgebied begrijpend lezen waarvoor dit veel lastiger is.

Om bruikbare en relevante feedback aan leraren te leveren moeten toetsen tot slot voldoen aan een tweetal basisvoorwaarden. Toetsen moeten valide zijn, de gegevens uit toetsen moeten met andere woorden echt iets relevants zeggen over de beheersing van de kennis die getoetst wordt. De rekentoets moet bijvoorbeeld informatief zijn over hoe goed een kind kan rekenen zonder dat de leesvaardigheid van het kind de eindscore sterk beïnvloedt. Daarnaast moeten toetsen betrouwbaar zijn, zodat gegevens stabiel zijn en dus niet sterk beïnvloed worden door externe factoren, zoals bijvoorbeeld het toetsmoment.

Feedback mechanismen

Feedback leidt volgens een aantal veronderstelde fasen, of mechanismen naar hogere leerresultaten (zie Figuur 1). Grotendeels komen de fasen uit Figuur 1 overeen met de bekendere Plan-Do-Check-Act cyclus. Een belangrijk verschil is echter dat niet gestart wordt met het stellen van doelen, maar met het analyseren van toetsgegevens, zodat op basis daarvan realistische doelen kunnen worden gesteld. Nadat doelen zijn geformuleerd is het noodzakelijk om de leerbehoeften van leerlingen te onderzoeken, zodat het onderwijs, of de strategie aansluit bij de huidige leerbehoeften van leerlingen. Nadat een strategie bepaald is aan de hand van de geanalyseerde feedback kan deze uitgevoerd worden. Daarna wordt feedback verzameld over het effect van de gekozen strategie. De leerresultaten worden weer geanalyseerd, en geëvalueerd om te bepalen of de gekozen strategie inderdaad geresulteerd heeft in het behalen van de leerdoelen. De cyclus wordt dan weer doorlopen vanaf de eerste fase. Zo ontstaat er een werkwijze waarbinnen feedback in het onderwijs systematisch en cyclisch gebruikt wordt.



Figuur 1 De veronderstelde fasen van de benutting van data uit digitale leerlingvolgsysteemsystemen (bron: Keuning & van Geel, 2012)

2. IMPLEMENTATIEBEVORDERENDE FACTOREN

In deze studie zijn naast (quasi-)experimentele onderzoeken veel niet-experimentele, maar wel relevante studies gevonden. De resultaten van laatstgenoemde studies gaven bijvoorbeeld inzicht in het feedbackmechanisme en de wijze waarop leraren DLVS-en gebruiken. De bevindingen van deze onderzoeken worden in dit hoofdstuk kort besproken evenals de resultaten van (quasi-) experimentele onderzoeken die niet voldeden aan de gestelde methodologische criteria.

Deze studies zijn vervolgens ook gebruikt voor het bepalen van de factoren die van invloed werden geacht op het effect van DLVS-gebruik. Deze implementatiebevorderende factoren worden aan het eind van dit hoofdstuk gepresenteerd.

Wayman, Cho, and Shaw (2009) onderzochten de effecten van Acuity. Acuity is een systeem dat bestaat uit vier diagnostische toetsen en drie voorspellende toetsen, leraren kunnen deze toetsen gebruiken om te voorspellen of leerlingen voldoende op de *state toetsen* zullen scoren. Het systeem bestaat daarnaast uit tools waarmee leraren behaalde leerresultaten kunnen analyseren. Na één schooljaar bleek dat het systeem weinig gebruikt werd, 44% van de leraren logde niet één keer in, en de overige 56% gebruikte het systeem minder dan één keer per maand. Een jaar later werd het systeem vaker gebruikt, gemiddeld elf keer per leerjaar per schooljaar, maar er werden geen effecten op leerresultaten gevonden (Wayman, Shaw, & Cho, 2011). Leraren gaven daarbij aan dat zij Acuity weinig gebruikten omdat het zorgde voor extra werkdruk, het systeem geen nieuwe informatie bood, en omdat zij zelf het gebruik van het systeem niet goed beheersten (Wayman et al., 2009). Zij konden de benodigde analysetools bijvoorbeeld niet terugvinden in het menu, of kozen verkeerde tools en konden de resultaten vervolgens niet interpreteren.

Veel leraren hebben moeite met het analyseren van toetsgegevens, het interpreteren van analyses en het vertalen van de bevindingen naar het lesgeven (Hellrung & Hartig, 2013; Inspectie van het Onderwijs, 2013; Williams & Coles, 2007). Ze hebben bijvoorbeeld moeite met het interpreteren van schalen en gemiddelden en kunnen hierdoor niet goed afleiden of leerlingen moeite hebben met bepaalde leerstofonderdelen (Nabrs Olah, Lawrence, & Riggan, 2010). Het is daarom belangrijk dat de software van een DLVS gebruikersvriendelijk is. De resultaten van analyses moeten bijvoorbeeld grafisch goed weergegeven worden en snel te koppelen zijn aan de bijbehorende leerlingen. Een gebruikersvriendelijk systeem zorgt er bovendien voor dat leraren niet onnodig tijd verliezen (Verhaeghe, Vanhoof, Valcke, & Van Petegem, 2010), en zo ook minder snel een negatieve attitude ten aanzien van het systeem en de bijbehorende werkwijze ontwikkelen (Wayman, 2007; Wayman, Cho, Jimerson, & Spikes, 2012).

De feedback uit formatieve toetsen wordt beter benut wanneer aan leraren ondersteuning wordt geboden (McCaffrey & Hamilton, 2007). Deze ondersteuning kan geboden worden vanuit een interventie zoals een training (Fuchs, Hamlett and Stecker, 1991), vanuit de schoolleiding (Wayman et al., 2012), en/of vanuit een schoolbestuur (Shaw & Wayman, 2012). Deze ondersteuning kan effecten op leerresultaten bevorderen. Het is van belang dat schoolleiders expertise bezitten over de inhoud van de feedback, zodat zij leraren kunnen stimuleren om kritisch en analytisch naar de feedback te kijken (Blanc et al., 2010; Wayman et al., 2012). Interventies zouden bovendien effectiever zijn wanneer ze gericht zijn op het gehele onderwijsteam (Wayman et al., 2011), binnen de school zelf plaatsvinden (en dus niet in een externe organisatie), en gericht zijn op de specifieke context waarin een school zich

bevindt (Blanc et al., 2010; Kelly, Downey, & Rietdijk, 2010; Verhaeghe et al., 2010). Het is ook van belang dat leraren de resultaten van analyses leren verbinden aan concrete aanpassingen van hun lessen. Leraren onderling kunnen elkaar onderling ondersteunen, samenwerking tussen leraren waarbij analyses en leerresultaten besproken worden kan het beoogde effect van formatieve toetsen bevorderen (Wayman et al., 2012).

Leraren prefereren korte informele toetsvormen die zij in hun lessen kunnen integreren, daaronder vallen niet alleen zelfgemaakte toetsen, maar ook observaties tijdens lessen en informatie die zij halen uit interacties met hun leerlingen. De meer formele toetsen die gebruikt worden in het geval van Acuity boden volgens leraren geen aanvulling op de informatie uit deze informele toetsvormen. Beide toetsvormen sloten elkaar uit volgens de leraren (Wayman et al., 2009). Leraren hebben daarnaast een voorkeur voor feedback over het beheersingsniveau en de ontwikkeling van een individuele leerling (Verhaeghe et al., 2010). De resultaten van de analyses op het individuele leerlingniveau worden grotendeels bepaald door de individuele kenmerken van een leerling, terwijl analyses op de resultaten van meerdere (groepen) leerlingen meer feedback geven over de resultaten van en de kwaliteit van het lesgeven. Als bijvoorbeeld een meerderheid van de leerlingen dezelfde toetsopgaven onvoldoende maakt, dan is de daarbij horende instructie waarschijnlijk niet op de juiste wijze aangeboden.

Acuity en de bijbehorende toetsen sloten onvoldoende aan bij de dagelijkse (toets) praktijk, waardoor het veel tijd zal kosten om dergelijke systemen goed te integreren. Shaw en Wayman (2012) betogen dat effecten van DLVS-en pas laat meetbaar zijn, omdat het langer dan twee jaar duurt voordat scholen erin geslaagd zijn om een dergelijke werkwijze goed te implementeren. Een andere verklaring voor uitblijvende effecten volgt uit de bevinding dat leraren juist minder differentiëren in de instructie wanneer zij gebruik maken van feedback uit toetsen (Chojnacki et al., 2013; Williams, et al., in press). Wanneer verwachtingen van leraren over de capaciteiten van hun leerlingen bevestigd worden zouden zij juist ook kunnen besluiten om de instructie niet aan te passen. De feedback bevestigt dan dat een leerling onvoldoende capaciteiten heeft om hogere resultaten te realiseren, dus waarom zouden ze hun instructie dan aanpassen? Deze bevinding kan ook veroorzaakt worden doordat leraren uit de toetsresultaten de *gemiddelde* leerbehoeften in de groep genereren en hun instructie op de gemiddelde leerbehoefte in plaats van op individuele behoeften richten.

In tegenstelling tot het onderzoek naar Acuity worden wel positieve effecten gevonden van het gebruiken van Accelerated Math (AM) (Burns, Klingbeil, & Ysseldyke, 2010; Spicuzza et al., 2001; Ysseldyke et al., 2003). AM is een flexibel digitaal systeem dat geïntegreerd kan worden met verschillende curricula, lesmaterialen en toetsen. In het onderzoek van Spicuzza kregen leraren een training over de mogelijkheden van AM. Elke leraar besloot vervolgens zelf hoe hij of zij het systeem ging integreren in het lesgeven. Spicuzza et al. (2001) onderzochten door welke mechanismen de positieve effecten verklaard kunnen worden. Een toename in het monitoren van de ontwikkeling van leerlingen, en het vaker toepassen van afgestemde leerstrategieën bieden volgens de auteurs een verklaring. AM kan deze mechanismen tot stand brengen, doordat het systeem een format bevat waarmee de ontwikkeling van leerlingen richting instructiedoelen in beeld gebracht wordt, en omdat het systeem leraren informatie verschaft over specifieke leerbehoeften. Overigens was er in het onderzoek van Spicuzza geen sprake van een aselecte toewijzing aan de control of treatment groep. Leraren hadden zich vrijwillig aangemeld en leerresultaten werden vergeleken met resultaten van leerlingen uit gematchte scholen.

Ook worden positieve effecten gevonden van het gebruik van het ASSISTment systeem (Koedinger, McLaughlin, & Heffernan, 2010). Van de leraren die dit systeem intensief gebruikten scoorden de laag

scorende leerlingen hoger, dan de laag scorende leerlingen van de andere leraren. Volgens de auteurs is ASSISTment een effectief systeem, omdat het de werkdruk van leraren nauwelijks verhoogt, de feedback specifiek aangeeft waardoor bepaalde leerlingen opgaven niet goed beantwoorden, en doordat zowel leerlingen als leraren direct na een toetsafname de feedback ontvangen.

Verbetering van de leerresultaten kan echter ook het gevolg zijn van andere mechanismen (Shepard, 2010). Hellrung & Hartig (2013) constateren dat leraren feedback uit externe rapportages vooral strategisch inzetten, leraren richten zich bijvoorbeeld meer op de leerlingen die nog niet voldoen aan normen, oefenen vaker met leerlingen voor toetsen, of sluiten te laag presterende leerlingen uit voor de toetsen (Rossi, Lipsey, & Freeman, 2004). Strategisch gebruik van feedback lijkt meer voor te komen in de UK en in de VS, dan in landen waarbinnen minder sterke consequenties verbonden zijn aan het niet behalen van gestelde normen (Hellrung & Hartig, 2013). De accountability context, de mate waarin er consequenties verbonden zijn aan te lage leerresultaten kan met andere woorden van invloed zijn op de wijze waarop scholen feedback uit toetsen gebruiken (Visscher & Coe, 2002).

Op basis van het voorgaande kunnen implementatiebevorderende factoren onderscheiden worden die van invloed zouden kunnen zijn op het effect van een DLVS op de leerresultaten. Voor dit onderzoek is een selectie gemaakt van implementatiebevorderende factoren waarvan verwacht wordt dat ze het meest bepalend zullen zijn voor genoemd effect. We onderscheiden de volgende zes factoren, en presenteren de bijbehorende hypothesen.

1. Frequentie van de feedback

Wanneer leraren voortdurend op de hoogte zijn van de leerbehoeften van leerlingen, en de veranderingen daarin, dan zullen leraren hun instructie beter afstemmen op de leerbehoeften. Leraren hebben daarom frequente feedback nodig om op de hoogte te blijven van leerbehoeften. *Er worden hogere effecten verwacht in de studies waarin leraren frequent feedback ontvangen, dan in de studies waarin de leraren minder frequent feedback ontvangen.*

2. Inhoud van de feedback

Op basis van de inhoud van de feedback moeten leraren hun instructie kunnen verbeteren. Voor het aanpassen van de instructie hebben leraren concrete, en bruikbare feedback nodig. Feedback waarin aangegeven wordt met welke leerstofinhoudelijke categorieën een leerling moeite heeft zal meer richting aan het handelen van een leraar geven, dan feedback waarin alleen een overall score voor een vakgebied gegeven wordt. De leraar kan uit het laatstgenoemde immers niet afleiden met welk onderdeel binnen het vakgebied de leerling moeite heeft. Er zijn DLVS-en waarin advies gegeven wordt over de benodigde instructie of leerdoelen, deze vorm van feedback geeft zeer concreet richting aan het handelen van de leraar. *In dit onderzoek wordt verwacht dat het effect van een DLVS op leerresultaten groter is wanneer de feedback meer richting geeft aan het handelen van een leraar.*

3. Interventie-intensiteit

Leraren die voldoende kennis en vaardigheden bezitten voor het gebruiken van een DLVS, en het toepassen van feedback zullen eerder hogere leerresultaten weten te realiseren, dan leraren die die kennis en vaardigheden onvoldoende bezitten. De intensiteit van een interventie zal mede bepalen in welke mate leraren de benodigde kennis en vaardigheden beheersen. De intensiteit van een interventie zal bepaald worden door:

- De omvang en duur van de interventie
  Hiermee wordt de tijdsbesteding van leraren aan het leren beheersen van de vereiste kennis en vaardigheden bedoeld evenals de duur van de interventie. *Verwacht wordt dat in de studies met een grote omvang en duur van de interventie hogere effecten worden gevonden dan in studies met een kleine en korte interventie.*
- De inhoud van de interventie
  Het effect van een interventie zal bovendien groter zijn wanneer leraren de geleerde werkwijzen uit de interventie langdurig implementeren in hun onderwijs. Dit zal niet bereikt worden met een interventie waarin alleen informatie gegeven wordt over de technische mogelijkheden van een DLVS. In de interventie zou er daarom ook aandacht moeten zijn voor de vertaling van de feedback naar aanpassingen in de instructie. Voorbeelden daarvan zijn dat leraren krijgen uitgelegd waarvoor welke functies dienen binnen het DLVS, en hoe zij bijvoorbeeld een ontwikkelingsgrafiek kunnen opvragen met behulp van het DLVS. *De interventies waarin de leraren leren hoe zij de feedback kunnen interpreteren en toepassen in het onderwijs, en daarbij begeleid worden zullen resulteren in hogere effecten, dan de studies waarin deze vormen van ondersteuning niet plaatsvinden.*

4. Doelgroep van de interventie

Niet alleen support vanuit een interventie is bevorderend voor het effect op leerresultaten, ook de support vanuit het schoolteam, de schoolleiding en/of het bestuur is daarbij van belang. *Er worden daarom hogere effecten verwacht van de interventies waarin het schoolteam, de schoolleiding en het bestuur betrokken is bij de interventie, dan van de interventies die alleen gericht zijn op individuele leraren.*

5. Feedback over het niveau van groep(en)

Als leraren de individuele leerresultaten kunnen omzetten naar het groepsniveau, dan kunnen zij beter conclusies trekken over het effect van de aangeboden klassikale lesmethode en instructie. Leraren kunnen hun klassikale instructie of lesmethode aanpassen wanneer blijkt dat de ontwikkeling van meerdere leerlingen in de groep niet naar verwachting verloopt. Bovendien kan de schoolleiding met de leerresultaten over het groepsniveau ook de kwaliteiten van leraren in beeld brengen. *In de studies waarin naast de individuele leerresultaten ook gewerkt wordt met leerresultaten op het groepsniveau worden hogere effecten verwacht, dan in studies waarin alleen gewerkt wordt met individuele leerresultaten.*

6. Benchmarking

Leraren kunnen beter een oordeel vormen over de ontwikkeling van hun leerlingen wanneer zij deze ontwikkeling kunnen vergelijken aan de ontwikkeling van een referentiegroep, of aan opgestelde normen. Leraren hebben hierdoor een referentiekader waardoor zij beter kunnen vaststellen of de ontwikkeling van een leerling niet, of juist boven verwachting verloopt. *In de studies waarin leraren de ontwikkeling van leerlingen kunnen beoordelen aan de hand van opgestelde normen/standaarden worden hogere effecten verwacht, dan in studies waar leraren deze mogelijkheid niet hebben.*

3. ONDERZOEKSMETHODE

In dit hoofdstuk wordt omschreven hoe gezocht is naar relevante studies, aan welke criteria de meegenomen studies moesten voldoen, en hoe deze vervolgens geanalyseerd zijn.

