

Op zoek naar het succes van inburgering: Methodologisch kader voor leerlastonderzoek

dr. Hans Kuhlemeier, drs. Kaatje Dalderop en drs. Laura van Hofwegen
Arnhem, Cito

Concept eindversie 3 december 2008

Inhoud

- 1 Inleiding
- 2 Algemeen methodologisch kader voor leerlastonderzoek
 - 2.1 Leerlast, niveau, rendement en benchmark
 - 2.2 Fasering van leerlastonderzoek
 - 2.3 Gebruik van een benchmark in het kader van leerlastonderzoek
- 3 Praktijkgericht leerlastonderzoek
 - 3.1 Stand van zaken
 - 3.2 Praktijkgericht leerlastonderzoek
 - 3.3 Initiëren, uitvoeren, monitoren en evalueren van praktijkgericht leerlastonderzoek
 - 3.4 Enkele aanbevelingen voor praktijkgericht leerlastonderzoek
- 4 De ontwikkeling van een benchmark NT2
 - 4.1 Inleiding
 - 4.2 Basale beslissingen
 - 4.3 Plan van aanpak voor het ontwikkelen van een benchmark NT2
 - 4.4. Enkele aanbevelingen voor benchmarkconstructie
- 5 (Quasi-)experimenteel leerlastonderzoek
 - 5.1 Inleiding
 - 5.2 Plan van aanpak voor (quasi-)experimenteel leerlastonderzoek
 - 5.3 Enkele aanbevelingen voor (quasi-)experimenteel leerlastonderzoek

Literatuur

Bijlagen

- 1 Andere gebruiksfuncties van een benchmark dan onderzoek
- 2 Kwaliteit en bruikbaarheid van bestaande bronnen voor benchmarkconstructie
- 3 Het gebruik van een benchmark NT2 in het kader van leerlastonderzoek

1 Inleiding

Medio 2008 heeft het Ministerie van VROM het Cito verzocht een kader op te stellen voor de uitvoerende fase van zogeheten leerlastonderzoek. Overeenkomstig het Programma van Eisen in de offerteaanvraag voor de voorbereidende fase van het leerlastonderzoek (kenmerk IKCnr38701) diende het kader te worden opgeleverd in twee producten: een theoretisch kader en een methodologisch kader.

Het theoretische kader geeft inzicht in de factoren die NT2-onderwijs in het kader van inburgering efficiënt en effectief maken (Dalderop e.a., 2008). Het benoemt de belangrijkste succesfactoren die bijdragen aan een zo hoog mogelijk taalniveau met een zo laag mogelijke leerlast voor vier groepen inburgeraars, te weten analfabeten, laag, midden en hoog opgeleiden.

De onderhavige publicatie bevat het methodologisch kader voor de uitvoering van leerlastonderzoek naar de succesfactoren uit het theoretisch kader. Het beschrijft de opzet van *practice based* onderzoek naar de invloed van potentiële succesfactoren op de leerlast en het taalniveau van inburgeraars.

Het methodologische kader voor de uitvoerende fase van leerlastonderzoek is als volgt opgebouwd. Na deze inleiding wordt eerst een algemeen kader geschetst waarbinnen leerlastonderzoek kan plaatsvinden (hoofdstuk 2). Daarna volgt een mogelijke opzet voor ontwikkelings- en praktijkgericht onderzoek naar de invloed van potentiële succesfactoren op de leerlast en het taalniveau van inburgeraars (hoofdstuk 3). Vervolgens wordt een voorstel gedaan voor een leerlastindicator ofwel benchmark aan de hand waarvan onderzoekers het effect van ontwikkelingsgerichte pilots op de leerlast en het taalniveau van inburgeraars kunnen evalueren (hoofdstuk 4). Het methodologisch kader besluit met het schetsen van de contouren van (quasi-) experimenteel onderzoek dat op termijn kan worden uitgevoerd naar de effectiviteit van NT2-interventies die zijn gebaseerd op de succesfactoren uit het theoretisch kader (hoofdstuk 5).

Bij de totstandkoming van dit advies is dankbaar gebruik gemaakt van de expertise van een begeleidingscommissie, bestaande uit Prof. Dr. Koen Jaspaert (KU Leuven, voorzitter van de begeleidingscommissie), Mw. Dr. Jeanne Kurvers (UVT), Mw. Drs. Marijke Huizinga (VU Amsterdam), Drs. Wim Coumou, Prof. Dr. Roel Bosker (RUG), Dr. Paul Tesser (Ministerie van VROM), Drs. Elwine Halewijn (Ministerie van VROM), Dhr. Ilhan Tuna (Ministerie van VROM) en Mw. Drs. Lieke Schmidt (Ministerie van SZW).

Onderdelen van dit advies zijn besproken met de deelnemers aan de workshop Onderzoeksopzet die gehouden is tijdens de Cito-werkconferentie “Op zoek naar het succes van inburgering” op 14 november 2008 te Utrecht. Tot slot is gebruik gemaakt van waardevolle opmerkingen en suggesties van dr. A. Béquin en Prof. dr. G. Maris (beiden Cito).

2 Algemeen methodologisch kader voor leerlastonderzoek

Dit hoofdstuk bevat het algemene kader voor de uitvoerende fase van leerlastonderzoek. Eerst worden enkele basale begrippen kort gedefinieerd (paragraaf 2.1). Vervolgens schetsen we een gefaseerd model van leerlastonderzoek dat bijdraagt aan innovatie en verbetering van het NT2-onderwijs (paragraaf 2.2). Tenslotte volgt een uiteenzetting over het gebruik van de benchmark in het kader van leerlastonderzoek (paragraaf 2.3).

2.1 Leerlast, niveau, rendement en benchmark

Vooruitlopend op de verdere conceptualisatie en operationalisatie elders in deze publicatie worden in deze paragraaf de begrippen leerlast, niveau, rendement en benchmark kort gedefinieerd.

Leerlast

Het begrip leerlast verwijst naar de inspanning die groepen inburgeraars zich getroosten om het taalniveau te bereiken dat nodig is voor het verkrijgen van een passend examen. Volgens het Deltaplan Inburgering geeft de leerlast aan “welke inspanning nodig is om een bepaald resultaat te behalen” en is de leerlast “een belangrijk gegeven als het gaat om het bepalen van de omvang van een passend programma voor de inburgeraar” (Ministerie van VROM, 2007, p. 16). Abbenhuis e.a. 1996) omschrijven leerlast als “het aantal onderwijsuren dat gebruikt is om van het ene niveau van taalvaardigheid naar het andere niveau te gaan” (p. 9). Leerlast wordt soms gedefinieerd in termen van duur en intensiteit. Onder de duur verstaan we de hoeveelheid tijd (bijvoorbeeld in weken) die de cursist gebruikt heeft vanaf het moment van instroom in het onderwijsprogramma. De intensiteit omvat het aantal uren en de spreiding van dat aantal in de tijd.

Niveau

Het niveau wordt gedefinieerd als het door de cursist bereikte taalniveau bij de start, op weg en na afronding van een passend examen (d.w.z. het inburgeringexamen of het Staatsexamen NT2 I of II). Daarbij wordt een onderscheid gemaakt in de taalvaardigheden luisteren, spreken, lezen en schrijven. De groei van de taalvaardigheid van de cursisten wordt uitgedrukt op de taalvaardigheidladder van het Common European Framework of Reference for Languages: Learning teaching, assessment (CEFR; Council of Europe, 2001).

Rendement

Indachtig hetgeen hierover vermeld is in het Deltaplan Inburgering (Ministerie van VROM, 2007) wordt rendement gedefinieerd als het daadwerkelijk door de cursist bereikte taalniveau in verhouding tot de door hem of haar geïnvesteerde tijd en lesuren (leerlast). Het is de opbrengst van een onderwijsprogramma bij de cursisten uitgedrukt in de hoeveelheid leerwinst per tijdseenheid en per aantal lesuren.

Benchmark

In navolging van Van der Linden (1996) definiëren we een leerlastindicator, tegenwoordig ook wel benchmark genoemd, als een instrument dat beschrijft ‘wie’ ‘welk niveau’ in ‘welke tijd’ onder ‘welke condities’ zou kunnen bereiken. Hierbij zij opgemerkt dat een benchmark in leerlastonderzoek uitsluitend een beschrijvende en geen vóórschrijvende functie heeft. Het instrument beschrijft hoeveel tijd/lesuren groepen inburgeraars gemiddeld nodig hebben, en niet in hoeveel tijd of lesuren zij dat niveau zouden kunnen of moeten behalen.

2.2 Fasering van leerlastonderzoek

Onderwijsvernieuwingen blijven te vaak “steken in uitwerking en voortijdige toepassing van het concept, zonder voldoende kritische reflectie en onderzoek naar effecten” (Onderwijsraad, 2006, p. 47). Alvorens een innovatie grootschalig in te voeren zijn “eerst vele jaren verkennend onderzoek, ontwikkelwerk en praktijkervaringen nodig om een beeld te krijgen van wat werkt en waarom en hoe” (p. 10). Dit lijkt ook te gelden voor een nog relatief onontgonnen en weerbarstig onderzoeksgebied zoals het NT2-onderwijs. In de uitvoeringsfase van het leerlastonderzoek zal van de belangrijkste succesfactoren uit het theoretisch kader worden bepaald wat het effect is op de leerlast in relatie tot het bereikte taalniveau. Hoe komen we nu te weten of deze succesfactoren daadwerkelijk uitvoerbaar en effectief zijn en in de vorm van interventies verder in het NT2-onderwijs verspreid kunnen worden? Alvorens een innovatie grootschalig uit te rollen kunnen verschillende stappen worden genomen (Onderwijsraad, 2006, p. 47):

- formuleren en uitwerken van het concept;
- uitwerken en verhelderen van het concept;
- aangeven welke negatieve effecten niet mogen optreden of moeten worden voorkomen;
- uitproberen van het concept op een beperkt aantal scholen, door leraren met een kritische, reflectieve, onderzoekende houding en met toetsing door onderzoekers;
- uitproberen en toetsen van het concept in een tweede tranche van scholen;
- herformuleren concept op basis van de ervaringen van leraren en onderzoekers;
- beproeven op grotere schaal, gaandeweg via hardere onderzoeksmethoden.

Aanvankelijk bestaat een innovatie vooral uit een verzameling uitgangspunten en goede bedoelingen. Om bruikbaar te zijn in de klas, moet de innovatie worden geconcretiseerd in onderwijsdoelen, didactische werkvormen, leeractiviteiten, lesmateriaal, media en toetsmateriaal (Tyler, 1949; De Kok-Damave, 1980; Van Gelder, Pool, Peters & Sixma; 1973; De Corte e.a., 1976). Ter verdere ontwikkeling van een innovatie zijn in eerste instantie kleinschalige pilots nodig waarin onderwijsgeveenden, onderzoekers en andere betrokken zoals leerplanontwikkelaars, begeleiders en toetsdeskundigen met elkaar samenwerken aan de vertaling van de veronderstelde succesfactoren in bijvoorbeeld lesmateriaal en docentenhandleidingen.