3.1. Zoekstrategie

Databases

In zes databases is systematisch gezocht naar (quasi-) experimentele studies. De databases *Educational Resources Information Center (ERIC), Web of Science, Scopus en PsycINFO* zijn gebruikt; daarnaast is gezocht naar proefschriften in de databases: *Narcis* (Nederlands) en *International Dissertation Abstracts*.

Om zoveel mogelijk relevante studies te vinden is ervoor gekozen om breed te zoeken, en daarom verschillende trefwoorden te gebruiken. Er is in drie stappen gezocht. In de ***eerste stap*** werd gezocht met trefwoorden die betrekking hebben op relevante typen toetsen:

> *accountability test, benchmark assessment, curriculum based assessment, diagnostic assessment, formative assessment, interim assessment, standardized achievement test, diagnostic test, high stakes test, low stakes test, summative assessment.*

In de ***tweede stap*** wordt gewerkt met trefwoorden die betrekking hebben op formatief toetsen:

> *assessment for learning, curriculum based measurement, data-driven, data-based, data analysis, feedback (response), formative evaluation, monitoring student progress, performance driven education, progress monitoring.*

De tweede stap werd tevens gecombineerd met trefwoorden die betrekking hebben op de leerprestaties van leerlingen, hiermee werd de zoekopdracht gespecificeerd en het aantal gevonden resultaten ingeperkt. Hiervoor werden de volgende trefwoorden gebruikt:

> *academic achievement, data feedback, learning outcome, student learning, student achievement, reading achievement, mathematics achievement, science achievement, writing achievement, outcomes of education.*

In de ***derde stap*** werd gezocht met trefwoorden over DLVS-en:

> *data analysis tool, data reporting system, electronic data management system, pupil monitoring system, school performance feedback system, student management system, student monitoring system, student progress system.*

Elke stap werd steeds gecombineerd met trefwoorden over een (quasi) experimentele onderzoeksmethode: *matching, regression discontinuity design, random, experiment, control group*. Daarnaast werden de woorden gecombineerd met trefwoorden over het type onderwijs: *elementary education, elementary secondary education, primary education, elementary school teachers, elementary school, grade 1/2/3/4/5/6/7/8/9/10/11/12, secondary education, secondary school teachers, secondary school*. Op die manier werd alleen naar studies binnen het onderwijs voor leerlingen van 4 tot 18 jaar gezocht. Studies vóór 1990 werden niet meegenomen, omdat er een kleine kans bestaat dat vóór deze periode de onderzochte digitale middelen ingezet werden.

Contactpersonen

Naast het zoeken in databases zijn wereldwijd via e-mail 126 contactpersonen en instanties in 24 verschillende landen benaderd die werkzaam zijn op het terrein van DLVS-en. Hen werd gevraagd naar informatie over studies die men kent en naar andere relevante contactpersonen op dit gebied (zie Bijlage 3).

Criteria

Van de gevonden studies werden de titel, het abstract; en in sommige gevallen ook een beschrijving van de interventie gelezen. Zo werd beoordeeld of de studie voldeed aan het inhoudelijke criterium voor opname, namelijk:

1. Het digitale leerlingvolgsysteem (DLVS) in het onderzoek komt overeen met de definitie voor DLVS-en, zoals deze in dit rapport opgesteld is.

Veel studies bleken voor de geplande meta-analyse niet relevant te zijn. Het ging daarbij vooral om studies naar het effect van feedback op scholieren, studies waarin geen gebruik gemaakt werd van digitale systemen, studies die zich vooral richten op de psychometrie van toetsen, dan wel studies waarin het effect op leerresultaten niet onderzocht werd. Het zoeken in de databases leverde uiteindelijk 38 inhoudelijk relevante onderzoeken op (zie Tabel 1). Het benaderen van contactpersonen heeft 32 studies opgeleverd, 12 daarvan waren ook al gevonden in de databases. In totaal werden dus 58 studies gevonden die inhoudelijk aansloten bij het onderzoek. Deze 58 studies zijn vervolgens nauwkeuriger gelezen door de twee onderzoekers, om te beoordelen of de studies voldeden aan de volgende vooraf opgestelde methodologische criteria:

2. De duur van het onderzoek is ten minste 12 weken (Slavin, 2008a). Als het onderzoek te kort duurt bestaat namelijk de kans dat het effect van de interventie niet (volledig) meetbaar is en hele korte studies kunnen kortstondig sterke effecten sorteren die op langere termijn niet haalbaar zijn.
3. De afhankelijke variabele bestaat uit de kwantitatieve resultaten van leerlingen op toetsen voor het vakgebied wiskunde-rekenen, lezen of taal. Het DLVS-gebruik en de interventie moet direct dan wel indirect gericht zijn op het verhogen van de resultaten op deze variabele. Onderzoeken waarin de effecten bepaald zijn op basis van de resultaten op normatieve landelijke eindtoetsen terwijl in de interventie de gegevens uit bijvoorbeeld interim- of benchmarktoetsen gebruikt zijn, zullen wel worden opgenomen.
4. In het onderzoek worden de resultaten van een experimentele groep vergeleken met de resultaten uit een controlegroep. In deze controlegroep vindt geen interventie plaats. De toewijzing aan de experimentele en controlegroep is random, dan wel bepaald door middel van propensity score matching, of het onderzoek wordt gekenmerkt door een regressie-discontinuïteit design.
5. In het onderzoek zijn voor- en nametingen gedaan op de afhankelijke variabele. Het verschil tussen de experimentele en controlegroep mag tijdens de voormeting niet meer dan 50% van een standaarddeviatie zijn, indien dit wel het geval is wordt het betreffende onderzoek niet meegenomen. De experimentele groep is dan niet vergelijkbaar met de controlegroep wat betref de spreiding in de leerprestaties en voor de verschillen is niet goed te corrigeren (Shadish, Cook, & Campbell, 2002).
6. In het onderzoek moeten in totaal ten minste 30 groepen of leraren zijn opgenomen (Kreft, 1998).
7. De interventie vindt plaats binnen een realistische school setting, en kan dus geïmplementeerd worden in het onderwijs. Onderzoeken waarin leraren bijvoorbeeld twee dagen per week besteden aan analyses met een DLVS worden bijvoorbeeld niet meegenomen, omdat deze tijdsbesteding niet haalbaar zal zijn binnen een realistische school setting.

Twintig studies voldeden bij nader inzien niet aan het inhoudelijke criterium (de definitie voor een DLVS), in tien studies stond het DLVS onvoldoende centraal in het onderzoek, daarnaast voldeden nog eens zestien studies niet aan alle methodologische criteria (zie Tabel 2). De referentielijsten van de geselecteerde onderzoeken zijn ook doorgenomen, hierin zijn nog 3 onderzoeken gevonden die voldeden aan alle criteria. In totaal zijn dus 15 onderzoeken meegenomen in de analyses. Een overzicht van de geselecteerde studies is terug te vinden in Bijlage 1.

Tabel 1 *Aantal gevonden studies in de databases (aantallen tussen haakjes zijn de aantallen zonder duplicaten)*

Database/data van raadplegen	**Eric (12-06-13)**	**Web of Science (19-06-13)**	**Scopus (21-06-13)**	**PsycINFO (24-06-13)**	**Narcis (15-07-13)**	**ProQuest (20-08-13)**
Trefwoorden stap 1	11.092	29.830	103.465	6.985	41	100
Onderzoeksmethode	1.197	3.461	10.436	846	-	11
Type onderwijs	406	35	78	90	-	5
Gepubliceerd na 1990	247	35	76	71	38	5
Voldeed aan criterium DLVS	***18***	***0***	***2(2)***	***3(1)***	***0***	***0***
Trefwoorden stap 2	32.850	88.194	219.482	20.623	203	551
Output onderwijs	5.538	niet gebruikt	niet gebruikt	niet gebruikt	-	niet gebruikt
Onderzoeksmethode	650	22.728	57.522	4.648	-	72
Type onderwijs	337	81	269	206	-	9
Gepubliceerd na 1990	243	81	159	188	-	9
Voldeed aan criterium DLVS	***19(12)***	***0***	***3(2)***	***9(3)***	***0***	***0***
Trefwoorden stap 3	75	773	1.100	86	0	2
Onderzoeksmethode	4	171	171	26	-	-
Type onderwijs	1	2	4	4	-	-
Gepubliceerd na 1990	-	-	-	-	-	-
Voldeed aan criterium DLVS	***0***	***0***	***0***	***0***	***0***	***0***
Totaal	***37(30)***	***0***	***5(4)***	***12(4)***	***0***	***0***

Tabel 2 *Onderzoeken die **niet** voldeden aan de methodologische criteria*

Auteurs en jaar van publicatie	***Toelichting***
1. Borman, Slavin, Cheung, Chamberlain, Madden and Chambers (2005) 2. Graney, and Shinn (2005) 3. Garet, Cronen, Eaton, Kurki, Ludwig, Jones, Uekawa, and Falk (2008) 4. Henderson, Petrosino, Guckenburg, and Hamilton (2007) 5. Henderson, Petrosino, Guckenburg, and Hamilton (2008) 6. McCaffrey, and Hamilton (2007) 7. McDowall, Cameron, Dingle, Gilmore, and MacGibbon (2007) 8. Parr, Timperley, Reddish, Jesson, and Adams (2007) 9. Phelan, Vendlinski, Choi, Dai, Herman, and Baker (2011) 10. Wijekumar, Hitchcock, Turner, Lei, and Peck (2009)	DLVS is te klein onderdeel van de interventie waardoor onduidelijk is wat de invloed van het DLVS is
11. Burns, Klingbeil, and Ysseldyke (2010) 12. Fuchs, Fuchs, Karns, Hamlett, Katzaroff and Dutka (1997) 13. Wayman, Cho, and Shaw (2009) 14. Williams, Swanlund, Miller, Konstantopoulos, Eno, van der Ploeg, and Meyers (in press)	Effecten op leerresultaten niet onderzocht (criterium 3)
15. Bolt, Ysseldyke, and Patterson (2010) 16. Betts, Youjin, and Zau (2011) 17. Cancino (2009) 18. Shaw, and Wayman (2012) 19. Wayman, Shaw, and Cho (2011)	Geen controlegroep (criterium 4)
20. Cole (2010) 21. Koedinger, McLaughlin, and Heffernan (2010) 22. Ysseldyke, Spicuzza, Kosciolek, Teelucksingh, Boys, and Lemkuil (2003) 23. Ysseldyke, Betts, Thill, and Hannigan (2004) 24. Calhoon and Fuchs (2003) 25. Vollands, Topping, and Evans (1999) 26. Spicuzza, Ysseldyke, Lemkuil, Kosciolek, Boys, and Teelucksingh (2001)	Geen random toewijzing aan experimentele en controlegroep (criteria 4), bovendien zijn er onvoldoende gegevens waaruit volgt dat de experimentele en controle- groep vergelijkbaar zijn (criterium 5) Kleine steekproef (criterium 6)

3.2. Analyse

Meta-analyse is een methode om de resultaten van verschillende kwantitatieve onderzoeken over hetzelfde onderwerp te combineren om een gemiddelde effectgrootte te bepalen. De resultaten uit verschillende onderzoeken kunnen gecombineerd worden wanneer ze gestandaardiseerd zijn. Deze gestandaardiseerde index is een effectgrootte. Voor het bepalen van deze effectgroottes zijn de formules: *Cohen's d,* en *Hedges' g* gebruikt (zie Bijlage 4), afhankelijk van het onderzoek (Lipsey & Wilson, 2000).

Voor het berekenen van de effectgroottes zijn de (*adjusted*) gemiddelden van de experimentele groep, de controlegroep, en de bijbehorende standaarddeviaties gebruikt, of zijn de waardes van de regressie coëfficiënten voor het treatment effect en de bijbehorende standaard fout gebruikt. In een aantal studies werden gemiddelden voor verschillende effect sizes. Er werd bijvoorbeeld een onderscheid gemaakt tussen het aantal correct geschreven woorden, en het aantal correct beantwoorde vragen.

Omdat in de meeste studies één gemiddelde effect size gehanteerd werd voor het vak taal zijn de gemiddelden voor de twee bovenstaande categorieën bijvoorbeeld gecombineerd. Ze werden alleen gecombineerd wanneer de berekende effect sizes betrekking hadden op hetzelfde DLVS en dezelfde interventie.

Voor het berekenen van de gepoolde standaarddeviatie waren gegevens nodig over de omvang van de experimentele, en controlegroep. Hiervoor werd het aantal leerlingen genomen waarvan de leerresultaten in de analyses waren gebruikt, tenzij in de studie de leerresultaten per leerjaar waren gemiddeld; in dat geval werd het aantal leraren genomen voor de bepaling van de omvang. In twee studies werden niet het aantal leerlingen gegeven maar het aantal opgenomen scholen en districten. Het gemiddelde aantal leerlingen per school werd wel gegeven, dit getal is gebruikt om te bepalen hoeveel leerlingen bij benadering in de treatmentgroep en in de controlegroep zaten.