Is de innovatie voldoende doordacht en uitgekristalliseerd, dan kunnen uitvoerbaarheid en effectiviteit in wat minder kleinschalige pilots onder “gunstige omstandigheden” worden vastgesteld (Raudenbush, 2003). Bij gunstige omstandigheden kan men denken aan een situatie waar onderzoekers de implementatie zorgvuldig regisseren en controleren, waar gewerkt wordt met hoog opgeleide en enthousiaste docenten, waar er veel tijd en geld beschikbaar is voor materiaalontwikkeling en scholing van docenten et cetera. In de pilots wordt ook de leerlast vastgesteld die inburgeraars met verschillende opleidingsachtergronden en instapniveaus gebruiken om een bij hen passend taalniveau te bereiken. Een methodologisch kader voor kleinschalig praktijkgericht leerlastonderzoek naar de succesfactoren uit het theoretisch kader wordt geschetst in hoofdstuk 3 van deze publicatie.

Om het effect van een pilot naar het effect van succesfactoren uit het theoretisch kader op de leerlast en het taalniveau van groepen inburgeraars te kunnen evalueren is een criterium nodig. Op dit moment ontbreken betrouwbare gegevens over de tijd die inburgeraars gemiddeld nodig hebben om een passend taalniveau te behalen. Wij adviseren dan ook een

zogeneten leerlastindicator ofwel benchmark te ontwikkelen aan de hand waarvan onderzoekers kunnen evalueren of de succesfactoren die in de pilots onderzocht worden inderdaad het veronderstelde positieve effect op de leerlast en het taalniveau van de cursisten hebben. Hoe onderzoekers deze benchmark kunnen ontwikkelen en toepassen is uiteengezet in hoofdstuk 4 en Bijlage 3 van deze publicatie.

Nadat het effect van de succesfactoren in relatief kleinschalige pilots onder gunstige omstandigheden is vastgesteld, kan de innovatie worden “opgeschaald”. Dit wil zeggen dat de onderzoekers de innovatie in andere settings en bij andere cursisten en docenten uittesten dan waarbij deze ontwikkeld is. Dit is het moment waarop “hard” onderzoek naar de effectiviteit van de innovatie in vergelijking met regulier onderwijs op zijn plaats is (Raudenbush, 2003). Een belangrijke vraag is daarbij in hoeverre de onder gunstige omstandigheden vastgestelde positieve effecten stand houden zonder dat aan het karakter van de innovatie afbreuk wordt gedaan. Bij het “exporteren” van de interventie kunnen zich allerlei problemen voordoen. Zo heeft de onderzoeker in de nieuwe situatie meestal minder controle over de implementatie van de interventie, zijn de leraren vaak minder enthousiast, zijn er minder faciliteiten beschikbaar voor materiaalontwikkeling, scholing en begeleiding en moeten docenten het materiaal dikwijls aan hun specifieke omstandigheden aanpassen. Dit kan het eerder onder gunstige omstandigheden vastgestelde positieve effect van de innovatie nadelig beïnvloeden. Een aanzet tot een methodologisch kader voor (quasi-)experimenteel leerlastonderzoek is opgenomen in hoofdstuk 5 van deze publicatie.

Nadat overtuigende bewijs voor de effectiviteit van de succesfactoren is verzameld en de uitvoerbaarheid is aangetoond in een brede range van toepassingssituaties is de tijd rijp voor grootschalige verspreiding van de interventie in het NT2-onderwijs.¹

De algemene fasering van leerlastonderzoek is besproken met de begeleidingscommissie en met degenen die tijdens de Cito-werkconferentie “Op zoek naar het succes van inburgering” deelnamen aan de workshop onderzoek en kon daarbij op algehele ondersteuning rekenen. Geopperd werd om na elke fase te bekijken in hoeverre het uitvoeren van de volgende fase opportuun is.

2.3 Gebruik van een benchmark in het kader van leerlastonderzoek

Mede naar aanleiding van overleg met leden van de begeleidingscommissie adviseren wij de benchmark de komende jaren alleen te gebruiken voor onderzoeksdoeleinden. De benchmark wordt geconstrueerd op basis van een representatieve steekproef van alle aanbieders die NT2-onderwijs verzorgen in Nederland (zie hoofdstuk 4). Deze steekproef van instellingen omvat zowel effectieve als ineffectieve aanbieders. De geconstrueerde benchmark representeert hiermee de uitgangssituatie voorafgaand aan de kwaliteitsverbetering die van brede implementatie van de succesfactoren uit het theoretisch kader in het NT2-onderwijs verwacht mag worden. Leden van de begeleidingscommissie hebben ervoor gewaarschuwd dat het niveau van het NT2-onderwijs sinds de invoering van de marktwerking sterk lijkt te zijn gedaald. Sinds het openen van de markt zijn er allerlei instellingen, bureaus en bureautjes op het toneel verschenen waar cursisten onder ongunstige omstandigheden les krijgen van slecht opgeleide docenten. Zo zijn er tegenwoordig voorbeelden van ongebruikelijke locaties waar

¹ Cronen, Silver-Paculla en Condelli (2006) definiëren een interventie als “a comprehensive program that includes curriculum development, adaptation of materials, teacher training, ongoing teacher support, and assessment” (p. 7).

nauwelijks een schoolbord is, waar cursisten uit kostenoverwegingen niet met bestaande leergangen werken, waar gewerkt wordt met via knip- en plakwerk tot stand gekomen lesmateriaal, waar cursisten les krijgen van personen die hooguit een korte cursus hebben gevolgd et cetera. De leden van de begeleidingscommissie spraken de vrees uit dat een op de huidige situatie geënte benchmark tot een relatief hoge schatting van de leerlast zal leiden. Sommigen leden verwachtten zelfs dat het aantal lesuren aanmerkelijk hoger zal uitvallen dan vóór de Wet Inburgering. Zolang de benchmark alleen voor onderzoeksdoeleinden wordt gebruikt, is dit niet bezwaarlijk. Om vast te stellen of de pilots en interventies daadwerkelijk effect gesorteerd hebben, is nu eenmaal een hard criterium nodig (en dan is de landelijke verdeling van de leerwinst een logische keuze). Stel nu echter dat de overheid deze deels op slecht onderwijs gebaseerde uitgangssituatiemeting zou gebruiken om de duur van de aangeboden trajecten te bepalen. Een mogelijk gevolg zou kunnen zijn dat cursisten in de toekomst nog meer onderwijsaanbod (d.w.z. modules en cursussen) voorgeschoteld krijgen dan in de huidige situatie reeds het geval is. Met een dergelijke verspilling van tijd en middelen zal noch de overheid noch de cursist erg gelukkig zijn (met uitzondering van een deel van de onderwijsaanbieders die er financieel garen bij spinnen). Wij adviseren dan ook de te construeren benchmark de komende jaren alleen voor onderzoeksdoeleinden te gebruiken. De ontwikkelingsgerichte pilots en interventies (zie hoofdstuk 3) zullen laten zien hoeveel sneller cursisten kunnen leren als NT2-onderwijs in overeenstemming met de succesfactoren uit het theoretisch kader wordt vormgegeven (door competente docenten onder gunstige omstandigheden). Deze onder betere dan de huidige omstandigheden verzamelde leerlastgegevens moeten naar de mening van de begeleidingscommissie worden verdisconteerd in het opstellen van leerlastadviezen voor de overheid, cursisten, onderwijsaanbieders en gemeenten. Het is dan ook zaak om de uitkomsten van de ontwikkelingsgerichte pilots en de interventies af te wachten alvorens realistische aanbevelingen te doen. Voor een uiteenzetting over het mogelijke gebruik van een benchmark voor andere doelen dan onderzoek wordt verwezen naar Bijlage 1. Daarin is een onderscheid gemaakt in gebruik van een benchmark ter ondersteuning van het overheidsbeleid, ter advisering van individuele cursisten bij de intake, ter begeleiding van cursisten tijdens het inburgeringprogramma, ter evaluatie van onderwijsprogramma's door onderwijsaanbieders en ter evaluatie van aanbieders door gemeenten.

3 **Praktijkgericht leerlastonderzoek**

In het voorgaande hoofdstuk is een gefaseerd model geschetst van leerlastonderzoek dat bijdraagt aan innovatie en verbetering van het NT2-onderwijs. Betoogd is dat een fasering van kleinschalig naar grootschalig leerlastonderzoek en van zachte naar harde bewijsvoering wenselijk is om het NT2-onderwijs meer in overeenstemming te brengen met de succesfactoren uit het theoretisch kader (Dalderop e.a., 2008). Dit hoofdstuk bevat een beschrijving van de opzet en de uitvoering van het ontwikkelings- en praktijkgerichte onderzoek dat nodig is alvorens een bewezen effectieve interventie grootschalig kan worden ingevoerd.

3.1 **Stand van zaken**

Kenmerkend voor het huidige NT2-onderwijs is dat de succesfactoren uit het theoretisch kader in de onderwijspraktijk meestal niet in hun zuivere en meest effectieve vorm voorkomen. NT2-onderwijs is in hoge mate eclectisch in de zin dat docenten een brede mix aan onderwijsvormen toepassen. Onder de noemer van eenzelfde innovatie gaan in de onderwijspraktijk zeer diverse aanpakken schuil. Zo zijn er van de innovatie “dual” vele varianten in omloop die sterk verschillen in theoretische achtergrond, uitwerking en uitvoering (Dalderop e.a., 2008). Hoewel er uitzonderingen zijn (bijv. Deutekom, 2007, 2008) moeten innovaties eerst verder doordacht en ontwikkeld worden alvorens ze in kleinschalige settings op hun uitvoerbaarheid en effectiviteit te kunnen toetsen. In reeds uitgevoerde en nog lopende pilots zijn nieuwe materialen en procedures ontwikkeld die meestal niet onmiddellijk geschikt zijn voor gebruik door docenten en cursisten van andere onderwijsaanbieders.² In sommige gevallen is er in de pilots al enige empirische evidentie verzameld over de uitvoerbaarheid en effectiviteit van de innovatie in termen van leerlast en taalniveau. De resultaten van deze kleinschalige pilots zijn vaak veelbelovend, zij het dat de gegevens meestal zijn verzameld in hooguit enkele lokale settings. “Kant-en-klare” interventies die bewezen effectief zijn en zich lenen voor grootschalige verspreiding in het NT2-onderwijs lijken op dit moment nog niet voorhanden.