Voor een meta-analyse kan voor de analyses gekozen worden uit twee verschillende modellen, het fixed effect model, en het random effect model. In het fixed model wordt er vanuit gegaan dat de steekproeven van de verschillende studies uit dezelfde populatie zijn getrokken en dat er dus één effectgrootte is. In het random model wordt er daarentegen vanuit gegaan dat de studies plaatsvonden binnen verschillende subpopulaties en dat de effecten gespreid zijn rond een gemiddeld effect. In deze meta-analyse is het random model gebruikt aangezien de studies in verschillende subpopulaties plaatsvonden, zoals het primair en secundair onderwijs, of het regulier en speciaal onderwijs.

In een meta-analyse worden gewichten toegekend aan de effecten uit de verschillende studies. Een nauwkeurig effect, dat wil zeggen een effect met een kleine variantie, krijgt een groter gewicht toegekend dan een minder nauwkeurig effect. In een random model wordt dit gewicht bepaald door de variantie binnen elke studie, en de variantie tussen de studies. Een studie met een kleinere variantie heeft dus een grotere invloed op het gemiddelde effect, dan een studie met een grotere variantie. Voor het toekennen van de gewichten en het bepalen van het gemiddelde effect is het programma Comprehensive Meta-analysis gebruikt. In dit programma worden de gewichten van de studies berekend met de standard error van de effecten.

3.2.1 Implementatiebevorderende factoren

Met behulp van Comprehensive Meta-analysis zijn ook de effecten van de zes implementatiebevorderende factoren bepaald. Voor de analyse van elke factor werden eerst verschillende subgroepen van studies samengesteld. Bijvoorbeeld een subgroep van studies met een hoge feedbackfrequentie, en een subgroep van studies met een lage feedbackfrequentie. Vervolgens werd dan met behulp van Comprehensive Meta-analysis bepaald of de gevonden effecten in beide subgroepen significant van elkaar afweken, en zo ja, in welke subgroep het effect significant hoger was.

4. RESULTATEN

Dit hoofdstuk start met een algemene beschrijving van de geselecteerde studies. De implementatiebevorderende factoren van elke studie zijn weergegeven in Tabel 3. Een nadere, meer uitgebreide inhoudelijke omschrijving van elke studie is terug te vinden in Bijlage 2. Nadat in paragraaf 4.1 een algemeen beeld geschetst is van de studies worden de resultaten van de meta-analyse beschreven.

4.1 Geselecteerde studies

Alle geselecteerde studies zijn afkomstig uit de USA en werden uitgevoerd tussen 1990 en 2013 binnen het primair onderwijs (3), het secundair onderwijs (1), of zowel in het primair als het secundair onderwijs (11). In zeven studies werden de resultaten van leerlingen uit het regulier onderwijs onderzocht, in vijf studies die van leerlingen uit het speciaal onderwijs. Daarnaast vonden drie studies weliswaar plaats in het reguliere onderwijs, maar waren daarvoor alleen leerlingen met leerachterstanden geselecteerd. De studies duurden tussen de vijftien weken en vier jaar. In elf studies werden effecten op rekenresultaten onderzocht, in zeven effecten op leesprestaties, drie studies waren gericht op de effecten op spelling.

In dertien studies werden de leraren, scholen of districten (schoolbesturen) random toegewezen aan een experimentele groep of een controlegroep. Binnen deze dertien studies waren er vijf waarin de random toegewezen leraren zelf leerlingen mochten selecteren voor de deelname aan het onderzoek. Uit analyses bleek dat er geen significante verschillen waren in belangrijke kenmerken (o.a. sekse, leeftijd, het verwachte beheersingsniveau, IQ en aantal jaren in speciaal onderwijs) tussen de leerlingen uit de experimentele groep en de controlegroep. In twee studies werd de experimentele groep gematcht aan een vergelijkbare controlegroep.

Met de vijftien studies zijn gegevens van meer dan 696.150 leerlingen, 2334 leraren, 2044 scholen en 366 schoolbesturen verzameld. Niet in alle onderzoeken werd het aantal leerlingen, leraren, scholen en besturen vermeld, deze gegevens zijn met andere woorden een schatting. In totaal zijn 72 effecten bepaald waarvan er in de analyses 40 effecten zijn opgenomen.

In het tweede hoofdstuk van dit rapport werden de implementatiebevorderende factoren gepresenteerd waarvan aangenomen wordt dat ze het effect van een DLVS op leerresultaten beïnvloeden. In Tabel 3 wordt per studie informatie gegeven over deze factoren. Binnen elke factor worden verschillende categorieën onderscheiden. De zeven factoren zijn per studie gescoord (per factor worden 2 á 3 categorieën aangegeven):

1. Feedbackfrequentie
    a. Laag: leraren ontvingen één keer per schooljaar feedback
    b. Midden: leraren ontvingen ten minste drie keer per schooljaar feedback
    c. Hoog: leraren ontvingen tenminste maandelijks feedback
2. Inhoud van de feedback
    a. Scores: alleen de scores werden teruggekoppeld
    b. Ontwikkeling: de leerontwikkeling over een langere periode werd teruggekoppeld
    c. Leerstofinhoudelijk: de mate van beheersing van specifieke leercategorieën werd teruggekoppeld
    d. Instructieadvies: in combinatie met de feedback werd instructieadvies teruggekoppeld

3. Omvang en duur van de interventie
    a. 1 á 2 keer per jaar een contactmoment
    b. 3 á 5 keer per jaar een contactmoment
    c. Maandelijkse contactmomenten
4. Inhoud van de interventie
    a. Technisch: er werd technische uitleg gegeven over het DLVS-en/of de toetsen
    b. Nadenken over de instructie: leraren werd uitgelegd hoe zij de feedback konden vertalen naar de instructie
    c. Begeleiding bij de uitvoering: lessen van leraren werden geobserveerd en aan de hand daarvan werden leraren begeleid bij het toepassen van de feedback
5. Doelgroep
    a. Leraren
    b. Schoolleiders: naast leraren waren ook schoolleiders betrokken bij het leren gebruiken van de feedback
    c. Bestuurders: naast leraren en schoolleiders waren ook de bestuurders betrokken bij het leren gebruiken van de feedback
6. Feedback over het niveau van groep(en) (groepsgemiddelden)
    a. Nee: de resultaten van individuele leerlingen werden niet omgezet naar het groeps-, en/of schoolniveau
    b. Ja: de resultaten van individuele leerlingen werden wel omgezet naar het groeps-, en/of schoolniveau
7. Benchmarking
    a. Nee: op basis van behaalde toetsresultaten konden leraren, of scholen niet met (landelijke) standaarden vergeleken worden
    b. Ja: op basis van behaalde toetsresultaten konden leraren, of scholen wel met (landelijke) standaarden vergeleken worden

Feedbackfrequentie

Uit Tabel 3 volgt dat in tien studies de leraren tenminste één keer per maand feedback ontvingen. In een aantal studies ontvingen de leraren twee keer per week feedback. In vier studies ontvingen de leraren tussen de drie en zeven keer per jaar feedback, en in één studie één keer per jaar.

Inhoud van de feedback

In twee studies bestond de feedback alleen uit scores op toetsen. Dit waren studies waarin de interventies niet alleen op leraren, maar vooral ook op schoolleiders en bestuurders gericht waren. In acht studies bestond de feedback niet alleen uit scores, maar ook uit informatie over de ontwikkeling van leerlingen, en informatie over de beheersing van specifieke leerstofcategorieën. In veel DLVS-en werd advies gegeven over instructievormen en verwerkingsmaterialen, bij vijf systemen waren dit adviezen die gezien de feedback zouden moeten resulteren in afgestemd onderwijs. In de andere systemen stond het advies los van de feedback.

Omvang, duur en inhoud van de interventie

Niet in elke studie wordt voldoende informatie gegeven over de omvang, de duur, en de inhoud van de interventie. In een aantal studies worden de effecten van commerciële DLVS-en onderzocht. De ontwikkelaars van deze systemen boden scholen vaak een aantal ondersteuningsmogelijkheden aan. Niet de onderzoekers, maar de ontwikkelaars van de DLVS-en verzorgden in dat geval de interventies. Dit verklaart misschien waarom de interventies niet systematisch aangeboden werden, en waardoor

niet bekend is hoe vaak, of hoe lang leraren daadwerkelijk contact hadden met de consulenten van de betreffende ontwikkelaar. Van twee studies is de omvang van de interventie daardoor onbekend.

In negen studies kregen leraren tenminste maandelijks training of begeleiding, in twee studies was de feedbackfrequentie drie tot vijf keer per jaar, en in twee studies was dit één á twee keer per jaar. De inhoud van de interventies bestond in acht studies uit lesobservaties en begeleiding bij het implementeren van de feedback. In vier studies ontvingen leraren training, of advies over hoe zij de feedback konden omzetten naar interventies, en in twee studies bestond de interventie alleen uit technische informatie over het DLVS. Dit laatste hield in dat leraren informatie ontvingen over de verschillende mogelijkheden van het DLVS, en welke functies van het systeem zij moesten gebruiken voor het invoeren, en/of analyseren van de toetsgegevens.

Doelgroep

In twaalf studies was de interventie alleen op leraren gericht. Dit is opvallend aangezien uit de literatuurbeschouwing in hoofdstuk twee blijkt dat schoolleiders een belangrijke invloed hebben op het succes van dergelijke systemen. In twaalf studies werd de schoolleider bij de beschrijving van de interventie niet genoemd. In drie studies waren schoolleiders en schoolbestuurders wel betrokken bij de interventie, in twee van deze studies hadden de leraren een beduidend kleinere rol dan de schoolleiders en bestuurders.

Groepsgemiddelden

Er waren zes studies waarin de leraren naast individuele leerresultaten van leerlingen ook leerresultaten op het niveau van de groep gebruikten voor het vormgeven van de instructie. Het gebruiken van groepsgemiddelden als feedback kwam vaker voor in de recente studies; de studies van May (2007) en Fuchs (1994) vormden hierop een uitzondering. Tevens waren in de studies waarin groepsgemiddelden als feedback gebruikt werden vaker schoolleiders en bestuurders betrokken. Dit was in drie studies het geval.

Benchmarking

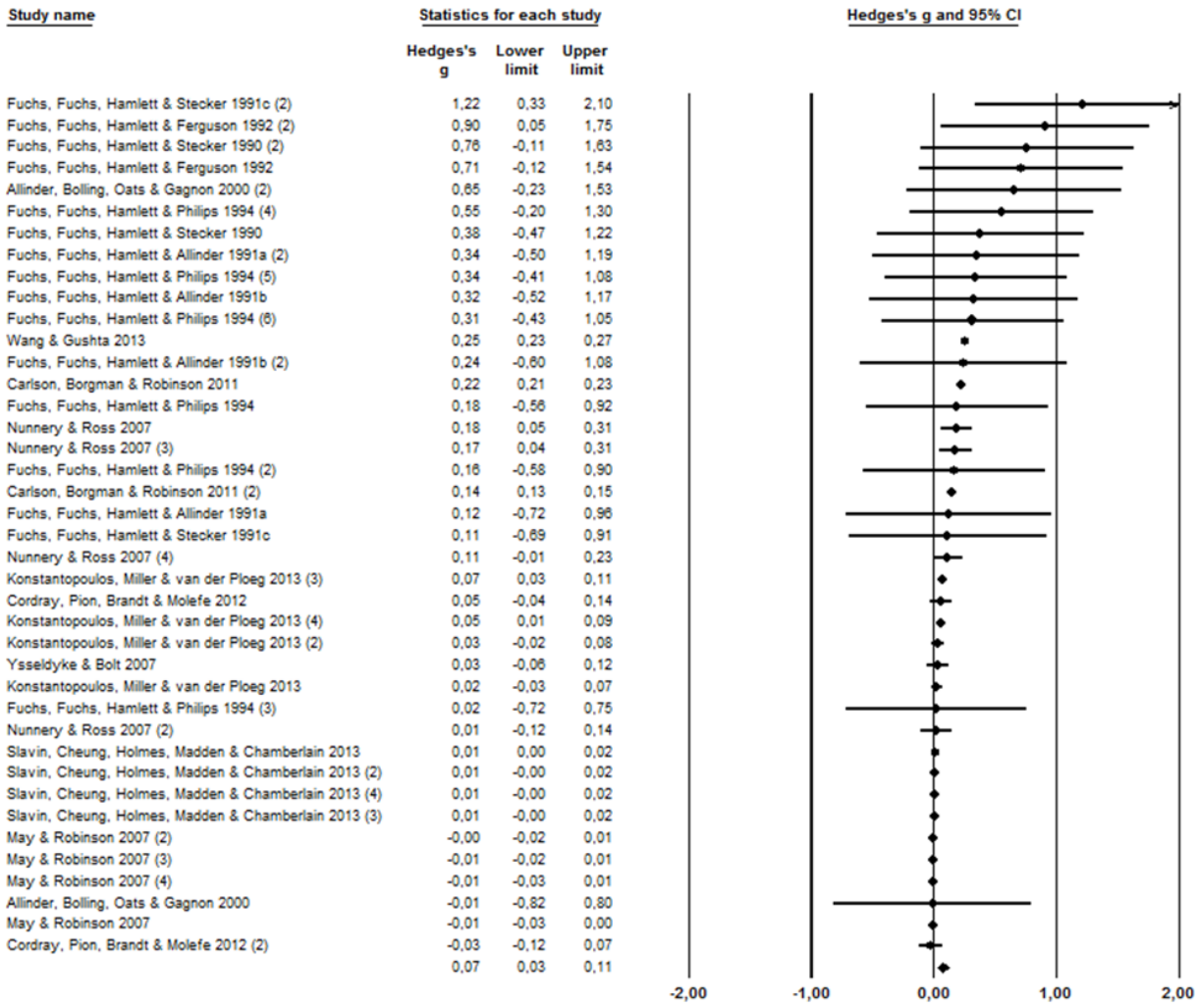
In zes studies werden de leerresultaten van leerlingen vergeleken met landelijke normen, of standaarden. Ook deze factor kwam vaker voor in recente studies, en in studies waarin schoolleiders en bestuurders betrokken waren bij de interventies. In vier studies maakten de leraren of scholen zowel van groepsgemiddelden als van benchmarking gebruik voor het vormgeven van het onderwijs.

Tabel 3 *Implementatiebevorderende factoren in de geselecteerde studies*

Studie	**Frequentie FB** a. laag b. midden c. hoog	**Inhoud FB** a. scores b. ontwikkeling c. leerstofinhoudelijk d. instructie advies	**Omvang interventie** a. 2-1x per jaar b. 3-5x per jaar c. maandelijks	**Inhoud interventie** a. technisch b. nadenken over instructie c. begeleiding bij uitvoering	**Doelgroep** a. bestuurders b. schoolleider c. leraren	**Groeps-gemiddelden** a. Ja b. Nee	**Bench-marking** a. Ja b. Nee	**Range gewogen effecten**
1. Allinder	hoog	leerstofinhoudelijk	maandelijks	nadenken over instructie	leraren	nee	nee	-0.01/0.68
2. Carlson	midden	scores	maandelijks	nadenken over instructie	+bestuurders	ja	ja	0.14/0.22
3. Cordray	midden	leerstofinhoudelijk	3-5x per jaar	begeleiding bij uitvoering	leraren	nee	ja	-0.03/0.05
4. Fuchs 1990	hoog	leerstofinhoudelijk	maandelijks	begeleiding bij uitvoering	leraren	nee	nee	0.39/0.79
5. Fuchs 1991a	hoog	leerstofinhoudelijk	maandelijks	begeleiding bij uitvoering	leraren	nee	nee	0.12/0.36
6. Fuchs 1991b	hoog	instructieadvies	maandelijks	begeleiding bij uitvoering	leraren	nee	nee	0.25/0.34
7. Fuchs 1991c	hoog	instructieadvies	maandelijks	begeleiding bij uitvoering	leraren	nee	nee	0.11/1.22
8. Fuchs 1992	hoog	instructieadvies	maandelijks	begeleiding bij uitvoering	leraren	nee	nee	0.74/0.94
9. Fuchs 1994	hoog	instructieadvies	maandelijks	begeleiding bij uitvoering	leraren	ja	nee	0.02/0.57
10.Konstantopoulos	midden	leerstofinhoudelijk	2-1x per jaar	onbekend	leraren	ja	ja	0.02/0.07
11. May	laag	leerstofinhoudelijk	2-1x per jaar	technisch	+besturen	ja	ja	-0.01/0
12. Nunnery	hoog	leerstofinhoudelijk	onbekend	nadenken over instructie	leraren	nee	nee	0.01/0.18
13. Slavin	midden	scores	maandelijks	begeleiding bij uitvoering	+bestuurders	ja	ja	0.01
14. Wang	hoog	instructieadvies	onbekend	technisch	leraren	ja	ja	0.25
15. Ysseldyke	hoog	leerstofinhoudelijk	3-5x per jaar	nadenken over instructie	leraren	nee	nee	0.03

4.2 Analyse resultaten

In de analyse zijn 40 gewogen effecten opgenomen die varieërden tussen de 1.22 en -0.03 (zie Figuur 2). Het gemiddeld gewogen effect van 0.072 (SE=0.021 en $p < .00$) wijkt significant af van nul (Tabel 4).



Figuur 2 Gewogen effectgroottes en betrouwbaarheidsintervallen van de geselecteerde studies

Uit Tabel 4 volgt daarnaast dat de variatie in de gewogen effecten groter is dan verwacht kan worden op basis van de standaardfout i.v.m. de steekproeftrekking (Q=2997.572, df=39 en $p < .00$). Uit deze gegevens volgt dat het random model beter aansluit bij de onderzoeksgegevens, dan het fixed model.

Tabel 4 *Gemiddeld gewogen effect*

						Test nul hypothese		**Test heterogeniteit**		
Model	**k***	**ES****	**SE*****	**95% betrouwbaarheidsinterval**		*z-waarde*	*p-waarde*	*Q-waarde*	*df (Q)*	*p-waarde*
Fixed	40	0.078	0.002	0.075	0.082	46.748	0.000	2997.572	39	.00
Random	40	0.072	0.021	0.030	0.114	3.378	0.001			

Noot: * k=aantal meegenomen effecten, ** ES=effectsize, *** SE=standard error effectsize.

De omvang van de steekproef van de studies verschilt sterk. Het gemiddelde effect is daarom ook bepaald voor de groep studies met een relatief kleine steekproefomvang (minder dan 1000 leerlingen) en voor de groep studies met een relatief grote steekproefomvang (meer dan 1000 leerlingen). De omvang van de steekproef in de 'kleine' studies varieerde tussen de 54 en 917 leerlingen en de omvang van de steekproef in de 'grote' studies varieerde tussen de 1880 en ± 300.000 leerlingen (zie Bijlage 1 voor de omvang van de steekproef in elke studie). Het gevonden effect in studies met een kleine steekproefomvang was 0.397 ($p < 0.001$) en in de studies met een grote steekproefomvang 0.057 ($p < 0.009$) (Tabel 5).

Tabel 5 *Gemiddeld gewogen effect naar steekproefgrootte (meer of minder dan 1000 leerlingen)*

Aantal leerlingen	k*	ES**	SE***	95% betrouwbaarheidsinterval		p-waarde ES	p-waarde verschil
Meer dan 1000	22	0.057	0.022	0.014	0.100	0.009	
Minder dan 1000	18	0.397	0.103	0.195	0.600	0.00	.001

Noot: * k=aantal meegenomen effecten, ** ES=effectsize, *** SE=standard error effectsize.

4.2.1 Implementatiebevorderende factoren

Tabel 6 biedt een overzicht van de resultaten betreffende de bestudeerde implementatiebevorderende factoren. In de tabel geeft de letter *k* het aantal meegenomen effecten weer.

De factor 'feedbackfrequentie' was onderverdeeld in drie categorieën: een hoge, een midden en een lage feedbackfrequentie. Uit Tabel 6 volgt dat de gemiddelde effectgroottes van de categorieën van elkaar afweken ($p < .007$), en dat de richting daarvan in overeenstemming is met de verwachting. Het grootste effect werd gevonden voor een hoge feedbackfrequentie (ES=0.171, *k*=24), kleinere effecten werden gevonden voor de middencategorie (ES=0.051, *k*=12), en de categorie lage feedbackfrequentie (ES=-0.008, *k*=4). Alleen het effect van een hoge feedbackfrequentie week significant af van nul ($p < .00$).

Binnen de factor 'inhoud van de feedback' werd een onderscheid gemaakt tussen vier categorieën. Uit de analyse volgt dat de gemiddelde effecten van deze categorieën significant van elkaar afweken ($p < .008$), de richting daarvan kwam echter niet volledig overeen met de verwachting. Het grootste effect werd gevonden voor de categorie 'ontwikkeling' (ES=0.421, *k*=3) in plaats van, zoals verwacht, voor de categorie 'instructieadvies' (ES=0.314, *k*=7). Het gevonden effect voor de categorie 'ontwikkeling' was echter niet significant ($p < .109$), het effect van 'instructieadvies' wel ($p < .00$). Gevonden effecten in de categorieën 'leerstofinhoudelijk' en 'scores' waren respectievelijk 0.045 (*k*=24) en 0.065 (*k*=6). Alleen het effect van de categorie instructieadvies week significant af van nul.

Voor de factor 'omvang van de interventie' werden geen statistisch significante verschillen gevonden tussen de drie categorieën ($p < .057$). De grootste effecten werden gevonden voor de studies waarin leraren elke maand ondersteuning ontvingen in het kader van een interventie (ES=0.103, *k*=24). Kleinere effecten werden gevonden voor de studies waarbinnen de leraren, of scholen drie tot vijf keer per jaar ondersteund werden (ES=0.018, *k*=3), of één á twee keer per jaar (ES=0.017, *k*=8). Alleen het effect van de categorie 'maandelijks' week significant af van nul ($p <.003$).

Voor het verschil tussen de gemiddelde effecten binnen de categorieën voor de vierde factor 'inhoud van de interventie' werd een *p*-waarde van < 0.016 gevonden. Gevonden effecten binnen de drie categorieën waren 0.029 (*k*=22) voor interventies waarin de inhoud ook gericht was op begeleiding bij de uitvoering, 0.135 (*k*=9) voor interventies waarin de inhoud van de interventie ook gericht was op de vertaling naar de instructie, en 0.044 (*k*=5) voor de interventies waarin alleen technische informatie gegeven werd over het DLVS en/of de toetsen. Alleen het effect van 'vertaling naar instructie' was significant ($p < .00$).

De factor ‘doelgroep’ was opgesplitst in drie categorieën. Geen enkele studie kon geplaatst worden in de categorie ‘schoolleider’, er waren dus geen studies met een interventie die gericht was op schoolleiders en leraren. Wanneer de schoolleiders betrokken werden bij de interventie dan waren de bestuurders ook altijd betrokken bij de interventie. Daarom werden alleen de categorieën leraren, en bestuurders opgenomen in de analyse, deze twee bleken niet significant van elkaar af te wijken ($p < .112$). Het gevonden effect voor interventies waarin zowel leraren als schoolleiders en bestuurders betrokken werden was 0.036 (k=10), en 0.104 (k=30) voor de interventies waarin alleen leraren betrokken werden. Alleen het laatste effect week significant af van nul ($p < .00$).

Tabel 6 *Resultaten analyses implementatiebevorderende factoren*

Factor	Categorie	k	ES	95% betrouwbaarheidsinterval		p-waarde ES	p-waarde verschil
Frequentie FB	hoog	24	0.171	0.093	0.249	.00	
	midden	12	0.051	-0,002	0.103	.060	
	laag	4	-0.008	-0.097	0.081	.864	
							.007
Inhoud FB	instructieadvies	7	0.314	0.156	0.472	.00	
	leerstofinhoudelijk	24	0.045	-0.004	0.094	.07	
	ontwikkeling	3	0.421	-0.094	0.937	.109	
	scores	6	0.065	-0.007	0.138	.076	
							.008
Omvang interventie	maandelijks	24	0.103	0.035	0.172	.003	
	3-5 keer per jaar	3	0.018	-0.097	0.133	.759	
	2-1 keer per jaar	8	0.017	-0.047	0.080	.607	
							.057
Inhoud interventie	begeleiding bij uitvoering	22	0.029	-0.017	0.076	.216	
	vertaling naar instructie	9	0.135	0.083	0.188	.00	
	technisch	5	0.044	-0.006	0.093	.082	
							.016
Doelgroep	bestuurders	10	0.036	-0.024	0.097	.241	
	schoolleiders	0	-	-	-	-	
	leraren	30	0.104	0.046	0.162	.00	
							.112
Groepsgemiddelden	ja	21	0.057	0.008	0.107	.023	
	nee	19	0.110	0.030	0.190	.007	
							.271
Benchmarking	ja	17	0.048	0.001	0.096	.045	
	nee	23	0.157	0.067	0.248	.001	
							.037

Studies waarin leraren zowel groepsgemiddelden als individuele leerresultaten gebruikten verschilden niet significant van de studies waarin alleen individuele leerresultaten werden gebruikt ($p < .271$). Bovendien werd de verwachte richting niet gevonden. In de studies waarin de leraren geen groepsgemiddelden gebruikten werden gemiddeld hogere effecten gevonden (ES=0.110, k=19), dan in de studies waarin dit wel het geval was (ES=0.057, k=21).

Voor het verschil binnen de laatste implementatiebevorderende factor werd een p-waarde van < .037 gevonden. De richting van het verschil stemde niet overeen met de verwachting. Het gemiddelde

effect in de studies waar leraren de behaalde leerresultaten konden vergelijken met normen, of met referentiegroepen was 0.048 (*k*=17). In de studies zonder benchmarking werd een hoger gemiddeld effect van 0.157 (*k*=23) gevonden.

4.2.2 Aanvullende analyses

Om de gevonden variantie tussen de studies nog nader te onderzoeken is een aantal aanvullende analyses uitgevoerd voor de volgende kenmerken:

- onderzoeksperiode
- vakgebied
- onderwijssoort
- de tijdsduur van het onderzoek.

Omdat in een groot gedeelte van de studies de leerresultaten van leerlingen uit het primair en secundair onderwijs waren samengenomen konden de effecten voor het primair en secundair onderwijs niet apart worden geanalyseerd. Tabel 7 bevat de resultaten van de aanvullende analyses. Significante verschillen tussen de afzonderlijke categorieën werden gevonden voor de onderzoeksperiode, de onderwijsssoort en de tijdsduur van het onderzoek.

In de studies die plaatsvonden vòòr 2000 (ES=0.397, *k*=18, $p < .00$) werden hogere effecten gevonden dan in de studies die daarna plaatsvonden.

Er werden geen significante verschillen gevonden tussen de verschillende vakgebieden. Voor het vakgebied rekenen werd wel een significant positief effect gevonden (ES=0.107, *k*=22, $p < .001$). De gevonden effecten voor lezen, spelling en science weken niet significant af van nul.

Tabel 7 *Resultaten analyses aanvullende kenmerken*

Kenmerk	**Categorie**	**k**	**ES**	**95% betrouwbaarheidsinterval**		**p-waarde ES**	**p-waarde verschil**
Periode	tot 2000	18	0.397	0.196	0.599	.00	
	tussen 2000 en 2010	9	0.041	-0.026	0.108	.230	
	na 2010	13	0.067	0.015	0.118	.012	
							.004
Vakgebied	rekenen	22	0.107	0.041	0.173	.001	
	lezen	12	0.051	-0.017	0.120	.143	
	spelling	5	0.043	-0.146	0.232	.655	
	science	1	-0.008	-0.216	0.200	.940	
							.555
Onderwijssoort	regulier onderwijs	22	0.040	0.013	0.066	.003	
	regulier: lage leerresultaten	6	0.187	0.115	0.260	.00	
	speciaal onderwijs	12	0.394	0.145	0.643	.002	
							.00
Duur onderzoek	korter dan een schooljaar	16	0.408	0.196	0.619	.00	
	een schooljaar	14	0.067	0.021	0.114	.004	
	langer dan een schooljaar	10	0.043	-0.013	0.099	.136	
							.005

Hoge significante effecten werden gevonden in de studies die plaatsvonden binnen het speciaal onderwijs (ES=0.394, *k*=12, *p* < .002), deze studies vonden grotendeels ook plaats voor het jaar 2000. De laagste effecten werden gevonden in de studies die plaatsvonden binnen het reguliere onderwijs (ES=0.040, *k*=22, *p* < .003).

In de onderzoeken die korter dan een jaar duurden werden de hoogste effecten gemeten, gemiddeld 0.408 (*k*=16, *p* < .00). De effecten in de onderzoeken die een schooljaar (ES=0.067, *k*=14, *p* < .004), of langer dan een schooljaar duurden (ES=0.043, *k*=10, *p* <.136) waren een stuk kleiner.

5. CONCLUSIE EN DISCUSSIE

Onderzoeksvraag 1: Wat is het effect van digitaal leerlingvolgsysteemgebruik door leraren op de prestaties van leerlingen?

Het effect van formatief toetsen op leerresultaten is sinds de jaren negentig veel onderzocht (Wiliam, 2011). In dit onderzoek is het effect van formatief toetsen in combinatie met het gebruiken van een DLVS door leraren onderzocht. In de analyses is een gemiddeld gewogen effect van 0.07 gevonden (Tabel 4). Om te bepalen wat dit getal zegt over het effect van het gebruik van een DLVS op leerresultaten is het noodzakelijk om een juiste normering te gebruiken. Het is daarbij van belang om rekening te houden met de verdeling van effectgroottes in vergelijkbare onderzoeken, dus onderzoeken waarin vergelijkbare interventies in vergelijkbare populaties onderzocht werden. Een scherp afgebakend onderzoeksonderwerp en bijbehorende selectiecriteria zijn daarom ook noodzakelijk in een meta-analyse. Hiermee wordt voorkomen dat in de analyses 'appels met peren' vergeleken worden.

In dit onderzoek is gezorgd voor een scherpe afbakening, door alleen studies te selecteren gericht op de analyse van de effecten van het gebruik van een DLVS: systemen waarmee toetsresultaten geanalyseerd worden en leraren de resultaten van die analyses vervolgens ontvangen, zodat ze deze feedback kunnen gebruiken voor het vormgeven van hun instructie. Er is niet geselecteerd op DLVS-studies waarin de interventies om de benutting van een DLVS te bevorderen inhoudelijk volstrekt identiek zijn. De interventies in de in deze meta-analyse opgenomen studies variëren daarom onderling.