3.2 **Praktijkgericht leerlastonderzoek**

Overeenkomstig het Programma van Eisen in de offerteaanvraag van het leerlastonderzoek (kenmerk IKCnr38701) dient onderzoek naar het effect van de potentiële succesfactoren op de leerlast en het taalniveau van inburgeraars praktijkgericht te zijn. In dit type onderzoek, ook wel design-based, ontwikkelingsgericht of ontwerpgericht onderzoek genoemd, werken onderzoekers intensief samen met praktijkdeskundigen aan het ontwikkelen, evalueren en het verbeteren van onderwijs. Wang en Hannafin (2005) definiëren het als “a systematic but flexible methodology aimed to improve educational practices through iterative analysis, design, development, and implementation, based on collaboration among researchers and practitioners in real-world settings, and leading to contextually-sensitive design principles and theories” (p. 6). Van den Akker e.a. (2008, p. 4) geven de volgende kenmerken van ontwikkelingsgericht onderzoek:

² In de voorbereiding van een interventieonderzoek naar het effect van directe instructie van taalvaardigheid aan laag opgeleide volwassenen in de Verenigde Staten is voor het ontwikkelen van de interventie een speciale openbare aanbesteding georganiseerd waar *intervention developers* op konden inschrijven (Cronen, Silver-Pacuilla & Condelli, 2006).

- Interventionist: the research aims at designing an intervention in the real world.
- Iterative: the research incorporates a cyclic approach of design, evaluation and revision.
- Process-oriented: a black box model of input-output measurement is avoided; the focus is on understanding and improving interventions.
- Utility-oriented: the merit of a design is measured, in part, by its practicality for users in real contexts.
- Theory-oriented: the design is (at least partly) based upon theoretical propositions; and field testing of the design contributes to theory building.

Ontwikkelingsgericht onderzoek naar de uitvoerbaarheid en effectiviteit van innovaties kan leiden tot praktijkgerichte kennis in de trant van "If you want to design intervention X [for the purpose/function Y in context Z], then you are best advised to give that intervention the characteristics A, B, and C [substantive emphasis], and to do that via procedures K, L, and M [procedural emphasis], because of arguments P, Q, and R." (Van den Akker, 1999, p. 9). Het grote voordeel van ontwikkelingsgericht onderzoek is dat onderwijsvernieuwingen vanwege de wisselwerking tussen theorievorming, praktijk en onderzoek meer kans van slagen hebben (Design-Based Research Collective, 2003; De Corte & Verschaffel, 2001).

Een nadeel van praktijkgericht onderzoek is de mogelijke vermenging van de rollen van ontwikkelaar en onderzoeker die een objectieve evaluatie van uitvoerbaarheid en effectiviteit in de weg kan staan. Verder is het vaak moeilijk om gevonden effecten aan afzonderlijke verklarende factoren toe te schrijven (zou men dit in het ontwerpstadium al willen). Ten slotte is generalisatie van de steekproef naar andere settings onder meer vanwege de kleine steekproefomvang beperkt mogelijk (Van den Akker, 1999). Is de interventie voldoende uitgekristalliseerd en doorontwikkeld, dienen de uitvoerbaarheid en effecten daarom in wat minder kleinschalige settings onder "gunstige omstandigheden" te worden vastgesteld (bijvoorbeeld in een situatie waar onderzoekers de implementatie zorgvuldig regisseren en controleren, waar gewerkt wordt met hoog opgeleide en enthousiaste docenten, waar er veel tijd en geld beschikbaar is voor materiaalontwikkeling en scholing van docenten, et cetera (Raudenbush, 2003).

3.3 Initiëren, uitvoeren, monitoren en evalueren van praktijkgericht leerlastonderzoek

Praktijkgericht onderzoek naar het effect van veronderstelde succesfactoren op leerlast en taalniveau veronderstelt dat docenten en aanbieders met onderzoekers willen en kunnen samenwerken aan de verbetering van het onderwijs in de klas. Hoe zou het ministerie van VROM onderwijsvernieuwing en onderzoek dat daar dienstbaar aan is, kunnen initiëren? Een eerste stap is het verspreiden van de kennis over de potentiële succesfactoren uit het theoretisch kader in het NT2-veld, bijvoorbeeld via publicaties op het Internet en scholingsbijeenkomsten voor docenten (Snoeken, 2008). Hiermee zal de beoogde wisselwerking tussen theorievorming, praktijkervaring en wetenschappelijk onderzoek vermoedelijk een onvoldoende krachtige impuls krijgen. Hieronder schetsen we een stapsgewijze aanpak die op dit moment wordt toegepast in het kader van het Actieprogramma Onderwijs Bewijs en die de beoogde wisselwerking waarschijnlijk wel tot stand kan brengen (OCW-SenterNovem, 2008):

- Reserveer een budget voor ontwikkelingsgericht onderzoek naar potentiële succesfactoren voor een periode van bijvoorbeeld vier à vijf jaar.
- Communiceer dat dit budget beschikbaar is en nodig onderwijsgevend, onderzoekers en andere betrokkenen uit onderzoeksvoorstellen in te dienen.

- Formuleer een Actieprogramma Ontwikkelingsgericht en Experimenteel Onderzoek NT2 (met als onderzoeksthema's de veronderstelde succesfactoren en de te ontwikkelen innovaties).
- Stimuleer partnerschap door te eisen dat alleen gezamenlijk ingediende voorstellen van aanbieders en onderzoeksinstellingen gehonoreerd zullen worden.
- Formuleer criteria voor de opzet en uitvoering van ontwikkelingsgericht onderzoek en communiceer dat ingediende voorstellen op deze criteria beoordeeld worden.
- Installeer een Onafhankelijke Toetsingscommissie (OTC) van deskundigen die ingediende voorstellen aan de hand van de vooraf opgestelde criteria zal beoordelen.
- Organiseer een of meer *dating* bijeenkomsten waarbij onderwijsgeevenden, onderzoekers en andere betrokkenen (leerplanontwikkelaars, onderwijsbegeleiders, nascholers en toetsdeskundigen) elkaar ontmoeten en op succesfactor of interventie aan elkaar "gematcht" kunnen worden.
- Lanceer een website met een onder meer een zogeheten *matching tool* waarmee onderzoekers en onderwijsgeevenden ideeën voor onderzoek kunnen uitwisselen en een samenwerkingspartner kunnen vinden.

Tijdens de workshop onderzoek van de veldraadpleging van 14 november 2008 werd de hierboven geschetste aanpak unaniem onderschreven.

Welke structuur is noodzakelijk om pilots te initiëren?

Tijdens de fase van het initiëren van praktijkgerichte pilots kan de Aanbestedende Dienst (het ministerie van VROM) een Stuurgroep in het leven roepen die beslist over de toewijzing van middelen voor de uitvoering van het Actieprogramma Ontwikkelingsgericht en Experimenteel Onderzoek NT2. Een Onafhankelijke Toetsingscommissie (OTC) kan de ingediende onderzoeksvoorstellen beoordelen en de Aanbestedende Dienst adviseren over de uitvoering van de ingediende onderzoeksvoorstellen. De praktische uitvoering van de aanbesteding van kan in handen gelegd worden van een Programmabureau dat de procedure in opdracht van de Aanbestedende Dienst uitvoert en begeleidt. In het hierboven beschreven voorbeeld was dat SenterNovum, een agentschap van het Ministerie van Economische Zaken.

Welke eisen moeten aan uitvoerders worden gesteld?

Hieronder geven we enkele suggesties voor criteria waaraan uitvoerders van praktijkgericht leerlastonderzoek moeten voldoen:

- De indiener van een onderzoeksvoorstel moet een samenwerkingsverband zijn van één of meerdere wetenschappers of wetenschappelijke instituten met één of meerdere aanbieders van NT2-onderwijs. De deelnemers aan de workshop Onderzoek van de Cito-werkconferentie "Op zoek naar het succes van inburgering" waren van mening dat behalve aanbieders en onderzoeksinstellingen ook gemeenten als een volwaardige partner bij het opstellen van onderzoeksvoorstellen betrokken moeten worden.
- Het samenwerkingsverband moet over de benodigde expertise beschikken om de pilot met succes te kunnen opzetten, uitvoeren en rapporteren. Deskundigheid kan onder meer blijken uit recente (internationale) wetenschappelijke publicaties. De deelnemers aan de workshop onderzoek van de Cito-werkconferentie "Op zoek naar het succes van inburgering" vonden dat NT2-deskundigheid voor onderzoekers een eerste vereiste moet zijn.
- Uitvoerders moeten voldoen aan de specifieke criteria van de Onafhankelijke Toetsingscommissie uitvoerders en de door hen ingediende onderzoeksvoorstellen beoordeelt en prioriteert.

- Personen en instellingen die onderzoek uitvoeren in opdracht van Ministeries moeten meestal voldoen aan de 'Algemene Rijksvoorwaarden voor het Verstreken van Opdrachten tot het Verrichten van Diensten' (Arvodi, 2008). Verder gelden eisen ten aanzien van voortgangsrapportage, evaluatie en kennisdeling.

Aan welke criteria moeten voorstellen voor ontwikkelingsgericht leerlastonderzoek voldoen?