Een tweede kenmerk waarop de studies van elkaar verschillen betreft de omvang van de steekproef. In de kleinste studie worden de leerresultaten van 54 leerlingen geanalyseerd, in de grootste studie de leerresultaten van ongeveer 300.000 leerlingen. In onderzoeken met een kleine steekproef worden vaak grotere positieve interventie-effecten gevonden, dan in de grotere studies. Dit heeft verschillende oorzaken. Allereerst lopen kleine onderzoeken een groter risico om niet gepubliceerd te worden wanneer daarin geen positieve effecten gevonden worden. Daarnaast kan in kleine onderzoeken vaak meer tijd besteed worden aan de implementatie van de interventie (Slavin & Smith, 2008b). Dit laatste geldt zeker ook voor de in deze meta-analyse opgenomen kleine onderzoeken. Bovendien werden in de kleine studies vaker toetsen gebruikt die door de auteurs zelf waren ontworpen, of door de ontwikkelaars van het onderzochte DLVS. De studies van Nunnery en Ross (2007) en Ysseldyke en Bolt (2007) vormen hierop een uitzondering, in de eerste 'kleine' studie wordt namelijk een *state* toets gebruikt en in de tweede 'grote' studie een eigen toets. In de grotere onderzoeken werden *state* toetsen gebruikt voor het bepalen van het effect. Een zelfgemaakte toets zal doorgaans beter aansluiten bij het DLVS, en de net onderwezen leerinhoud, dan een algemene toets zoals een *state* toets. Onderzoeken waarin zelfgemaakte toetsen gebruikt werden kunnen daarom eerder resulteren in grotere effecten. In Bijlage 1 wordt vermeld welke toetsen zijn gebruikt in de geselecteerde onderzoeken.

Omdat de interventies, en de omvang van de steekproeven in de geselecteerde studies sterk verschillen is het zinvol om de gemiddelde effecten te bepalen in twee verschillende groepen studies. Namelijk één groep studies met een relatief kleine steekproef en een intensieve interventie voor een kleinere groep leraren en leerlingen, en één groep studies met een minder intensieve interventie en een grotere steekproef. Het gemiddeld effect in de eerste groep is 0.397 (met een standard error van 0.103), het gemiddelde effect in de tweede groep is 0.057 (met een standard error van 0.022) (Tabel 5). Het eerste effect wijkt significant af van nul ($p < .00$) de p-waarde van het tweede effect ($p < .009$) is groter.

Uit het rapport van Lipsey (2012) en collega's volgt dat voor onderwijsinterventies gericht op:

- *individuele leerlingen* een gemiddeld effect van 0.4 gevonden wordt,
- *kleine groepen leerlingen* een gemiddeld effect van 0.26 wordt gevonden,
- *klassen* een gemiddeld effect van 0.18 gevonden wordt, en
- dat voor interventies gericht op *scholen* een effect van gemiddeld 0.10 wordt gevonden.

Het eerste effect kunnen we beoordelen door het te vergelijken met het gemiddelde effect van onderwijsinterventies die gericht waren op kleine groepen leerlingen, dit was immers ook het geval in de meerderheid van deze studies. Uit deze vergelijking volgt dat we kunnen spreken van een gemiddeld tot hoog effect van een DLVS op de leerresultaten van leerlingen.

In de tweede groep studies is een effect van 0.057 gevonden. Dit is in vergelijking met andere interventies die gericht waren op gehele scholen een relatief klein effect (daar was gemiddeld sprake van effecten van 0.10). Effecten van interventies op schoolniveau lijken een minder grote impact te hebben op leerresultaten dan effecten van interventies die gericht zijn op individuele leerlingen. De impact kan op een andere manier echter op schoolniveau ook groot zijn aangezien er veel meer leerlingen zijn die door de interventie op schoolniveau hogere leerresultaten behalen. Een aantal studies was niet gericht op scholen maar zelfs op gehele schoolbesturen (districten), voor deze effecten is geen goede normering gevonden. De interventies die op dit niveau gericht zijn zullen waarschijnlijk resulteren in gemiddeld kleinere effecten, dan de interventies die gericht zijn op het niveau van de school.

De relatief hoge effectgrootte van een DLVS in de studies aangaande een interventie gericht op kleine groepen leerlingen werd niet gevonden in de studies met een interventies gericht op alle leerlingen binnen scholen, of besturen. Het is de moeite waard om nader te onderzoeken hoe de relatief succesvolle aanpakken van DLVS-gebruik voor kleine groepen leerlingen vertaald kunnen worden naar bruikbare en nog effectievere aanpakken op het school- en/of bestuursniveau.

Geringere effecten op schoolniveau kunnen allereerst veroorzaakt worden door het feit dat het moeilijker is om intensief met alle leraren van een school te werken. Een andere verklaring voor het relatieve kleine effect op het niveau van de school zou kunnen zijn dat scholen in de controlegroepen ook feedback op basis van toetsgegevens benutten, en/of in het bezit zijn van een ander DLVS. Zeker in de USA (waar de geanalyseerde studies plaatsvonden) gebruiken veel scholen tegenwoordig systemen voor de opslag en analyse van feedback op basis van toetsen. Controlescholen bevatten nooit precies het DLVS waarvan het effect onderzocht werd, maar onderzoekers konden vaak ook niet geheel uitsluiten dat deze scholen geen enkel ander DLVS gebruikten.

Onderzoeksvraag 2: *Welke factoren belemmeren dan wel bevorderen het beoogde effect van digitaal leerlingvolgsysteemgebruik op leerprestaties?*

Voor het beantwoorden van de tweede onderzoeksvraag geeft de variatie in DLVS-en en interventies extra mogelijkheden, omdat hierdoor de verschillende DLVS-en interventies vergeleken konden worden om te bepalen welke varianten effectief zijn.

De resultaten in Tabel 6 laten zien dat de hoogste effecten op leerresultaten gevonden werden voor DLVS-en:

- die frequent feedback gaven (tenminste maandelijks),
- die gepaard gingen met een interventie die minimaal één keer per maand plaatsvond,

- die gecombineerd werden met een interventie waarin advies werd gegeven over de benodigde instructie,
- en die daarnaast feedback gaven met daarin advies over benodigde instructie en verwerkings-opdrachten.

Het laatste punt komt overeen met de bevindingen uit een meta-analyse naar de effecten van feedback in een computergestuurde omgeving. Hierin werd gevonden dat uitgebreide feedback, waarin bijvoorbeeld uitleg gegeven werd over waarom een antwoord juist of onjuist is, tot hogere effecten leidt dan feedback welke alleen betrekking had op de correctheid van een antwoord (Van der Kleij, Feskens & Eggen, in press).

Voor de interventies waarin advies gegeven werd over de vertaling van de feedback naar de instructie, werd een hoger effect gevonden dan voor andere interventies. Maar het verschil met de andere categorieën is statistisch niet significant ($p < .016$).[1]

De analyses betreffende de overige onderzochte implementatiebevorderende factoren resulteerden niet in significante verschillen; geen van de categorieën binnen deze factoren gaat samen met de verwachte, significant hogere leerresultaten. Van hogere leerresultaten was wel sprake in het geval van de factor 'benchmarking', maar de richting van het effect kwam niet overeen met de verwachting. Er zijn juist hogere effecten gevonden in de studies waarin leraren de leerresultaten niet vergelijken met normen, of standaarden. De verwachting was dat leraren door het vergelijken van leerresultaten beter kunnen beoordelen of de ontwikkeling van leerlingen naar verwachting verloopt, tijdig risico's signaleren, en, wanneer wenselijk, de instructie aanpassen. Een mogelijke verklaring voor het vinden van een significant hoger effect in de studies waar leraren de leerresultaten niet vergelijken met normen, zou kunnen zijn dat er geen sprake van benchmarking was in de studies met een kleine steekproef en een intensievere interventie, en juist wel in de grotere studies die gericht waren op het school-, of bestuursniveau. Wat het gebruik van 'groepsgemiddelden', of 'feedback over het niveau van de groep' betreft geldt iets soortgelijks als bij 'benchmarking'. Omdat ook deze factor voornamelijk gebruikt wordt in de grotere onderzoeken met een minder intensieve interventie is het effect van een DLVS kleiner wanneer groepsgemiddelden gebruikt worden, dan wanneer deze niet gebruikt worden.

Binnen de implementatiebevorderende factor 'inhoud van de feedback' is onderscheid gemaakt tussen vier categorieën. Voor de categorieën 'ontwikkeling', 'leerstofinhoudelijk' en 'scores' zijn geen effecten gevonden die significant afwijken van nul. Dit is wel het geval voor de categorie 'instructieadvies'. Zoals verwacht worden de hoogste effecten gevonden in de studies waarin het DLVS in aansluiting op de feedback advies geeft over de gewenste instructie. Het effect van DLVS-en waarmee alleen feedback gegeven wordt over de leerstofinhoudelijke beheersing resulteert in een veel lager effect. Binnen deze categorie ontvangen leraren feedback waarmee aangeven wordt welke type vragen leerlingen niet goed beantwoorden, en wat het leerstofonderdeel was dat met vragen getoetst werd. Hieruit kunnen leraren dus afleiden welk leerstofonderdeel men nog aan bod moeten laten komen in de instructie. Een mogelijke verklaring voor het lage effect van deze categorie is dat het leraren te weinig (nieuwe) informatie geeft over waardóór leerlingen geen goede antwoorden geven, en daarmee dus ook te weinig aangrijpingspunten hebben voor hoe de instructie beter op de

[1] Voor het beantwoorden van de onderzoeksvragen zijn diverse analyses uitgevoerd op dezelfde data, er is m.a.w. sprake van *multiple testing*. De kans op het vinden van een statistisch significant effect door toeval neemt toe, het is daarom gebruikelijk om een strenger significantieniveau dan 0.05 aan te houden. Bij het trekken van conclusies op basis van de resultaten is hiermee rekening gehouden.

onderwijsbehoeften van leerlingen kan orden afgestemd. Een ander belangrijk punt dat hier een rol kan spelen is de mate waarin leraren in staat zijn om bij geconstateerde achterstanden van leerlingen adequate 'remedies' te bepalen en aan te bieden .

Feedback waarin alleen informatie wordt gegeven over de beheersing van verschillende leerstofcategorieën leidt dus in mindere mate tot een verhoging van de leerresultaten, dan feedback waarin daarnaast ook concrete adviezen over de gewenste instructie, en verwerkingsopdrachten wordt gegeven.

Naast de analyses op de genoemde implementatiebevorderende factoren is nog een aantal aanvullende analyses uitgevoerd. De resultaten van deze analyses zijn terug te vinden in Tabel 7. Uit de analyses volgt dat DLVS-en die binnen het speciaal onderwijs ingezet worden hogere effecten hebben, dan DLVS-en die in het regulier onderwijs worden gebruikt. In de studies die voor het jaar 2000 plaatsvonden, en korter dan een jaar duurden zijn significante hogere effecten gevonden, dan in de studies die na 2000 plaatsvonden en langer dan een jaar duurden. Deze bevinding overlapt deels met de eerdere analyse, omdat de studies binnen het speciaal onderwijs ook allemaal voor, of in 2000 plaatsvonden en een jaar, of korter duurden. Wanneer de DLVS-en gebruikt worden voor het verzorgen van onderwijs aan leerlingen met lage leerresultaten, dan resulteren deze in hogere effecten, dan wanneer het systeem gebruikt wordt voor leerlingen in het reguliere onderwijs. Uit de analyses volgt tot slot dat DLVS-gebruik voor het vakgebied rekenen de hoogste effecten oplevert, dit effect wijkt echter niet significant af van de effecten gevonden voor de vakgebieden lezen en spelling.

Samenvattend:

- DLVS-gebruik resulteert in vergelijking met andere interventies die gericht zijn *op kleine groepen leerlingen* in een positief, en relatief groot effect op leerresultaten.
- DLVS-gebruik resulteert in vergelijking met andere interventies die gericht zijn op *scholen* in een iets lager dan gemiddeld effect op leerresultaten.
- DLVS-en die minimaal *maandelijks feedback* aan leraren geven hebben een grotere positieve invloed op de leerresultaten.
- DLVS-en die naast feedback *ook advies geven over de instructie*, en over de *verwerkingsopdrachten* die aansluiten bij de feedback, hebben een sterkere positieve invloed op de leerresultaten.
- Een goede interventie is bepalend voor het effect. Uit de analyses volgt dat de *duur en omvang van de interventie* van belang is. Korte intensieve interventies zijn waarschijnlijk succesvoller, omdat meer controle mogelijk is ten aanzien van de implementatie van het DLVS, maar de vraag is of de effecten blijvend zijn over de periode na de interventie. In de interventie zou daarnaast minimaal aandacht moeten zijn voor *hoe leraren de ontvangen feedback kunnen vertalen naar de instructie.*
- Het succes van een DLVS hangt waarschijnlijk af van *een juiste combinatie van implementatiebevorderende factoren.* Als de doelstelling van de interventie het verhogen van leerresultaten op schoolniveau is, dan zal het bijvoorbeeld belangrijker zijn dat schoolleiders en bestuurders betrokken worden bij de interventie. Wanneer het het doel is om individuele leerresultaten te verhogen in de klassen van geselecteerde leraren, dan is dit waarschijnlijk minder noodzakelijk.
- Het belang van de factoren 'benchmarking' en 'groepsgemiddelden' zal waarschijnlijk afhangen van de doelstelling, doelgroep en de focus van de interventie. Deze factoren zullen minder van belang zijn in interventies die plaatsvinden binnen het speciaal onderwijs waar de verschillen tussen leerlingen binnen de groep veel groter zullen zijn, en maar in beperkte mate kunnen worden vergeleken met landelijke gemiddelden. Deze factoren zullen meer van belang zijn wanneer het doel het verhogen van leerresultaten op school- en bestuursniveau is.

- DLVS-en die gebruikt worden in combinatie met het vakgebied rekenen resulteren in vergelijking met de andere vakgebieden in een groter effect op leerresultaten. De gevonden verschillen tussen de vakgebieden waren echter niet significant, bovendien zou het beeld kunnen veranderen wanneer er meer studies in de analyses waren opgenomen waarin de effecten van lezen, spelling of science zijn onderzocht.

LITERATUURLIJST

*Allinder, R. M., Bolling, R. M., Oats, R. G., & Gagnon, W. A. (2000). Effects of teacher self-monitoring on implementation of curriculum-based measurement and mathematics computation achievement of students with disabilities. *Remedial and Special Education, 21*(4), 219-226.

Berkemeyer, N., & Van Holt, N. (2012). Leistungsruckmeldungen im Langsschnitt. Erste Erfahrungen mit dem Schuler-Monitoring-System (SMS). *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung*. (pp. 109-130): Wiesbaden: Springer VS.

Black, P., & Wiliam, D. (1998a). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. doi: 10.1080/0969595980050102

Black, P., & Wiliam, D. (1998b). Inside the Black Box: Raising Standards Through ClassroomAssessment. *Phi Delta Kappan, 80*(2), 139-144.

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to Learn from Data: Benchmarks and Instructional Communities. *Peabody Journal of Education, 85*(2), 205-225.

Bulkley, K. E., Oláh, L. N., & Blanc, S. (2010). Introduction to the Special Issue on Benchmarks for Success? Interim Assessments as a Strategy for Educational Improvement. *Peabody Journal of Education, 85*(2), 115-124. doi: 10.1080/01619561003673920

Burns, M. K., Klingbeil, D. A., & Ysseldyke, J. (2010). The Effects of Technology-Enhanced Formative Evaluation on Student Performance on State Accountability Math Tests. *Psychology in the Schools, 47*(6), 582-591.

*Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. Educational Evaluation and Policy Analysis, 33(3), 378-398.

Chojnacki, G., Eno, P., Liu, F., Meyers, C., Konstantopoulos, S., Miller, S., & van der Ploeg, A. (2013). *Do Interim Assessments Influence Instructional Practice in Year One? Evidence from Indiana Elementary School Teachers*. Abstract presented at the SREE Fall 2013 Conference, Washington

*Cordray, D., Pion, G., Brandt, C., & Molefe, A. (2012). *The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement*. (NCEE 2013-4000). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Faber, M., Van Geel, M., & Visscher, A. (2013). *Digitale Leerlingvolgsystemen als basis voor Opbrengstgericht werken in het Primair Onderwijs: een analyse van de wijze waarop scholen en besturen de mogelijkheden van digitale leerlingvolgsystemen kunnen benutten*. Enschede: Universiteit Twente Opgehaald 2 september, 2013 van

http://www.kennisnet.nl/fileadmin/contentelementen/kennisnet/Passend_Onderwijs/Kennisnetonderzoeksanalyse_LVS.pdf

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional children*. 53, pp. 199-208.

*Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review,* 19(1), 6-22

Fuchs, L. S., Hamlett, D. F. C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations.*American educational research journal*, *28*(3), 617-641.

*Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Allinder, R. M. (1991a). The contribution of skills analysis to curriculum-based measurement in spelling. *Exceptional Children*. 57(7), 443-452.

*Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Allinder, R. M. (1991b). Effects of expert system advice within curriculum-based measurement on teacher planning and student achievement in spelling. *School Psychology Review*. 20(1), 49-66.

*Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991c). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American educational research journal*, 28(3), 617-641.

*Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement, using a reading maze task. *Exceptional Children*. 58(5), 436-450

*Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Phillips, N. B. (1994). Classwide curriculum-based measurement: Helping general educators meet the challenge of student diversity. *Exceptional Children*. 60(6), 518-537

Ginsburg, H. P., Cannon, J., Eisenband, J., & Pappas, S. (2006). Mathematical thinking and learning. *Blackwell handbook of early childhood development*, 208-229.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81-112.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of educational statistics, 6*(2), 107–128.

Hellrung, K., & Hartig, J. (2013). Understanding and using feedback – A review of empirical studies concerning feedback from external evaluations to teachers. *Educational Research Review,9*(0), 174-190. doi: http://dx.doi.org/10.1016/j.edurev.2012.09.001

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement.* Issues & Answers Report REL(039).

Heritage, M., & Yeagley, R. (2005). Data Use and School Improvement: Challenges and Prospects. *Yearbook of the National Society for the Study of Education, 104*(2), 320-339.

Inspectie van het Onderwijs. (2012). *Toezichtkader po/vo 2012.* Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs. (2013). *De staat van het onderwijs. Onderwijsverslag 2011/2012.* Utrecht: Inspectie van het Onderwijs.

Kingston, N., & Nash, B. (2011). Formative Assessment: A Meta-Analysis and a Call for Research. *Educational Measurement: Issues and Practice, 30*(4), 28-37.

Kelly, A., Downey, C., & Rietdijk, W. (2010). *Data dictatorship and data democracy: understanding professional attitudes to the use of pupil performance data in schools*. Reading: CfBT Education Trust

Keuning, T., & Van Geel, M. J. M. (2012). *Focus projects II and III. The effects of a training in 'achievement oriented work' for primary school teams*. Poster presented at the International ICO fall school, Girona, Spain.

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. doi: 10.1037/0033-2909.119.2.254

Koedinger, K. R., McLaughlin, E. A., & Heffernan, N. T. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational ComputingResearch, 43*(4), 489-510.

*Konstantopoulos, S., Miller, S. R., & van der Ploeg, A. (2013). The Impact of Indiana's System of Interim Assessments on Mathematics and Reading Achievement. *Educational Evaluation and Policy Analysis*, 35(4), 481-499.

Kreft, I. G., & de Leeuw, J. (2002). *Introducing multilevel modelling*. London Thousand Oaks New Delhi: Sage Publications.

Locke, E.A., & G. Latham (2002). Building a Practically Useful Theory of Goal Setting and Task Motivation. *The American Psychologist*, *57*(9), 705-17.

Lipsey, M. W., & Wilson, D. (2000). *Practical meta-analysis (applied social research methods).*London Thousand Oaks New Delhi: Sage Publications.

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms.* (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

*May, H., & Robinson, M. A. (2007). *A randomized evaluation of Ohio's Personalized Assessment Reporting System (PARS)*. Philadelphia: Consortium for Policy Research in Education

McCaffrey, D. F., & Hamilton, L. S. (2007). *Value-Added Assessment in Practice: Lessons from the Pennsylvania Value-Added Assessment System Pilot Program* (Vol. 506): Rand Corporation.

Muralidharan, K., & Sundararaman, V. (2010). The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India. *The Economic Journal, 120*(546), 187-203.

Nabrs Olah, L., Lawrence, N. R., & Riggan, M. (2010). Learning to Learn from Benchmark Assessment Data: How Teachers Analyse Results. *Peabody Journal of Education, 85*(2), 226-245.

*Nunnery, J. A., & Ross, S. M. (2007). The effects of the School Renaissance program on student achievement in reading and mathematics. *Research in the Schools*, 14(1), 40-59.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments. *Educational Measurement: Issues and Practice, 28*(3), 5-13.

Quint, J., Sepanik, S., & Smith, J., (2008). *Using student data to improve teaching and learning:Findings from an evaluation of the Formative Assessments of Student Thinking in Reading (FAST-R) Program in Boston Elementary Schools.* New York: MDRC.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioural Science, 28*(1), 4-13. doi: 10.1002/bs.3830280103

Rennie Center for Education Research & Policy. (2006). *Data-Driven Teaching: Tools and Trends.* Cambridge: Rennie Center for Education Research & Police.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* Thousand Oaks: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Andover*:* Cengage Learning.

Shaw, S., & Wayman, J. C. (2012). *Third-Year Results From an Efficacy Study of the Acuity DataSystem.* Austin, TX: The University of Texas.

Shepard, L. A. (2010). What the Marketplace Has Brought Us: Item-by-Item Teaching with Little Instructional Insight. *Peabody Journal of Education, 85*(2), 246-257.

Slavin, R. E. (2008a). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5-14.

Slavin, R. E., & Smith, D. (2008b). *Effects of Sample Size on Effect Size in Systematic Reviews in Education*. Paper presented at the Society for Research on Effective Education, Virginia.

*Slavin, R. E., Cheung, A., Holmes, G., Madden, N. A., & Chamberlain, A. (2013). Effects of a Data-Driven District Reform Model on State Assessment Outcomes. *American Educational Research Journal*, 50(2), 371-396.

Spicuzza, R., Ysseldyke, J., Lemkuil, A., Kosciolek, S., Boys, C., & Teelucksingh, E. (2001). Effects of Curriculum-Based Monitoring on Classroom Instruction and Math Achievement. *Journal of School Psychology, 39*(6-), 521-542.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (submitted). *Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis.* Manuscript submitted for publication.

Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2010). Using School Performance Feedback: Perceptions of Primary School Principals. *School Effectiveness and School Improvement, 21*(2), 167-188.

Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2011). Effecten van ondersteuning bij schoolfeedbackgebruik. *Pedagogische Studiën, 88*(2), 90-106.

Visscher, A., & Coe, R. (2002). *School improvement through performance feedback*: Swets & Zeitlinger.

Visscher, A. J., & Coe, R. (2003). School Performance Feedback Systems: Conceptualization, Analysis, and Reflection. *School Effectiveness and School Improvement, 14*(3), 321-349.

*Wang, Y., & Gushta, M. (2013). *Improving student outcome with mClass: Math, a technology-enhanced CBM and Diagnostic Interview Assessment*. Abstract presented at the SREE Fall 2013 Conference, Washington

Wayman, J. C., Stringfield, S., & Yakimowski, M. (2004). *Software enabling school improvementthrough analysis of student data.* Johns Hopkins University and Baltimore City Public School System: Report No. 67

Wayman, J. C. (2007). *Student data systems for school improvement: The state of the field.* Paper presented at the TCEA educational technology research symposium.

Wayman, J. C., Cho, V., & Shaw, S. (2009). *First-year results from an efficacy study of the Acuity data system.* Austin, TX: The University of Texas.

Wayman, J. C., Shaw, S. M., & Cho, V. (2011). *Second-year results from an efficacy study of the Acuity data system.* Austin: Authors.

Wayman, J. C., Cho, V., Jimerson, J. B., & Spikes, D. D. (2012). District-Wide Effects on Data Use in the Classroom. *Education Policy Analysis Archives, 20*(25).

Williams, D., & Coles, L. (2007). Teachers' Approaches to Finding and Using Research Evidence: An Information Literacy Perspective. *Educational Research, 49*(2), 185-206.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, (37), 3-14.

Williams, R. T., Swanlund, A., Miller, S., Konstantopoulos, S., Eno, J., van der Ploeg, A., & Meyers, C. (in press). Measuring Instructional Differentiation in a Large Scale Experiment

Ysseldyke, J., Spicuzza, R., Kosciolek, S., Teelucksingh, E., Boys, C., & Lemkuil, A. (2003). Using a curriculum-based instructional management system to enhance math achievement in urban schools. *Journal of Education for Students Placed at Risk, 8*(2), 247-265.

*Ysseldyke, J., & Bolt, D. M. (2007). Effect of technology-enhanced continuous progress monitoring on math achievement. *School Psychology Review*, 36(3), 453.

BIJLAGE 1: GESELECTEERDE STUDIES

	Auteurs	*Titel*	*Jaar*	*Onderzoek*	*Type publicatie*	*Vakgebied*	***Populatie*** type onderwijs leeftijd leerlingen	***Steekproef*** leerlingen, leraren, scholen, besturen	*Toets*
1.	Allinder, Bolling, Oats en Gagnon	Effects of teacher self-monitoring implementation of curriculum-based measurement and mathematics computation achievement of students with disabilities	2000	experiment	artikel	rekenen	speciaal 9 -10 jaar	54 30 - -	eigen
2.	Carlson, Borgman en Robinson	A Multistate District-Level Cluster Randomized Trial of the Impact of Data-Driven Reform on Reading and Mathematics Achievement	2011	experiment	artikel	rekenen lezen	regulier (lage leerresultaten) 8 tot 14 jaar	≈276148 - 514/524 57/59	state
3.	Cordray, Pion, Brandt en Molefe	The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement	2012	experiment	rapport	lezen	regulier 9 – 11 jaar	3720 172 32 5	state en eigen
4.	Fuchs, Fuchs, Hamlett en Stecker	The role of skills analysis in curriculum-based measurement in math	1990	experiment	artikel	rekenen	speciaal 8 - 15 jaar	91 30 16 -	eigen
5.	Fuchs, Fuchs, Hamlett en Allinder	The contribution of skills analysis to curriculum-based measurement in spelling	1991a	experiment	artikel	spelling	speciaal 8-15 jaar	92 30 16 -	eigen
6.	Fuchs, Fuchs, Hamlett en Allinder	Effects of expert system advice within curriculum-based measurement on teacher planning and student achievement in spelling.	1991b	experiment	artikel	spelling	speciaal 7 – 14 jaar	59 30 16 -	eigen
7.	Fuchs, Fuchs, Hamlett en Stecker	Effects of Curriculum-Based Measurement and Consultation on Teacher Planning and Student Achievement in Mathematics Operations	1991c	experiment	artikel	rekenen	regulier leerlingen lage resultaten 7-14 jaar	63 33 15 -	eigen

Auteurs	*Titel*	*Jaar*	*Onderzoek*	*Type publicatie*	*Vakgebied*	*Populatie* type onderwijs leeftijd leerlingen	*Steekproef* leerlingen, leraren, scholen, besturen	*Toets*
8. Fuchs, Fuchs, Hamlett en Ferguson	Effects of expert system consultation within curriculum-based measurement, using a reading maze task	1992	experiment	artikel	lezen	speciaal 6-15 jaar	63 33 15 -	eigen
9. Fuchs, Fuchs, Hamlett, en Phillips	Class wide Curriculum-Based Measurement: Helping General Educators Meet the Challenge of Student Diversity	1994	experiment	artikel	rekenen	regulier (leerlingen met lage tot gemid-delde resultaten) 6-11 jaar	120 40 11 1	eigen
10. Konstantopoulo, Miller en van der Ploeg	The Impact of Indiana's System of Interim Assessments on Mathematics and Reading Achievement	2013	experiment	artikel	rekenen lezen	regulier 5-8 jaar 8-14 jaar	≈20.000 - 59 -	state
11. May en Robinson	A Randomized Evaluation of Ohio's Personalized Assessment Reporting System (PARS)	2007	experiment	rapport	lezen spelling rekenen	regulier 15 -16 jaar	51.580 - 100 60	state
12. Nunnery en Ross	The effects of the School Renaissance program on student achievement in reading and mathematics	2007	matching	artikel	lezen rekenen	regulier 8-14 jaar	≈917 - 18/4 -	state
13. Slavin, Cheung, Holmes, Madden en Chamberlain	Effects of a Data-Driven District Reform Model on State Assessment Outcomes	2013	experiment	artikel	lezen rekenen	regulier 8-17 jaar	≈300.000 - 608 59	state
14. Wang en Gushta	Improving student outcome with mClass: Math, a technology-enhanced CBM and Diagnostic Interview Assessment	2013	matching	niet gepubliceerd	rekenen	regulier 7-8 jaar	41.363 1856 606 175	state
15. Ysseldyke en Bolt	Effect of Technology-Enhanced Continuous Progress Monitoring on Math Achievement	2007	experiment	artikel	rekenen	regulier 5-14 jaar	1880 80 8 7	eigen

BIJLAGE 2: BESCHRIJVING PER STUDIE

1. Allinder et al., (2000)

In de studie van Allinder wordt het effect van *Curriculum-Based Measurement* (CBM) onderzocht. CBM is een methode waarbij korte, formatieve toetsen worden afgenomen om de voortgang van leerlingen te registeren, zodat de ontwikkeling daarin in beeld gebracht wordt. Deze methode gebruiken leraren daarnaast ook om de effectiviteit van hun instructie te evalueren.