De Onafhankelijke Toetsingscommissie beoordeelt en prioriteert de ingediende onderzoeksvoorstellen. Daartoe dienen gunningcriteria beschikbaar te zijn die in de vorm van een Leidraad voor indieners van onderzoeksvoorstellen gepubliceerd kunnen worden. Het opstellen van deze criteria is tijdrovend en luistert nauw. Hieronder volgen enkele eerste suggesties voor de eisen waaraan onderzoeksvoorstellen voor praktijkgericht leerlastonderzoek zouden moeten voldoen (vgl. OCW-SenterNovum, 2008):

- Betreft het voorstel een praktijk- en ontwikkelingsgericht onderzoek?
- Heeft het onderzoeksvoorstel betrekking op een of meer van de succesfactoren of interventies uit het theoretisch kader?
- Bevat het onderzoeksvoorstel een nadere conceptualisatie en operationalisatie van de te onderzoeken succesfactoren?
- Is de vraag- of probleemstelling helder omschreven en goed uitgewerkt?
- Bevat het voorstel een nadere theoretische onderbouwing van de vraag- en probleemstelling op basis van relevante literatuur of bestaand onderzoek?
- Zijn de deelnemende onderzoekers voldoende competent, bijvoorbeeld blijkend uit recente (internationale) publicaties?
- Zijn de docenten op de werkvloer betrokken geweest bij de totstandkoming van het onderzoeksvoorstel? Tijdens de workshop onderzoek werd opgemerkt dat betrokkenheid van docenten voor het welslagen van de pilots van doorslaggevende betekenis is. Voorkomen moet worden dat directies over de hoofden van de docenten heen over deelname aan het praktijkgericht leerlastonderzoek beslissen.
- Is aangegeven hoe de indiener het gewenste partnerschap denkt op te bouwen en te onderhouden?
- Is aangegeven hoe de indiener docenten en cursisten voor de pilots denkt te enthousiasmeren?
- Is aangegeven op welke terreinen de samenwerking tussen praktijkdeskundigen en onderzoekers zal plaatsvinden?
- Is aangegeven hoe de samenwerking en de verantwoordelijkheidsverdeling tussen de verschillende partijen wordt vormgegeven?
- Is aangegeven hoe de indiener van het voorstel ruimte voor ontwikkeling en onderzoek in het takenpakket van de docenten denkt te creëren?
- Bevat het voorstel een operationalisatie van de te onderzoeken variabelen?
- Is de verwachte opbrengst van de pilot goed onderbouwd? Zijn deze verwachtingen realistisch? Is aangegeven wat de veronderstelde positieve effecten van de pilot op leerlast en taalniveau zijn? Idem dito voor mogelijk ongewenste neveneffecten?
- Is de interventie duidelijk omschreven? Omvat het voorstel het ontwikkelen van materiaal of procedures en zo ja op welke wijze deze ontwikkeld worden?
- Is aangegeven hoe de indieners willen bevorderen dat docenten de succesfactoren overeenkomstig de uitgangspunten implementeren (bijvoorbeeld via het gezamenlijk uitwerken van innovatie- en gebruikersprofielen)?
- Is aangegeven hoe de indiener de check op de implementatie wil uitvoeren?

- Bevat het voorstel een uiteenzetting over de zogeheten middel-doel relaties? Is met ander woorden aangeven op welke wijze de indiener van het onderzoeksvoorstel de cursisten uit de pilot sneller naar een hoger taalniveau denkt te brengen?
- Is uitgelegd hoe en hoe vaak de veronderstelde positieve effecten op leerlast en taalniveau gemeten zullen worden?
- Is toegelicht hoe groot de pilotgroep zal zijn en hoe deze ongeveer zal zijn samengesteld (aantal scholen, klassen, leerlingen, samenstelling naar opleidingsniveau, beginniveau, land en taal van herkomst et cetera)?
- Is aangegeven welke maatregelen de indiener wil treffen om uitval en afwezigheid van cursisten te voorkomen?
- Is voorzien in meting van de leerlast en de taalvaardigheid aan ten minste het begin en einde van der pilot (en bijvoorkeur ook in de periode daartussen)? Is de indiener desgevraagd bereid om daartoe de instrumenten uit het onderzoek naar de constructie van de benchmark te gebruiken en om externe proefleiders op zijn of haar school toe te laten?
- Is aangegeven welke onderzoeksmethoden worden gebruikt (bijvoorbeeld: observatie, interview, vragenlijst en/of inhoudsanalyse van documenten)?
- Is aangegeven welke faciliteiten, materialen en hulpmiddelen nodig zijn om de pilot te realiseren?
- Is aangegeven welke procedures en richtlijnen binnen en buiten de pilot nodig zijn om de pilot te laten slagen?
- Is de planning adequaat? Wanneer start het project en wanneer eindigt het? In welke fasen wordt het projectresultaat gerealiseerd? Wanneer worden de tussenresultaten opgeleverd? Wie gaat wanneer welke activiteit uitvoeren? Zijn er tussen de verschillende activiteiten relevante afhankelijkheden en zo ja, welke?
- Is de begroting adequaat?
- Is er een risico-inventarisatie van de haalbaarheid van (onderdelen) van de pilot en een beschrijving van eventuele maatregelen om de risico's te beheersen?
- In hoeverre is er sprake van een redelijke verhouding tussen de te verwachten opbrengsten en de te maken kosten?
- Is aangegeven hoe de pilot, mits succesvol, kan worden opgeschaald?

Welke structuur is noodzakelijk om de pilots te begeleiden, monitoren en evalueren?

De hierboven beschreven aanbestedingsprocedure zal naar verwachting resulteren in een aantal gehonoreerde voorstellen voor praktijkgericht onderzoek naar het effect van de succesfactoren uit het theoretisch kader op leerlast en taalniveau. Een landelijke overkoepelende projectorganisatie zou de pilots gedurende een aantal jaren kunnen coördineren, begeleiden, monitoren en evalueren. Tot de taken van deze landelijke projectorganisatie zouden kunnen behoren:

- Het bevorderen van de communicatie en uitwisseling van kennis en ervaring tussen de pilots;
- Het begeleiden en ondersteunen van de pilots, bijvoorbeeld via het inrichten van een helpdesk, het inrichten van een website waarin betrokken met elkaar kunnen communiceren en elkaar kunnen adviseren inzake gemeenschappelijk ervaren problemen, en het verzorgen van trainingen voor docentonderzoekers;
- Het bewaken en evalueren van de voortgang en financiering van de pilots;
- Het bevorderen van de verspreiding van de opbrengsten van de pilots naar andere scholen.

Bij de landelijke overkoepelende projectorganisatie valt te denken aan de volgende constructie:

- een Stuurgroep die de opdrachtgever adviseert inzake het coördineren, ondersteunen, monitoren en evalueren van de pilots;
- een Begeleidingscommissie die de Stuurgroep adviseert inzake inhoudelijke, onderzoekstechnische en innovatiestrategische kwesties;
- een Projectbureau dat in opdracht van de Stuurgroep zorg draagt voor de dagelijkse communicatie, ondersteuning, monitoring en evaluatie van de pilots.

3.4 Enkele aanbevelingen voor praktijkgericht leerlastonderzoek

Mede op grond van de op 14 november in Utrecht georganiseerde veldraadpleging en de daarop volgende schriftelijke commentaarronde kunnen de volgende aanbevelingen voor de opzet van ontwikkelings- en praktijkgericht leerlastonderzoek worden gedaan.

Opbouwen en onderhouden van partnerschap

Wellicht de belangrijkste voorwaarde voor een succesvol NT2-interventieonderzoek is het opbouwen en onderhouden van “partnerschap” tussen de aanbieders, onderzoekers en gemeenten die aan het onderzoek deelnemen. Praktijkgericht onderzoek waarin onderwijsgevenenden samen met onderzoekers werken aan het vergaren van evidence-based kennis vergt een langere en intensievere voorbereiding dan regulier onderzoek. Een voorwaarde voor een vruchtbare samenwerking is dat de pilot voor scholen zinvol is en praktische informatie oplevert die bijdraagt aan het oplossen van een praktisch probleem op de school. In de voorbereidingsfase van een pilot kunnen onderzoekers een onderwijsaanbieder daarom niet “cold turkey” benaderen. Zij moeten ruimschoots te tijd nemen om het doel van het onderzoek, de innovatie, het onderzoeksontwerp en de wijze van gegevensverzameling aan directie en onderwijsgevenenden uit te leggen. Als er ruime aandacht is voor het opbouwen van wederzijds vertrouwen en open communicatie is de kans op een getrouwe uitvoering van de innovatie het grootst.

In de schriftelijke commentaarronde na de workshop onderzoek op de Cito-werkconferentie “Op zoek naar het succes van inburgering” wees een van de respondenten op het belang van het communiceren van het algemene belang van ontwikkelingsgericht leerlastonderzoek (we willen immers weten wat ertoe doet als het gaat om het ondersteunen van inburgering). Verder wees zij erop dat marktwerking, en de daarmee gepaard gaande onderlinge concurrentie tussen instellingen, het dienen van het eigen belang kan stimuleren. Sommige instellingen zullen niet mee willen werken omdat de pilot niet in het lopende contract past of omdat het niet past in de strategie waarmee men de markt wil veroveren. Andere instellingen zullen wel mee willen werken en de pilot willen gebruiken om zich te profileren in de markt. Daar hoeft niets mis mee te zijn, maar toch: de instelling kan bij het uitproberen van de innovatie het initiatief naar zich toe trekken en dingen gaan doen die gelet op het doel van het onderzoek niet wenselijk of nodig zijn.

Gezamenlijke ontwikkeling van innovatieprofielen

Aanvankelijk bestaat een innovatie vooral uit uitgangspunten en goede bedoelingen, maar nog niet uit uitgewerkt lesmateriaal, docentenhandleidingen, media, toetsmateriaal en dergelijke. Het gezamenlijk opstellen en uitwerken van een innovatieprofiel is een effectieve methode om een interventie te ontwikkelen en docenten ertoe te brengen de succesfactoren uit te voeren overeenkomstig de uitgangspunten van de innovatie (Leithwood & Montgomery, 1987). Een innovatieprofiel omvat onder meer een gedetailleerde beschrijving van wat de docent kan of moet doen en laten bij het initiëren en implementeren van effectief onderwijs in

de klas of daarbuiten. Het geeft tevens aan wanneer er sprake is van een onvoldoende, minimale, acceptabele en volledige ofwel ideale implementatie. Drempelcriteria zijn onderdeel van een innovatieprofiel en geven aan vanaf welk niveau van het profiel er sprake is van implementatie (Van den Akker & Voogt, 1994). Het gezamenlijk met docenten ontwikkelen en bijstellen van innovatieprofielen verhoogt de kans dat de interventie praktisch relevant is en wordt uitgevoerd overeenkomstig de uitgangspunten van de vernieuwing.

Gezamenlijk werken aan een getrouwe implementatie

Kleinschalig ontwikkelingsgericht leerlastonderzoek heeft tot doel eerste empirische gegevens te verzamelen op basis waarvan de innovatie verbeterd kan worden. Een belangrijke vraag daarbij is in hoeverre getrouwe uitvoering van de potentiële succesfactoren in de klas mogelijk is. Docenten kunnen hun steentje bijdragen door de interventie zoveel mogelijk volgens het gezamenlijk vastgestelde innovatieprofiel uit te voeren, door mee werken aan kleinschalig onderzoek in hun klas en door mee te schrijven aan bijvoorbeeld lesmateriaal, toetsmateriaal en handleidingen.

Check op de implementatie

Docenten neigen ertoe een innovatie anders uit te voeren dan de ontwikkelaars ervan bedoeld hebben (bijvoorbeeld als zij vast houden aan oude routines omdat zij erg tevreden zijn met hun “eigen” manier van lesgeven). De onderzoeker dient dan ook een zorgvuldige check op de implementatie uit te voeren. Het nauwkeurig observeren en registreren van de manier waarop docenten de succesfactoren implementeren is ook van belang om te begrijpen waarom de innovatie al dan niet succes heeft gehad.