In het onderzoek van Allinder maken leerlingen twee keer per week een korte toets die tussen de één en vijf minuten duurt. Leraren evalueren twee keer per maand de toetsresultaten met behulp van een DLVS. Het DLVS maakt de ontwikkeling van leerlingen zichtbaar, doordat de scores van de verschillende toetsmomenten in een grafiek gezet worden. Tijdens het evalueren van de toetsresultaten moesten leraren een vaste set van *decision rules* gebruiken, namelijk de volgende*:*

- wanneer de ontwikkeling van leerlingen te langzaam verloopt om de leerdoelen te behalen, dan passen leraren hun instructie aan,
- wanneer de ontwikkeling sneller dan verwacht verloopt passen leraren de leerdoelen aan,
- en wanneer de ontwikkeling van een leerling naar verwachting verloopt vervolgen leraren de eerder toegepaste instructie, en lesmethoden .

Voor het aanpassen van de instructie kunnen leraren gebruik maken van een *skills analysis*. Met deze functie van het DLVS wordt het beheersingspercentage aangegeven op de verschillende leerstofcategorieën.

De interventie in deze studie bestond uit één trainingsbijeenkomst van drie uur. In deze bijeenkomst werden de mogelijkheden van het DLVS gepresenteerd, en konden leraren oefenen met het DLVS. Elke leraar had daarnaast ongeveer twee keer per maand contact met een onderzoeksassistent. Onderzoeksassistenten waren afgestudeerde studenten onderwijspsychologie die ervaring hadden opgedaan met CBM. Zij ondersteunden leraren bij het interpreteren van de feedback, en lieten ze werken met de verschillende mogelijkheden binnen het systeem. Acht leraren ontvingen daarnaast een extra interventie waarin ze een monitoringssysteem leerden toepassen. Met behulp van dit systeem maakten leraren gestructureerd gebruik van de *skills analysis* functie.

2. Carlson et al., (2011)

In de studie van Carlson worden de effecten van een *data-driven reform* onderzocht. Kort gezegd houdt dit in dat scholen en schoolbesturen getraind worden in hoe zij verschillende databronnen kunnen inzetten voor het signaleren van risico's die betrekking hebben op de kwaliteit van het onderwijs, en voor het vormgeven aan onderwijsbeleid.

In de studie speelden schoolbegeleiders of consulenten een belangrijke rol. Zij maakten voor de scholen en besturen een feedbackrapportage en ze begeleiden maandelijks een schoolactieteam''. Voor het opstellen van de feedbackrapporten gebruikten de consulenten verschillende bronnen, zoals de resultaten op benchmarktoetsen, de resultaten op *state toetsen*, doorstroomgegevens van leerlingen, gegevens over verwijzingen naar het speciaal onderwijs, dan wel gegevens uit lerarenenquêtes. Het was de bedoeling dat de scholen vier, of vijf keer per jaar een benchmarktoets zouden afnemen. In het eerste onderzoekscohort nam echter maar 70% van de scholen de toetsen twee keer in het jaar af. In het tweede en derde onderzoekscohort steeg dit percentage, toen nam 90% van de scholen de toetsen

drie, of vier keer af in het schooljaar. De scholen konden een DLVS gebruiken om de resultaten van de benchmarktoetsen te analyseren. Met dit DLVS konden scholen de behaalde resultaten afzetten tegen *state* standaarden, en individuele leerresultaten aggregeren naar leeftijdsgroepen, cohorten of overige subgroepen.

Maandelijks kwamen de schoolactieteams bijeen onder begeleiding van een consulent. In deze teams zaten bestuurders, schoolleiders en leraren met leidinggevende taken. In de teams werd het feedbackrapport besproken en overlegd hoe de feedback vertaald kon worden naar interventies in de onderwijspraktijk.

In de studie zijn alleen de eerste componenten van een driejarige interventie onderzocht. De onderzochte componenten waren: het systematisch afnemen van benchmarktoetsen, het evalueren van verschillende databronnen, en het trainen van de schoolactieteams. De effecten van de twee daarop volgende componenten, het beschikbaar maken van onderzoeksrapporten over de effecten van schoolprogramma's, en het begeleiden van scholen bij de selectie en implementatie van schoolprogramma's werden niet onderzocht in deze studie.

3. Cordray et al., (2012)

Cordray en collega's onderzochten de effecten van het programma *Measures of Academic Progress* (MAP). Dit programma bestaat uit twee onderdelen: het systematisch afnemen van computerondersteunde adaptieve toetsen, en een interventie die bestaat uit een vierdaagse training en schoolconsultaties.

Drie keer per jaar maakten leerlingen de MAP toetsen. De resultaten van deze toetsen konden op een schaal geplaatst worden waardoor de ontwikkeling over meerdere leerjaren in beeld werd gebracht. De resultaten van deze toetsen waren gestandaardiseerd, zodat het verschil tussen twee scores steeds dezelfde betekenis had, ongeacht of de resultaten zich aan het begin of eind van de schaal bevonden. De resultaten op de toetsen gaven tevens een goede indicatie van de resultaten op de *state toetsen*. Na de toetsafname kregen leerlingen meteen hun scores. Leraren hadden 24 uur na de afname toegang tot de resultaten. In het DLVS konden zij rapporten opvragen. Leraren kregen daarin feedback over de studievoortgang en de beheersing van specifieke leerdomeinen.

De interventie bestond uit een training van vier dagen. Deze dagen waren verspreid over het schooljaar. Het onderwerp dat centraal stond in de eerste dag waren de MAP toetsen, op de tweede dag stond het gebruiken van feedback centraal, en tijdens de derde en vierde dag stonden respectievelijk het differentiëren in de instructie en het beoordelen van de leerontwikkeling centraal. In de training leerden leraren bijvoorbeeld hoe zij subgroepen van leerlingen met dezelfde leerbehoeften moesten vormen, of hoe zij verwerkingsmethodes en leermaterialen konden selecteren die aansloten bij de leerbehoeften. Naast deze training hadden leraren ook toegang tot een website waar zij informatie konden opvragen over bijvoorbeeld instructiematerialen die aansloten bij specifieke leerbehoeften, of over het formuleren van leerdoelen.

Naast de training hadden scholen de mogelijkheid om vier schoolconsultaties aan te vragen. De aanvraag voor een consultatie vond vaak op initiatief van de schoolleider plaats. Deze consultaties duurden ongeveer 1 tot 2 uur en waren gericht op de specifieke behoeften van de school. In het eerste jaar hadden de consulenten 90% van de scholen tenminste één keer bezocht.

4. Fuchs et al., (1990)

In de studie van Fuchs worden de effecten van CBM onderzocht. Het DLVS, de toetsen en de interventie komen daardoor deels overeen met die in de studie van Allinder.

In de studie van Fuchs moesten leraren gedurende een periode van vijftien weken twee keer per week een korte toets afnemen bij leerlingen. In deze toetsen werden steeds dezelfde leerstofcategorieën getoetst. Deze leerstofcategorieën zou een leerling aan het einde van het schooljaar moeten kunnen beheersen, dit houdt dus in dat een leerling aan het begin van het schooljaar nog veel fouten mocht maken. Uit deze fouten kon de leraar vervolgens afleiden voor welke leerstofcategorieën meer instructie, of een andere instructie nodig was.

Met een DLVS werden de scores van de leerlingen in een grafiek gezet. Door de scores werd een ontwikkelingslijn en een doellijn getrokken. De leraar kon hieruit afleiden of een leerling de opgestelde leerdoelen aan het eind van het jaar wel of niet zou gaan behalen. Het DLVS gaf daarnaast een aantal *decision rules* weer. Nadat er acht toetsen waren afgenomen gaf het systeem aan of het nodig was om de leerdoelen aan te passen, of dat het nodig was om de instructie aan te passen. Welke *decision rule* aangegeven werd was afhankelijk van de ontwikkelings- en doellijn. Ook de leraren in deze studie konden weer gebruik maken van een *skills analysis*.

De interventie van deze studie was verspreid over een periode van acht weken. De interventie bestond uit twee workshops van twee uur, en lesobservaties. Onderzoekers bezochten elke drie weken de leraren om samen de voortgang van de implementatie van CBM te bespreken.

5. Fuchs et al., (1991a)

Deze studie van Fuchs komt voor een groot gedeelte overeen met de hiervoor beschreven studie. Alleen worden in deze studie niet de effecten op de rekenresultaten maar op spelling onderzocht.

Ook in deze studie moesten leraren gedurende een periode van vijftien weken tenminste twee keer per week een korte toets afnemen. Aan het begin van de periode bepaalden leraren het spellingsniveau van een leerling. Dit niveau bepaalde vervolgens welke categorieën woorden de leerling aan het eind van het schooljaar moest beheersen. Uit die categorieën werden woorden geselecteerd die gebruikt waren voor het samenstellen van de toetsen. De toetsen werden digitaal afgenomen en de resultaten werden automatisch in een grafiek gezet. In deze grafiek was de doellijn zichtbaar (de ontwikkeling waarnaar gestreefd werd), en een lijn *of best fit*, die de werkelijke ontwikkeling weergaf. Het DLVS gaf de *decision rules* weer, en bovendien was een *skills analysis* functie beschikbaar.

De interventie kwam ook grotendeels overeen met de interventie uit de voortgaande studie van Fuchs. Een toevoeging was dat er tijdens de individuele contactmomenten ook instructieadvies gegeven werd. Door de onderzoekers waren 27 instructiesuggesties ontwikkeld die aansloten bij de 27 soorten spellingfouten die met de *skills analysis* functie konden worden bepaald.

6. Fuchs et al., (1991b)

In deze studie worden de effecten van CBM in combinatie met een *expert systeem* onderzocht.

Gedurende een periode van 18 weken namen leraren ten minste twee keer per week een korte spellingstoets af. Het DLVS gaf weer automatisch de ontwikkeling van leerlingen weer in een grafiek, door middel van een doellijn, en een lijn of *best fit*. Het systeem gaf de *decision rules* en voerde automatisch een *skills analysis* uit.

Leraren moesten wekelijks de resultaten uit de analyses raadplegen. Wanneer leraren de *decision rule*: "Oh-oh. Make a teaching change" op hun scherm zagen moesten zij ook het *expert system* raadplegen. Voordat leraren vanuit het *expert system* advies ontvingen over de instructie moesten zij eerst een aantal gegevens invoeren. Leraren moesten gegevens over de ontwikkeling van de leerling, gegevens uit de *skills analysis*, gegevens over de gegeven instructie, en gegevens over het presteren van de leerling op andere opdrachten invoeren. Het systeem gaf vervolgens advies over de benodigde instructie en verwerkingsopdrachten.

De interventie bestond uit twee workshops van elk twee uur en individuele contactmomenten. Eén keer per twee weken bezochten onderzoekmedewerkers de leraren om te bewaken dat zij de interventie goed implementeerden. Gedurende het onderzoek had elke leraar gemiddeld tien keer contact met een projectmedewerker.

7. Fuchs et al., (1991c)

Net zoals in de voorgaande studie worden ook hier de effecten van CBM in combinatie met een *expert systeem* onderzocht. Alleen werden in deze studie de effecten op rekenresultaten onderzocht in plaats van op spelling.

In een periode van twintig weken namen leraren ten minste twee keer per week korte toetsen af. Net zoals in de voorgaande studie gebruikten leraren daarvoor een DLVS dat:

- automatisch de behaalde toetsresultaten in een grafiek zette,
- in deze grafiek een doellijn en ontwikkelingslijn weergaf,
- de *decision rules* aangaf,
- een *skills analysis* kon uitvoeren,
- en tot slot een *expert system* bevatte.

De interventie bestond weer uit twee workshops en individuele contactmomenten. Eén keer per twee weken bezochten onderzoekmedewerkers de leraren. Gedurende het onderzoek had elke leraar gemiddeld tien keer contact met een medewerker van het project.

8. Fuchs et al., (1992)

Fuchs en collega's onderzochten ook in deze studie de effecten van CBM in combinatie met een *expert systeem.* In deze studie werden de effecten op lezen onderzocht.

Gedurende een periode van 17 weken namen leraren tenminste twee keer per week korte leestoetsen af. Net zoals in de voorgaande studie gebruikten leraren daarvoor een DLVS dat de ontwikkeling van leerlingen grafisch in beeld bracht, een ontwikkelingslijn, doellijn en *decision rules* aangaf, en een *expert systeem* bevatte. In deze studie hadden leraren dus geen beschikking over een *skills analysis*. De interventie kwam wel overeen met die uit de eerdere studies van Fuchs, et al.

9. Fuchs et al., (1994)

Ten opzichte van de voorgaande Fuchs-studies wijkt deze studie enigszins af. Toetsen werden in deze studie klassikaal afgenomen in plaats van individueel, en de studie vond niet plaats in het speciaal onderwijs, maar in het regulier onderwijs.

Gedurende een periode van 25 weken namen leraren elke week een korte toets af. Op basis van de toetsresultaten ontvingen leraren twee keer per maand een feedbackrapport. Dit rapport bestond uit de volgende onderdelen:

- een grafiek waarin de ontwikkeling van elke individuele leerling werd weergegeven,
- een *skills profiel* waarin aangegeven werd of een leerling een specifieke *skill*:
  - nog niet geoefend had,
  - wel geoefend had, maar nog niet beheerste,
  - wel geoefend had en deels beheerste,
  - grotendeels beheerste,
  - volledig beheerste.
- een grafiek waarin de ontwikkeling van de gehele groep weergegeven werd,
- een *skills profiel* voor de gehele groep,
- en, indien de leraar geplaatst was binnen de experimentele conditie met instructieadvies, dan gaf het DLVS ook:
  - instructieadvies,
  - en advies over de samenstelling van instructiegroepjes.

De interventie bestond uit één gezamenlijke trainingsbijeenkomst en wekelijkse individuele begeleiding van 10 tot 15 minuten door een projectmedewerker. Tijdens deze begeleidingsmomenten ontvingen leraren de feedbackrapportages en werden ze op basis van lesobservaties begeleid.

10. Konstantopoulous et al., (2013)

In deze studie zijn de effecten van twee online assessmentsystemen onderzocht. De systemen bestaan uit periodieke diagnostische toetsen die door leerlingen in de state Indiana min of meer op hetzelfde moment gemaakt worden. Er is een systeem voor leerlingen van zeven tot acht jaar (*mClass*), en een systeem voor leerlingen van acht tot veertien jaar (*Acuity*). Met beide systemen kunnen leraren direct na de toetsafnames feedbackrapportages inzien en toetsresultaten analyseren.

In het artikel wordt niet aangegeven hoe vaak, en wanneer leraren de toetsen van *mClass* afnamen. De auteurs benoemen wel de volgende mogelijkheden die het bijbehorende DLVS biedt:

- het signaleren van leerproblemen,
- het onderzoeken van de mogelijke oorzaken voor leerproblemen,
- het in beeld brengen van de leerstrategieën van leerlingen,
- het in beeld brengen van de ontwikkeling van leerlingen,
- individuele leerresultaten genereren naar groepsniveau,
- en het geven van toegang tot instructiemateriaal.

Acuity bestaat uit twee typen toetsen. Namelijk uit drie voorspellende toetsen die hoofdzakelijk ingezet worden om de resultaten op *state toetsen* te voorspellen, en uit vier diagnostische toetsen die ingezet worden om leerbehoeften te identificeren. De toetsen bestaan uit 30 tot 35 digitale multiple choice vragen die klassikaal worden afgenomen. Het DLVS van Acuity biedt leraren de volgende mogelijkheden:

- benchmarks waarmee leraren de individuele- en de groepsresultaten af kunnen zetten tegen *state standaarden*,
- het kunnen uitvoeren van analyses op itemniveau (specifieke leerbehoeften),
- de mogelijkheid om zelf uit items aanvullende toetsen samen te stellen waarmee leerlingen extra moeten oefenen,
- toegang tot verschillende instructie- en lesmaterialen.