Docenten betrekken bij verdere verspreiding van de innovatie

Nadat de uitvoerbaarheid en effectiviteit van de succesfactoren zijn vastgesteld kunnen de docenten uit de pilots bijdragen aan de verspreiding van de interventie buiten de ontwikkel- en onderzoeksscholen. Bijvoorbeeld door docenten van andere scholen op hun school uit te nodigen en door trainingen in de getrouwe uitvoering van de interventie te verzorgen.

Ter beschikking stellen van onderzoeksmateriaal

Directies en onderwijsgeveden besluiten vaak tot deelname aan ontwikkelingsgericht onderzoek als zij verwachten dat het onderzoek positief aan de onderwijskwaliteit in hun school of klas zal bijdragen. Als “tegenprestatie” kunnen de onderzoekers de school na afloop van het onderzoek helpen bij de verdere invoering van de vernieuwing op de school, bijvoorbeeld door het in de pilot ontwikkelde materiaal gratis ter beschikking te stellen of door docenten die niet bij het onderzoek betrokken waren te scholen en te begeleiden bij de invoering van de succesfactoren in hun klassen. Hiermee krijgen de onderzoekers op hun beurt weer informatie die bij het landelijk “uitrollen” van de interventie van belang kan zijn.

Creëren van gunstige voorwaarden voor ontwikkelingsgericht onderzoek

Tijdens de veldraadpleging kwam naar voren dat de randvoorwaarden voor ontwikkelings- en praktijkgericht leerlastonderzoek in het NT2-onderwijs ongunstig zijn. De van boven opgelegde marktwerking werd unaniem gezien als een factor die leerlastonderzoek ernstig belemmert. Het verdient dan ook aanbeveling om in de pilots gunstige voorwaarden voor onderzoek te creëren, bijvoorbeeld door met gemeenten en aanbieders duidelijke afspraken te maken over de looptijd van het onderzoek en door in het takenpakket van de deelnemende docenten ruimte voor ontwikkeling en onderzoek te creëren. Naar de mening van de deelnemers aan de veldraadpleging moeten de huidige regels voor de open aanbesteding tijdelijk terzijde worden geschoven, zoals de gemeente Utrecht dat bijvoorbeeld deed bij de

opzet van de pilots in het project Participatie van Vrouwen uit Etnische Minderheidsgroepen (Bavet) Taal Totaal.

4 De ontwikkeling van een benchmark NT2

Een benchmark kan een belangrijke rol spelen bij het evalueren van praktijkgerichte pilots naar het effect van potentiële succesfactoren op leerlast en taalniveau. Dit vierde deel van het methodologisch kader beschrijft eerst de aanleiding tot het ontwikkelen van een benchmark, waarbij ook summier wordt ingegaan op hetgeen er al bekend is over de leerlast van inburgeraars (paragraaf 4.1). Daarna worden enkele elementaire beslissingen beschreven die aan het ontwerp van een benchmark ten grondslag liggen (paragraaf 4.2). Ten slotte wordt een plan van aanpak voor de ontwikkeling van de benchmark gepresenteerd waarmee de relatie tussen leerlast en taalniveau van verschillende groepen inburgeraars beschreven kan worden (paragraaf 4.3).

4.1 Inleiding

De afgelopen jaren zijn jaarlijks gemiddeld 35.000 cursisten met een inburgeringprogramma gestart. In de beleidsnotitie Deltaplan Inburgering (Ministerie van VROM, 2007) wordt gesteld dat het rendement van deze programma's laag is. De resultaten van de inburgering van oudkomers en nieuwkomers blijven achter wat van hen gevraagd wordt. Zowel analfabeten als laag, midden en hoog opgeleiden boeken op het terrein van de beheersing van het Nederlands vaak te weinig voortgang. In 2006 kwam bijvoorbeeld ruim de helft van de oudkomers na het volgen van een inburgeringprogramma niet verder dan het zeer beperkt taalniveau van NT2 A1. Van de oudkomers boekte iets meer dan de helft sinds de start van het programma geen meetbare vooruitgang (zij bleven op hetzelfde NT2 niveau). De nieuwkomers deden het slechts marginaal beter.

Aanleiding tot het ontwikkelen van een benchmark

Volgens het Deltaplan Inburgering geeft de leerlast aan “welke inspanning nodig is om een bepaald resultaat te behalen” en is de leerlast “een belangrijk gegeven als het gaat om het bepalen van de omvang van een passend programma voor de inburgeraar” (Ministerie van VROM, 2007, p. 16). In Nederland is nog weinig kwantitatief onderzoek gedaan naar de leerlast van onderwijsprogramma's voor inburgeraars. Op dit moment bestaan er geen betrouwbare schattingen van de investering die nodig is om tot een gewenst taalniveau te komen of om de voortgang in termen van bijvoorbeeld CEFR-taalniveaus vast te stellen. Om de leerlast te kunnen bepalen moet een relatie worden gelegd tussen de duur van een programma en het behaalde taalniveau. Het aantal uren van een programma moet worden gerelateerd aan de niveaustijging die de cursist heeft behaald. Daartoe dienen de leerlast en de start-, tussen- en eindniveaus van de cursisten nauwkeurig te worden vastgesteld. Naast het inrichten van innovatietrajecten en experimenten wil het Ministerie van VROM dan ook middelen ter beschikking stellen voor onderzoek naar de duur van inburgeringprogramma's voor verschillende groepen inburgeraars (Ministerie van VROM, 2007, p. 44). Dit leerlastonderzoek moet in 2009 van start gaan.

Wat is er al bekend over de leerlast van inburgeraars?

Een eerste poging tot het ontwikkelen van een benchmark, destijds nog leerlastindicator genoemd, is gedaan in het onderzoek van Abbenhuis e.a. (1996). Het instrument moest beschrijven onder welke condities cursisten bepaalde taalniveaus behalen. Leerlast werd daarbij omschreven als “het aantal onderwijsuren dat gebruikt is om van het ene niveau van taalvaardigheid naar het andere niveau te gaan” (p. 9). Behalve over de leerlast en de

niveaustappen verzamelden de onderzoekers een grote hoeveelheid gegevens over kenmerken van de cursist, het onderwijs en het onderwijsprogramma. Bij de meting van de taalvaardigheid van de cursisten gebruikten de onderzoekers zowel door de instellingen verzamelde toetsgegevens als zelf afgenomen toetsinstrumenten. Voor dit laatste werd uit bestaande NT2-toetsen een nieuw toetsinstrument samengesteld. Op basis van een responsegroep van 310 cursisten bepaalden de onderzoekers het percentage cursisten dat na 150, 300, 450, 600 en meer dan 600 uur al dan niet een niveaustap maakte. Uit het onderzoek bleek onder meer dat grote groepen cursisten geen meetbare progressie boekten, zelfs niet na vele honderden lesuren. Zo bleek ook na meer dan 600 uur ruim een kwart van de cursisten die op taalniveau 0 instapten nog geen niveaustap gerealiseerd te hebben. Cursisten die de stap van taalniveau 0 naar 1 realiseerden, hadden daartoe gemiddeld meer uren les gebruikt dan cursisten die de stap van taalniveau 1 naar 2 (en van 2 naar 3) zetten. De grootste toename van het aantal cursisten dat een stap zette, vond plaats vóór 450 uur (basiseducatie) of 300 uur (vavo). Meer uren les leidde in verhouding tot een relatief geringe verhoging van het percentage cursisten dat een stap zette. Dit wees er volgens Abbenhuis e.a. (1996) op dat “cursisten die na 450 resp. 300 uur nog geen stap hebben gezet, niet erg gebaat zijn bij meer van hetzelfde” (p. 2 van Bijlage 2). Een van leden van de begeleidingscommissie wees er overigen op dat onduidelijk is in hoeverre deze constatering voor het huidige NT2-onderwijs situatie nog steeds opgeld doet.

Bij de uitvoering van het onderzoek van Abbenhuis e.a. (1996) deed zich een groot aantal problemen voor. Voor een bloemlezing uit de onderzoekstechnische problemen waarmee de onderzoekers te kampen hadden, wordt verwezen naar Bijlage 2. Deze problemen geven aanleiding tot twijfel aan de representativiteit, betrouwbaarheid en validiteit van de onderzoeksuitkomsten. De ontwikkelde leerlastindicator is geëvalueerd tijdens een expertmeting. Een belangrijk kritiekpunt was het ontbreken van “voorspellingen over de kans ‘wie’ ‘welk niveau’ in ‘welke periode’ onder ‘welke condities’ zou kunnen bereiken, terwijl dat juist de essentie van een leerlastindicator zou moeten zijn” (Van der Linden, 1996, p. 4).

Een recentere leerlastindicator is afkomstig van de Commissie Franssen (2004). Deze commissie heeft een ruwe schatting opgesteld van het aantal lesuren dat nodig is voor het bereiken van de niveaus A1 en A2 onder optimale omstandigheden, dat wil zeggen: goed onderwijs onder leiding van een docent. De commissie baseerde de schatting van de leerlast niet op empirisch onderzoek, maar op een bevraging van deskundigen (Drs. W. Coumou, persoonlijke communicatie). De leerlast voor het bereiken van het niveau A1 werd voor laagopgeleiden geschat op 250 contacturen (waarbij de uren besteed aan huiswerk moeten worden opgeteld; zelfstudie is bij de meeste laagopgeleiden geen optie) en voor hoogopgeleiden op 150 contacturen (exclusief huiswerk en zelfstudie). Voor het bereiken van het niveau A2 werd de leerlast voor laagopgeleiden geschat op een uren-aantal van 750 uur en voor hoogopgeleiden op 450 uur. Zoals de Commissie Franssen zelf aangeeft, betreft het ruwe schattingen die slechts zeer ten dele zijn gebaseerd op empirisch onderzoek. Zo is de leerlast voor A2 gebaseerd op de vuistregel van Stoks en De Jong (1999) die stellen dat voor het bereiken van een naast hoger taalniveau telkens twee keer zoveel tijd nodig is als voor het eraan voorafgaande niveau. Deze vuistregel is gebaseerd op cijfers over onderwijs aan hogeschoolde jonge taalleerders die een vreemde taal leren van binnen de eigen taalgroep en is mogelijk niet zonder van toepassing op de benodigde aantallen lesuren NT2-onderwijs van volwassen inburgeraars.

Waarin verschilt de nieuwe benchmark van eerdere schattingen?

De beide indicaties van de leerlast zijn op verschillende momenten onder verschillende condities verzameld (waardoor ook onduidelijk is in hoeverre de onderscheiden groepen

inburgeraars volledig vergelijkbaar zijn). Bovendien zijn de schattingen opgesteld onder de randvoorwaarden van het “oude” inburgeringstelsel (WIN). Dat kende slechts een inspanningsverplichting. Sinds 1 januari 2007 is er sprake van een nieuw inburgeringstelsel (WI). Dit kent een resultaatverplichting in de zin dat inburgeraars binnen een bepaalde tijd een inburgeringsexamen moeten behalen. Daarnaast is er tegenwoordig sprake van marktwerking. Deze heeft tot doel de kwaliteit door onderlinge concurrentie te verbeteren, maar volgens leden van de begeleidingscommissie en deelnemers aan de veldraadpleging lijkt marktwerking in de praktijk ook ongewenste neveneffecten te hebben. Voor zover ons bekend is er nog geen onderzoek gedaan naar de relatie tussen leerlast en rendement in de situatie van het nieuwe inburgeringstelsel onder de condities van marktwerking. Een instrument dat duidelijk maakt hoeveel lesuren groepen cursisten in de huidige situatie gebruiken om bepaalde taalniveaus te bereiken is op dit moment niet voorhanden.

Anders dan de ruwe schattingen van de commissie Franssen, reikt de te construeren benchmark NT2 verder dan de taalniveaus A1 en A2. Het bestrijkt de volledige range van niveau 0 tot en met niveau B2. De leerlast wordt niet alleen geschat voor laag en hoog opleiden, maar ook voor analfabeten en middenopgeleiden. Indien daar aanleiding toe is en de informatie beschikbaar is, kan een onderscheid naar oudkomer en nieuwkomer worden gemaakt (eventueel binnen de vier categorieën van het opleidingsniveau). Verder beschrijft de benchmark NT2 de gebruikte tijd/lesuren conditioneel op het startniveau van de cursist (gegeven een bepaald opleidingsniveau). Immers, een groep laagopgeleide inburgeraars die instromen op niveau 0 zal bijvoorbeeld meer tijd/uren gebruiken dan een in alle overige opzichten vergelijkbare groep laag opgeleiden die op niveau A1 binnenkomt. Ten slotte wordt alleen gebruik gemaakt van gegevens die met betrouwbare en valide meetinstrumenten onder gecontroleerde omstandigheden verzameld zijn (waarbij naast de gemiddelde leercurven van de onderscheiden groepen inburgeraars ook de spreiding rond dit gemiddelde in de constructie van de benchmark betrokken wordt).

4.2 Basale beslissingen

In deze paragraaf worden enkele basale beslissingen beschreven die ten grondslag liggen aan het plan van aanpak voor het ontwikkelen van een benchmark dat in paragraaf 4.3 gepresenteerd wordt.

Aan welke eisen moet een benchmark voldoen?

Alvorens het plan van aanpak voor de constructie van een benchmark nader uit te werken, worden eerst enkele kwaliteitseisen geformuleerd waaraan een benchmark moet voldoen. De kwaliteitseisen zijn deels afhankelijk van het doel waarvoor en de situatie waarin de benchmark gebruikt gaat worden. Voor een bespreking van andere functies van een benchmark dan onderzoek wordt verwezen naar Bijlage 1. We beperken ons hier tot kwaliteitseisen die gelden voor het gebruik van een benchmark voor onderzoeksdoeleinden, d.w.z. om het effect van een ontwikkelingsgerichte pilot of een veldexperiment op de leerlast en het taalniveau van groepen inburgeraars te evalueren. Daartoe moet de benchmark onder meer aan de volgende algemene kwaliteitseisen voldoen: voldoende gedifferentieerd, landelijk representatief, betrouwbaar, valide en efficiënt.

Voldoende gedifferentieerd. De gemiddelde cursist is een niet bestaande abstractie. In de praktijk van het inburgeringsonderwijs hebben intakers, docenten, begeleiders en onderzoekers te maken met verschillende soorten inburgeraars. Inburgeraars verschillen op allerlei variabelen die van belang zijn voor het beschrijven van de tijd cursisten gebruiken voor het behalen van het overeengekomen streefniveau. Dit betekent dat de benchmark

voldoende gedifferentieerd moet zijn om de leerlast voor al deze subgroepen te kunnen beschrijven. De steekproef, het onderzoeksontwerp en de instrumenten moeten het mogelijk maken om de parameters (d.w.z. intercept, regressiegewichten en varianties) per subgroep betrouwbaar te schatten. Een relevante vraag is dan hoeveel en welke groepen inburgeraars te onderscheiden zijn. Eerder onderzoek (zie ook Bijlage 2) doet vermoeden dat de leersnelheid van inburgeraars sterk afhankelijk is van de opleidingsachtergrond, het taalvaardigheidsniveau bij aanvang van het inburgeringonderwijs (o.a. Abbenhuis e.a., 1996; Duijm, 1996; Van der Linden, 1996) en wellicht ook van de vraag of de inburgeraar oud- of nieuwkomer is (Significant, 2007). De benchmark dient dan ook bruikbaar te zijn voor ten minste acht groepen inburgeraars (d.w.z. hoog opgeleide nieuwkomers, hoog opgeleide oudkomers, et cetera). Aan de hand van de uitkomsten van het benchmarkonderzoek kan worden bepaald of er met één benchmark voor allen kan worden volstaan dan wel dat er voor afzonderlijke groepen afzonderlijke benchmarks geconstrueerd moeten worden.

Landelijk representatief. De benchmark NT2 moet de leerlast beschrijven voor de gemiddelde inburgeraar (per subgroep) gegeven gemiddeld NT2-onderwijs. Dit betekent dat de steekproef waarop de benchmark ontwikkeld wordt landelijk representatief moet zijn.

Betrouwbaar. De benchmark NT2 moet de leerlast voor het bereiken van een passend taalniveau nauwkeurig beschrijven (dit wil zeggen met een zo klein mogelijke standaardfout/betrouwbaarheidsinterval). Daartoe moet de steekproef voldoende groot zijn. De noodzakelijke omvang van de steekproef is onder meer afhankelijk van het aantal groepen waarvan de benchmark de leerlast moet beschrijven en het aantal taalniveaus i.c. niveaustappen dat voor deze groepen relevant is (zie ook Bijlage 2). Daarnaast moeten de gemiddelde leerlast en taalniveau en hun varianties betrouwbaar worden gemeten. Meetfouten verlagen immers de correlatie tussen leerlast en rendement en hiermee de voorspelbaarheid van de tijd/lesuren die nodig zijn om een passend eindniveau te bereiken.

Valide. De instrumenten om leerlast en rendement te meten moeten onder meer voldoen aan eisen van inhoudsvaliditeit, begripsvaliditeit en predictieve validiteit. Wat betreft de laatstgenoemde vorm van validiteit moet de benchmark NT2 de daadwerkelijk benodigde tijd/lesuren ook buiten de onderzoekssituatie kunnen beschrijven. Het moet een goede beschrijving geven van de daadwerkelijk gebruikte tijd/lesuren in de dagelijkse inburgeringspraktijk.

Efficiënt. De ontwikkeling van een benchmark NT2 mag niet onnodig tijdrovend, arbeidsintensief of kostbaar zijn.

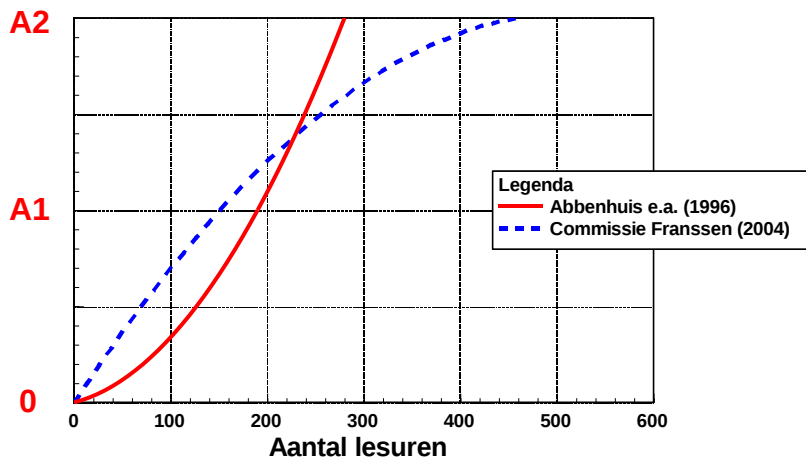
Hoeveel metingen van leerlast en taalniveau zijn minimaal noodzakelijk?

Er zijn ten minste twee theoretische modellen ter beschrijving van de relatie tussen leerlast en taalniveau, te weten dat van lineaire groei (ook wel het *incremental model* genoemd) en van non-lineaire groei (waarbij het *diminishing returns model* ofwel het *learning plateau model* veelvuldig genoemd wordt).

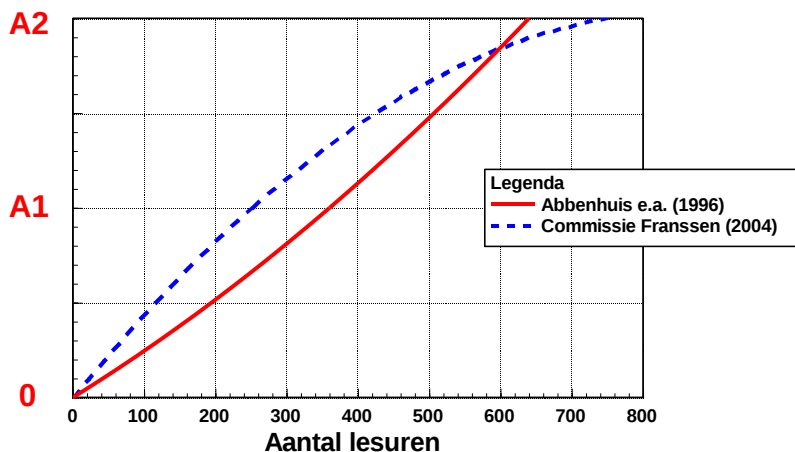
Overeenkomstig het model van lineaire groei is er sprake van een gestage rechtlijnige groei van het taalniveau als een functie van de tijd/lesuren. Het bereiken van het naastgelegen hogere taalniveau kost dan telkens ongeveer eventueel tijd/lesuren als het eraan voorafgaande niveau (bijv. Watt & Lake, 2004).

Volgens het model van non-lineaire groei is voor het bereiken van het naastliggende hogere taalniveau in toenemende mate meer of juist minder tijd/lesuren nodig. De twee hiervoor besproken leerlastindicatoren gaan beide uit van een kromlijning verband tussen leerlast en taalniveau. De aard van de relatie is echter fundamenteel verschillend. Zo kost de stap van niveau 0 naar A1 volgens Abbenhuis e.a. (1996) meer lesuren dan de stap van niveau A1 naar A2 (althans, cursisten die de stap van niveau 0 naar A1 zetten, gebruikten daarvoor meer lesuren dan cursisten die de stap van niveau A1 naar A2 realiseerden). Volgens de

schatting van de Commissie Franssen (2004) kost het bereiken van A1 juist veel minder tijd dan de stap van A1 naar A2 (namelijk precies de helft minder). In Figuur 1 en 2 is te zien hoe de schattingen van de leerlast voor hoog respectievelijk laag opgeleiden uiteenlopen.