In de studie is een *train-the-trainer model* gebruikt. Deze interventie werd gegeven door de ontwikkelaars van de assessmentsystemen. Het *train-the-trainer model* hield in dat er per school één

tot vier leraren getraind werden. De training bestond uit twee bijeenkomsten in de zomerperiode, en twee bijeenkomsten in de herfstperiode. Vervolgens werd van de getrainde leraren verwacht dat zij twee, tot drie trainingen zouden verzorgen binnen hun eigen school. In de publicatie wordt geen informatie gegeven over de mate waarin leraren daadwerkelijk gebruik maakten van *mClass* of *Acuity*, noch over de mate waarin de interventie binnen scholen doorgegeven was aan collega's.

11. May et al., (2007)

In de studie van May werden de eerstejaars effecten van het *Personalized Assessment Reporting System (PARS)* onderzocht. Het doel van PARS is om leerlingen, ouders, scholen, en schoolbestuurders feedback te geven op basis van de prestaties van leerlingen op *state toetsen*. In deze studie konden leraren, of scholen de resultaten van de huidige groep 10 (15-16 jarigen) alleen gebruiken om het onderwijs van de toekomstige groep 10 te verbeteren.

De feedback werd gegeven in vier (papieren) *grow* rapporten en er was een online DLVS beschikbaar. Met het DLVS konden scholen toetsresultaten analyseren, bovendien werd informatie geboden over verschillende instructiemogelijkheden, en het systeem bevatte professionele ontwikkelingstools. Met de analysetools konden leraren onder andere de ontwikkeling van leerlingen over een langere periode in beeld brengen, en een overzicht maken waarin het percentage leerlingen weergegeven werd, dat een specifieke vraag van de *state toets* juist, of onjuist beantwoord had.

Scholen ontvingen vier verschillende *grow* rapporten per jaar. In het eerste rapport werden per vakgebied de behaalde toetsresultaten weergegeven. Deze werden vergeleken met de state standaarden en konden weergegeven worden per subgroep (bijvoorbeeld de subgroep 'leerlingen met een indicatie voor het speciaal onderwijs'). In het tweede rapport werden de toetsresultaten van leerlingen die de *state toets* niet hadden gehaald verder geanalyseerd. Het derde rapport bevatte een *student roster*: een tabel waarin per individuele leerling het beheersingsniveau van de vijf categorieën binnen de *state toets* aangegeven wordt. Dit niveau werd met vijf verschillende kleurcoderingen weergegeven. Het laatste rapport betrof een interventieplan.

Ook in deze studie werd een *train-the-trainer model* ingezet en werd de interventie ook uitgevoerd door de producenten van PARS. In de eerste bijeenkomst van een uur ontvingen de deelnemende leraren vooral technische informatie over de rapporten en de website, in de tweede bijeenkomst konden leraren twee uur oefenen met de analysetools van het DLVS.

12. Nunnery et al., (2007)

In de studie van Nunnery werden de effecten van *Accelerated Reader* (AR) en *Accelerated Math* (AM) onderzocht. Dit betreft systemen die digitale toetsen bevatten, en op basis van de toetsresultaten digitale feedbackrapportages ontwikkelen. Leraren konden met behulp van beide systemen dagelijks korte formatieve toetsen afnemen bij leerlingen.

AR bevatte toetsen waarmee de woordenschat van leerlingen in beeld gebracht werd. Nadat leerlingen een boek hadden gelezen maakten ze een digitale toets waarin vragen gesteld werden over het begrip van woorden die in het boek aan de orde waren gekomen. Het DLVS genereerde vervolgens automatisch een rapport over de woordenschatontwikkeling van elke leerling. De leraar kon hiermee de leesontwikkeling van leerlingen monitoren en ervoor zorgen dat leerlingen boeken lazen die aansloten bij hun niveau.

AM toetsen bestonden uit multiple choice vragen. Deze toetsen gaven leerlingen de gelegenheid om frequent te oefenen met praktische problemen. Leraren ontvingen feedbackrapportages met daarin diagnostische informatie over het niveau van elke leerling. Leraren kregen informatie over de algehele beheersing van rekenen, en informatie over de beheersing van specifieke rekenonderdelen. Daarnaast gaf AM voor specifieke leerstofcategorieën passende instructiemogelijkheden en verwerkingsopdrachten aan.

De producenten van beide systemen boden scholen een cursus en coaching aan. Deze coaching was naast technische ondersteuning ook gericht op de vertaling van feedback naar de instructie. In de studie wordt geen informatie gegeven over de precieze omvang van deze interventie.

13. Slavin et al., (2013)

In de studie van Slavin werden de effecten van dezelfde interventie onderzocht als die in de studie van Carlson. In beide studies werd het effect van een *district-level reform* onderzocht, een interventie die ontwikkeld is door het *Center for Data-Driven Reform in Education* (CDDRE). Consulenten van CDDRE voerden de interventie uit en bezochten daarvoor 30 keer de geselecteerde schoolbesturen. In de studie van Slavin was de interventie volledig uitgevoerd terwijl in de studie van Carlson alleen de eerste drie componenten uitgevoerd waren. De belangrijkste toevoeging in de studie van Slavin is het implementeren van schoolprogramma's waarvan de effecten op leerresultaten wetenschappelijk bewezen zijn.

In het artikel van Slavin wordt beschreven dat de interventie uit vier verschillende onderdelen bestaat. Het eerste onderdeel betrof een *data review*. In bijeenkomsten met bestuurders en schoolteams werd feedback uit verschillende bronnen besproken, deze bijeenkomsten werden ondersteund door de consultants. Het tweede onderdeel van de interventie waren de benchmark assessments die vier keer per jaar werden afgenomen. De inhoud van deze toetsen sloot grotendeels aan bij de inhoud van de *state* toetsen. Het primaire doel van deze benchmarktoetsen was dan ook het voorspellen van resultaten op de *state* toets. Voor leraren en schoolleiders was een DLVS beschikbaar waarmee zij de individuele leerresultaten konden aggregeren naar het (sub)groepsniveau. Het derde onderdeel van de interventie bestond uit *school walk-through*. De consulenten bezochten samen met de bestuurders scholen om zich een beeld te vormen van de kwaliteit van het onderwijs. In het vierde onderdeel van de interventie begeleiden de consultenten de bestuurders en schoolleiders bij het kiezen, en implementeren van schoolprogramma's die aansloten bij de geïdentificeerde problemen. Slechts een derde van de scholen implementeerde een schoolprogramma gericht op lezen, en acht procent een verbeteringsprogramma gericht op rekenen.

14. Wang et al., (2013)

In de studie van Wang worden de effecten van *mClass*. De effecten van *mClass* werden ook onderzocht in de studie van Konstantopoulos.

In de studie van Wang wordt *mClass* beschreven als een CBM methode. Het systeem bestond uit korte digitale toetsen van ongeveer twee minuten. Met de toetsen werden drie rekenonderdelen getoetst, de gevoeligheid voor getallen, rekenen, en automatiseren. Elk onderdeel bestond weer uit een aantal verschillende leerdomeinen. Leerlingen moesten tijdens de toets niet alleen antwoorden geven, ze gaven daarnaast ook aan hoe zij tot hun antwoord waren gekomen. Daarvoor konden ze kiezen uit verschillende symbolen die in *mClass* werden weergegeven. Er was bijvoorbeeld een symbool voor optellen met chips en een symbool voor hoofdrekenen. Nadat de toetsen afgenomen waren ontvingen leraren rapporten met daarin feedback over de ontwikkeling van de leerling afgezet tegen *state* standaarden, feedback over de beheersing van een leerling op de verschillenden leerstofcategorieën, en feedback over de toegepaste leerstrategieën. Het systeem gaf op basis van deze feedback ook advies

over de instructie, over het formuleren van subgroepen, en over lesmaterialen. De toetsresultaten konden geaggregeerd worden naar het groepsniveau.

Naast de toetsen beschikt mClass ook over materiaal voor het afnemen van diagnostische interviews. Tijdens deze interviews van vijf tot tien minuten onderzocht de leraar de leer- en oplossingsstrategieën van een leerling. Ook de bekendheid met rekenkundige concepten werd onderzocht. De effecten van de interviews zijn niet opgenomen in de meta-analyse, omdat de toewijzing van de controle en experimentele groep niet voldeed aan het in dit onderzoek opgestelde criterium (zie hoofdstuk 3).

De producent van mClass leverde het systeem in combinatie met een support service. Deze service bestond uit begeleiding bij het implementeren van mClass en een training over het gebruiken van mClass in de onderwijspraktijk. Daarnaast konden scholen contact opnemen met mClass voor technische ondersteuning, of vragen over de toepassing van feedback in het onderwijs. De mate waarin de geselecteerde scholen gebruik maakten van deze services werd niet vermeld.

15. Ysseldyke et al., (2007)

Ysseldyke en collega's onderzochten de effecten van *Accelerated Math* (AM). De effecten van dit systeem werden ook in de studie van Nunnery onderzocht. In de studie van Ysseldyke wordt AM beschreven aan de hand van een aantal principes waarop het systeem gebaseerd is, namelijk het efficiënt besteden van de tijd aan het leren van essentiële vaardigheden, het afstemmen van de instructie op leerbehoeften, het frequent feedback verschaffen aan leerlingen en leraren, het gestructureerd werken aan de hand van leerdoelen, en, tot slot, het gebruiken van technologie.

Voordat leraren konden gaan werken met AM werd eerst een pre-test afgenomen, zodat het startniveau van elke leerling bekend was. In aansluiting op dit niveau genereerde AM een aantal opdrachten waarmee de leerling ging oefenen. Leerlingen maakten de opdrachten, en de resultaten daarvan werden gescand. Op basis van deze resultaten gaf AM meteen feedback, zowel aan de leerling als aan de leraar. Wanneer de resultaten voldoende waren gaf het systeem aan dat de leerling kon worden getoetst. Maakte de leerling de toets vervolgens ook voldoende, dan gaf AM een reeks opdrachten die aansloten bij het volgende leerstofdomein. Wanneer bleek dat een leerling een leerstofdomein na verschillende oefeningen onvoldoende beheerste, dan werd de leraar daarvan op de hoogte gesteld.

Het programma bevatte een reviewfunctie waarmee leraren toetsen, en opdrachten konden opvragen waarin leerlingen gelijktijdig met de al behandelde leerstofcategorieën kon oefenen. AM hield automatisch de vorderingen van leerlingen bij, zodat leraren zicht hadden op de leerontwikkeling. Leraren ontvingen daarnaast feedback waarmee zij vorm konden geven aan een afgestemde instructie, en waarmee zij verschillende instructiesubgroepen konden samenstellen.

Het systeem bood 197 verschillende leerstofcategorieën aan. De onderzoekers hebben in beeld gebracht hoeveel van de leerstofcategorieën aangeboden waren binnen de geselecteerde scholen. Ze konden een onderscheid maken tussen scholen die gemiddeld minder dan 9 leerstofcategorieën hadden aangeboden, scholen die gemiddeld tussen de 10 en 36 hadden aangeboden, en scholen die gemiddeld meer dan 36 leerstofcategorieën hadden aangeboden.

Gedurende het schooljaar ontvingen leraren drie tot vijf consultaties van een medewerker van de producent van AM. Deze consulent adviseerde leraren hoe zij hun gebruik van AM zo konden verbeteren, dat het systeem beter in het onderwijs geïntegreerd kon worden. Elke leraar had daarnaast toegang tot een helpdesk voor het stellen van technische vragen.

BIJLAGE 3: CONTACTPERSONEN

Inhoud e-mail

We have started a study on the effects of digital student monitoring systems like for example the Dutch Cito student monitoring system: digital systems for analyzing and reporting the results of (quality) student assessments (taken twice a year in primary education either on paper or by means of the computer). We study the use and impact of these systems from the perspective of **providing feedback to teachers:** supporting teachers in fine tuning their instruction to the (varying) needs of students that have been determined by means of the assessments (and other data). It is of course probable that the assessment results also will be used for feeding them back to their **students**.

I noticed that my description makes some people think of digital systems that students use for learning subject matter content on their own while the system analyses their progress (assessment for learning) and provides feedback based on that. That is not what I am looking for although it is also a very interesting and promising field in my view.

My study is about the evidence that we have from good studies (ideally RCT's or designs close to that) that feeding back student achievement data (benchmark assessments) to teachers improves student achievement because the feedback drives the instruction of those teachers.

We all suppose that feedback works but the evidence in this particular case is not that impressive. Searching for all the quality studies in this area around the world that are available may form a basis for drawing conclusions about what we know about this and what not, and how we should go further.

We are working on a review (if possible a quantitative one) of the effects of these digital monitoring systems on teaching quality and student performance. We ourselves are executing several interventions in 150 schools (3000 school staff) to promote the use of these systems in schools and to study the effects scientifically and therefore do this worldwide review now.

We systematically search for studies that meet our criteria but we expect that not that many studies are available.

We would like to ask you as an expert whether you know:
1. Studies by yourself and/or others that we could include in our review.
2. Other people we should contact for other studies/relevant information on this area of research.

Aantal contactpersonen per land

Land	***Aantal benaderde contactpersonen***
Australia	3
Austria	1
Belgium	2
Canada	4
Cyprus	1
Denmark	1
France	1
Finland	1
Germany	13
Israel	1
Italy	7
Japan	1
Mexico	1
New Zealand	5
Nederland	7
Northern Ireland	1
Norway	1
Portugal	1
Slovenia	1
South Africa	1
Switzerland	2
Turkey	1
UK	10
USA	59
Totaal	126

BIJLAGE 4: FORMULES EFFECTGROOTTE

Voor het berekenen van de effectgroottes zijn de volgende formules gebruikt:

(1) Cohen's d

$$ES = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{S_{pooled}}$$

Met deze formule wordt de effectgrootte (ES) bepaald door het verschil tussen gemiddelden te delen door de gepoolde standaarddeviatie.

(2) Gepoolde standaarddeviatie

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2}{n_1 + n_2 + \cdots + n_k - k}$$

$$S_p = \sqrt{S_p^2}$$

Voor het berekenen van de gepoolde standaarddeviatie zijn de bovenstaande formules gebruikt', met daarin de omvang van de steekproef (*n*), de standaarddeviaties (*s*) en het aantal groepen (*k*). Met de tweede formule wordt vanuit de gepoolde variantie de gepoolde standaarddeviatie berekend.

(3) Hedges g

$$ES' = \left[1 - \frac{3}{4(n_1 + n_2 - 9}\right] ES$$

(4) Correctie SE

$$SE = \sqrt{\frac{n_{G1} + n_{G2}}{n_{G1} n_{G2}} + \frac{(ES')^2}{2(n_{G1} + n_{G2})}}$$

Om de effectgrootte en bijbehorende standaard error (SE) te corrigeren voor een kleine steekproef zijn de formules van Hedges (1981) gebruikt.