Figuur 1
Twee schattingen van de leerlast voor hoog opgeleiden



Figuur 2
Twee schattingen van de leerlast voor laag opgeleiden

In statistische termen komt het ontwikkelen van de benchmark neer op het vaststellen van de empirische relatie tussen leerlast (tijd, aantal lesuren) en taalniveau. Dit wil zeggen: het schatten van de parameters (intercept; regressiegewichten; varianties) van een zo goed mogelijk bij de data passende vergelijking voor de regressie van leerwinst op leerlast. In zijn meest eenvoudige vorm laat de regressievergelijking zien hoe het bereikte taalniveau voor de gemiddelde cursist met een bepaald startniveau en opleidingsniveau verandert als de tijdsduur en het aantal lesuren toenemen. Voor de constructie van de benchmark NT2 is het belang vast te stellen in hoeverre de relevante groepen inburgeraars verschillende groeicurves hebben, dit wil zeggen in hoeverre de intercepten, lineaire, kwadratische en/of kubische termen (en de

bijbehorende varianties tussen leerlingen en scholen) significant van elkaar afwijken.³ Het individuele groeitraject wordt daarbij beschreven met een polynoom waarbij het intercept en de lineaire term respectievelijk het beginniveau en de lineaire toename van de groei beschrijven, de kwadratische term de verandering in de leersnelheid aangeeft en een eventuele kubische term veranderingen in de versnelling of afzwakking van de leersnelheid beschrijft. Als de kwadratische groei een positieve waarde aanneemt, neemt de groeisnelheid na verloop van tijd toe. Is de waarde van de kwadratische term negatief, vlakt de groeisnelheid na verloop van tijd af. Als er slechts twee metingen zijn, kan alleen de lineaire groei worden bepaald. Wil men ook weten of de groei met het toenemen van de tijd/lesuren afzwakt (of toeneemt), zijn ten minste drie meetmomenten noodzakelijk (zodat ook de kwadratische groei kan worden bepaald). Vandaar dat de leerlast en het taalniveau in het leerlastonderzoek waarin de benchmark geconstrueerd wordt ten minste drie keer gemeten moeten worden.

Welke aspecten van leerlast worden ten minste onderzocht?

Aan het begrip leerlast kunnen verschillende onderling samenhangende aspecten worden onderscheiden die al dan niet in leerlastonderzoek gemeten kunnen worden:

- A De programma-uren: het aantal geprogrammeerde uren volgens de lessentabel of het rooster van de onderwijsaanbieder. Hierbij kunnen drie deelaspecten worden onderscheiden:
 - Bruto versus netto programma-uren (d.w.z. exclusief lesuitval door bijvoorbeeld ziekte van de docent);
 - Geprogrammeerde uren op school met de docent (contacturen) versus uren op school zonder de docent (bijv. zelfstudie of oefening school achter de computer);
 - Geprogrammeerde uren binnen versus buiten de school (bijv. stages; buitenschoolse opdrachten).
- B De gerealiseerde cursisturen, d.w.z. het geprogrammeerde aantal uren minus afwezigheid van de cursist. Hierbinnen kan nog een onderscheid worden gemaakt in:
 - Gerealiseerde uren op school versus buiten de school.
- C De duur van de cursus, bijvoorbeeld acht weken of een half jaar.
- D De intensiteit van de cursus, bijvoorbeeld drie of tien lesuren per week.
- E De tijdsduur van de les, bijvoorbeeld een regulier lesuur, verlengde lestijden of dagdelen.
- F Het nagestreefde aantal lesuren zoals afgesproken tussen intaker en inburgeraar tijdens het inburgeringonderzoek
- G De lestijd die gebonden is aan groepsactiviteiten en leertijd die gebonden is aan activiteiten van het individu (Van der Linden, 1996).

In de schriftelijke commentaarronde die volgde op de workshop onderzoek van de Cito-werkconferentie “Op zoek naar het succes van inburgering” bleken de meeste respondenten van mening dat vrijwel alle genoemde aspecten van leerlast zowel relevant als meetbaar zijn.⁴ Uitzonderingen betroffen de tijdsduur van de les (minder relevant), het nagestreefde aantal lesuren zoals afgesproken tussen intaker en inburgeraar tijdens het inburgeringonderzoek (onbetrouwbaar) en het onderscheid tussen de geprogrammeerde uren binnen versus buiten de school (onbetrouwbaar). Naar verwachting is de tijdbesteding binnen de school gemakkelijker vast te stellen dan de tijdbesteding buiten de school (en waarschijnlijk ook meer betrouwbaar

³ Andere functies die mogelijk de moeite waard zijn om te beschouwen zijn exponentieel ($\text{niveau} = 2^{\text{tijd}}$) en logaritmisch ($\text{niveau} = \log(\text{tijd})$).

⁴ Niet alle aspecten van leerlast zijn overigens even makkelijk meetbaar. Voor een beschrijving van enkele onderzoekstechnische problemen die zich bij het meten van leerlast kunnen voordoen wordt verwezen naar Bijlage 2.

en valide). Een uitzondering geldt wellicht de tijdbesteding aan stages die mogelijk wel goed meetbaar is. Mede op basis van de reacties van de deelnemers aan de workshop onderzoek kiezen we er vooralsnog voor om zo mogelijk alle genoemde aspecten van leerlast in de constructie van de benchmark te betrekken.

Welke gegevens over de condities worden ten minste onderzocht?

Welke gegevens over de condities waaronder cursisten inburgeringonderwijs volgen moeten tijdens de ontwikkeling van een benchmark worden verzameld? Een antwoord op deze vraag hangt ten dele af van de gewenste functionaliteit van de benchmark en hiermee van de vraag hoe smal of breed de benchmark behoort te zijn. De benchmark kan voornamelijk beperkt blijven tot een beschrijving van de leercurven van de vier groepen inburgeraars. Volgens deze smalle definitie hoeft er naast de leerlast en de taalniveaus slechts een beperkt aantal cursistgegevens verzameld te worden (hoofdzakelijk opleidingsachtergrond, startniveau van taalvaardigheid, leeftijd, land en taal van herkomst en zo mogelijk ook oud- versus nieuwkomer en het trajectadvies voor de domeinen OGO of Werk).

Er lijkt echter onder NT2-deskundigen, uitvoerders en onderzoekers overeenstemming te bestaan over een bredere opvatting waarbij een benchmark wordt gezien als een instrument dat voorspelt ‘wie’ ‘welk niveau’ in ‘welke tijd’ onder ‘welke condities’ zou kunnen bereiken (Van der Linden, 1996). Volgens deze bredere definitie moet niet alleen de gemiddelde leerlast van de onderscheiden groepen inburgeraars worden beschreven, maar ook de mate waarin deze beschrijving varieert met andere verklarende factoren dan opleidingsniveau en startniveau. Daarbij kan gedacht worden aan aanvullende achtergrondkenmerken van de cursisten, de faciliterende en beperkende randvoorwaarden waarmee zij te maken hebben (bijv. kinderopvang), de NT2-docenten, het gegeven NT2-onderwijs (inhoud, organisatie, vormgeving en didactiek), de onderwijsaanbieder et cetera. Voor de constructie van de benchmark betekent dit dat de onderzoekers aanvullende informatie moeten verzamelen over (de vele) andere verklarende factoren die van invloed zijn op leerlast en taalniveau dan met name leerlast, opleidingsachtergrond en startniveau.

Evidente voordelen van een brede benchmark zijn de grotere nauwkeurigheid en bruikbaarheid van de beschrijvingen. Omdat meer predictoren in de constructie van de benchmark betrokken worden, kunnen meer gedifferentieerde beschrijvingen worden gegeven (bijvoorbeeld niet alleen per opleidingsniveau per startniveau, maar ook voor oud- en nieuwkomers). Bovendien stijgt de precisie van de beschrijving als deze op meer invloedrijke predictoren gebaseerd wordt.

De ontwikkeling van een brede benchmark is echter arbeidsintensiever, tijdrovender, duurder en belastender voor het onderwijs dan een smalle benchmark. Het verzamelen van de aanvullende gegevens over de condities gaat doorgaans gepaard met grote onderzoekstechnische problemen (vgl. Abbenhuis e.a., 1996; Duijm, 1996; Van der Linden, 1996). In de bestudeerde secundaire analyses lijken de voor de constructie van een brede benchmark benodigde (andere) kenmerken van inburgeraars, hun docenten, hun opleidingen, het gegeven NT2-onderwijs (inhoud, organisatie, vormgeving en didactiek) en de faciliterende en beperkende randvoorwaarden nagenoeg te ontbreken. Het is mogelijk dat deze aanvullende gegevens in de een of andere vorm in bestanden van de gemeenten, intakers of instellingen voorhanden zijn. De verwachting is echter dat gemakkelijk toegankelijke bronnen, bestanden en rapportages weinig bruikbaar zijn, terwijl het verkrijgen van wel bruikbare gegevens tijdrovend, arbeidsintensief en kostbaar is en er tevens vraagtekens te plaatsen zijn bij de actualiteit, volledigheid, nauwkeurigheid en validiteit ervan (zie ook Bijlage 2). Daarnaast moeten onderzoekers tegenwoordig bij analyse van individuele cursistgegevens ook rekening houden met de beperkingen van de Wet Bescherming Persoonsgegevens (waarbij cursisten om toestemming moet worden gevraagd).

Op basis van een afweging van de voor- en nadelen stellen wij na raadpleging van de begeleidingscommissie voor de onderzoeksinspanningen te concentreren op het maximaliseren van de volledigheid, betrouwbaarheid en validiteit van de meting van de leerlast (tijd/lesuren) en het taalniveau van de cursisten en bij het verzamelen van aanvullende kenmerken van de cursist, de docent, het onderwijs en de instelling een zo groot mogelijke terughoudendheid te betrachten.

Taalvaardigheidmeting: nieuwe of oude taaltoetsen?

Zoals uiteengezet in Bijlage 2 blijkt een groot deel van de thans beschikbare taaltoetsen niet of slechts beperkt bruikbaar voor leerlastonderzoek. Het instrumentarium voor het meten van de groei in taalvaardigheid moet aan verschillende eisen voldoen:

- alle vier vaardigheden meten (omdat cursisten kunnen verschillen in de leerwinst voor de vier vaardigheden en de effectiviteit van pilots of interventies per vaardigheid verschillend kan zijn);
- hoge toetstechnische kwaliteit (o.a. nauwkeurig, betrouwbaar en valide, ook voor analfabeten);
- meetbereik bij voorkeur tot C1, maar minimaal B2 (gerelateerd aan de niveaus van het CEFR);
- leerwinst meten en rapporteren in ruwe scorepunten en in hele niveaustappen (zodat de vooruitgang ook zichtbaar is als de cursist geen volledige niveaustap maakt);
- voldoende sensitief om de groei van de taalvaardigheid van de cursisten ook in het eerste onderzoeksjaar vast te stellen;
- geschikt voor afname door getrainde proefleiders onder gestandaardiseerde omstandigheden;
- beschikbaar in verschillende parallelle toetsversies;
- qua moeilijkheid afgestemd op het vaardigheidsniveau van de cursist (ook afgestemd op verschillen in vaardigheidsniveau van cursisten binnen de klas);
- niet bekend zijn bij de cursisten (geheimhouding!).

Vooralsnog wordt voorgesteld voor de taalvaardigheidmeting gebruik te maken van de Toolkit Onderwijs en Arbeidsmarkt (TOA) van Bureau ICE (Bureau ICE, 2008). Voor de meting van het beginniveau tijdens de intake bevat deze toolkit onder meer behalve een Voorschatting taalvaardigheid (een cloze-toets) voor elk van de vier vaardigheden toetsen voor de gemengde niveaus A1/A2, A2/B1, B1/B2 en B2/C1. Voor de meting van de tussen- en eindniveaus zijn er voor niveaus A1 tot en met C1 eveneens toetsen beschikbaar, met voor de niveaus A1 en A2 zelfs verschillende varianten voor de profielen OGO en Werk. De inhoud van de TOA is gerelateerd aan de beschrijvingen in het CEFR. De teksten en taken zijn beoordeeld door experts op het gebied van het CEFR en pas opgenomen als de experts overeenstemden in hun beoordeling van het niveau. Ongeveer de helft van de opgaven van de TOA is geselecteerd op basis van gangbare psychometrische eisen; de andere helft betreft volledig nieuwe opgaven waarvan de meetkwaliteit nog niet is vastgesteld. Het aantal opgaven per toets is of dertig (intake) of twintig (tussen- en eindniveaus). Bij alle toetsen die één CEFR-niveau meten, moet de kandidaat minstens 80% van de opgaven goed maken om te bewijzen dat hij of zij het betreffende niveau beheerst (dit wil zeggen 16 van de 20 opgaven). Bij alle toetsen die twee CEFR-niveaus meten, moet de kandidaat zowel 80% op het lagere niveau beheersen als 50% op het hogere niveau; hij beheerst het hogere niveau als hij zowel 80% van dat hogere niveau correct maakt als 90% van het lagere niveau. Bij een toetslengte van 30 items betekent dit dat het lagere niveau wordt beheerst vanaf 20 punten en het hogere

niveau vanaf 26 punten. De handleiding bij de TOA bevat geen wetenschappelijke verantwoording. Er lijkt nog geen wetenschappelijk onderzoek te zijn gedaan naar de betrouwbaarheid, validiteit en normering van de toetsen voor inburgeraars. Alvorens een definitieve keuze voor toetsen uit de TOA te maken, adviseren wij eerst een onderzoek naar de betrouwbaarheid, schaalbaarheid, validiteit en normering uit te voeren (waarbij geprobeerd zal worden de verschillende toetsen te equivaleren). Ter vaststelling van de concurrente en discriminante validiteit kan daarbij ook de relatie met de toetsen van het inburgeringsexamen worden bepaald. Het inburgeringsexamen is immers de ultieme maatstaf waartegen de leerwinst van cursisten wordt afgezet.

Mocht de TOA in zijn huidige vorm onvoldoende bruikbaar, betrouwbaar of valide blijken voor doeleinden van leerlastonderzoek, dient aanvullende toetsconstructie in overweging te worden genomen. Mogelijk kan daarbij ook gebruik worden gemaakt van materiaal uit de Profieltoets NT2, met als nadelen dat het meetbereik van deze toets niet verder reikt dan B1 (waardoor de toets minder geschikt is voor hoog opgeleiden) en dat de opgaven na 2008 niet meer ververst worden.

De TOA-toets en de Profieltoets NT2 zijn niet geschikt voor analfabeten. Geadviseerd wordt om voor analfabeten de begin- en voortgangstoetsen Alfa te gebruiken. De Alfatoets is nog niet direct gelinkt aan de TOA- of Profieltoets. Om de leercurven van de analfabeten te kunnen vergelijken met die van de laag opgeleiden (en de beide andere groepen inburgeraars), adviseren wij begin 2009 een kleinschalig equivaleringonderzoek uit te voeren. Er bestaat overigens wel een kans dat de beide groepen niet op één schaal gebracht kunnen worden. Enerzijds omdat bij analfabeten deels andere vaardigheden een rol spelen en deze groep ook qua samenstelling verschilt van de laag opgeleiden en de beide andere groepen cursisten. Anderzijds omdat de afnamecondities van de beide toetsen niet gelijk zijn (de afnamecondities van de TOA sluiten minder goed aan bij alfaleerders).⁵

Een probleem is dat de Alfatoetsen alleen lees- en schrijfboekjes bevatten. Er moet dus nog een oplossing gezocht worden voor de toetsing van gesprekken/spreken (dat kan waarschijnlijk wel met TOA) en luisteren (dat kan waarschijnlijk niet met TOA).

Tijdens de schriftelijke commentaaronde die volgde op de Cito-werkconferentie “Op zoek naar het succes van inburgering” waarschuwden twee NT2-onderzoekers voor het gebruik van de Profieltoets NT2. Deze toets is ontwikkeld binnen het oude referentiekader NT2, hetgeen volgens een van de onderzoekers onherroepelijk vergelijkbaarheidsproblemen oproept.

Secundaire of primaire gegevensverzameling?

Een benchmark NT2 kan op verschillende manieren worden ontwikkeld die ieder hun eigen mogelijkheden en beperkingen hebben: secundaire analyse van bestaande bestanden of rapportages en primaire gegevensverzameling via empirisch onderzoek. Bij secundaire analyse heranalyseren onderzoekers door anderen verzamelde bestanden of rapportages en bij primaire gegevens verzamelen de onderzoekers de gegevens zelf. Overeenkomstig het Programma van Eisen in de offerteaanvraag van het leerlastonderzoek (kenmerk IKCnr38701) hebben wij een haalbaarheidsstudie uitgevoerd naar de beschikbaarheid en bruikbaarheid van bestaande rapportages en databestanden voor constructie van een benchmark NT2. In Bijlage 2 is uitgebreid verslag gedaan van de kwaliteit en bruikbaarheid van onder meer de rapportages die gemeenten jaarlijks aan de overheid verstrekken in het kader van de financiële verantwoording (zie paragraaf 2.1 van Bijlage 2), de cliëntsystemen waarover gemeenten, intakers en onderwijsaanbieders beschikken (zie paragraaf 2.2. van Bijlage 2) en van de onderzoekstechnische problemen van primaire gegevensverzameling (zie paragraaf 2.3 van

⁵ De alfaleerder zou ook beschouwd kunnen worden als een afzonderlijke categorie die pas vergeleken wordt met de overige groepen cursisten op het moment dat ze hebben leren lezen en schrijven.

Bijlage 2). De voor- en nadelen van primaire en secundaire gegevensverzameling tegen elkaar afwegend kan een duidelijke keuze voor het verzamelen van primaire gegevens worden gemaakt. In het plan van aanpak voor de constructie van een benchmark is er dan ook vanuit gegaan dat de onderzoeker de gegevens zelf verzamelt (zie ook hoofdstuk 3). De keuze voor primaire gegevensverzameling kon rekenen op instemming van zowel de begeleidingscommissie als de deelnemers aan de workshop onderzoek.

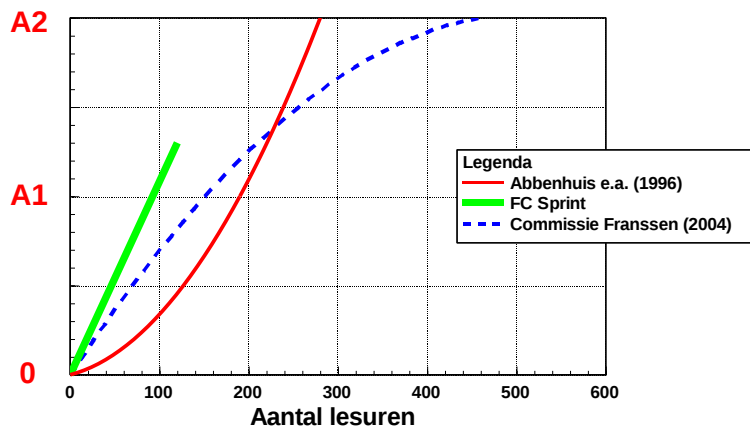
Evaluatie van het effect van pilots met controlegroepen of een benchmark?

Naast het inrichten van innovatietrajecten en experimenten is het ministerie van VROM voornemens leerlastonderzoek uit te zetten naar de benodigde duur van inburgeringsprogramma's voor verschillende groepen inburgeraars (Ministerie van VROM, 2007). Daarnaast wil het Ministerie praktijkgericht onderzoek stimuleren naar het effect van potentiële kenmerken van effectief NT2-onderwijs op de leerlast en het taalniveau (zie hoofdstuk 2). In de uitvoeringsfase van het leerlastonderzoek zal van de belangrijkste succesfactoren uit het theoretisch kader het effect op leerlast en taalniveau worden vastgesteld. Daartoe worden naar verwachting vele pilots uitgevoerd waarin docenten en onderzoekers samenwerken aan het realiseren van NT2-onderwijs overeenkomstig de succesfactoren uit het theoretisch kader. Een van de criteria voor het bepalen van de effectiviteit van een pilot is de leerwinst die cursisten die aan de innovatie deelnemen in een bepaalde periode realiseren. In ontwikkelingsgericht onderzoek kan de leerwinst van de cursisten door middel van een kleinschalige en relatief goedkope pilots worden vastgesteld, maar ook minder harde onderzoeksontwerpen kunnen in deze situatie bruikbare informatie leveren. Webbink en Van der Steeg (2008) zijn in dit verband van mening dat pilots te weinig worden benut voor het vergaren van evidence-based kennis over wat werkt en wat niet werkt in het onderwijs. Hun advies is om aan toekomstige pilots een controlegroep toe te voegen. Vóór de start van een pilot zouden scholen geworven kunnen worden voor deelname aan de pilot. Vervolgens zouden geïnteresseerde scholen via loting aan de pilot- of controlegroep kunnen worden toegewezen. Hiermee wordt een eerste stap gezet van een pilot naar een gecontroleerd experiment. Behalve voordelen heeft het voorstel van Webbink en Van der Steeg (2008) echter ook enkele innovatiestrategische nadelen:

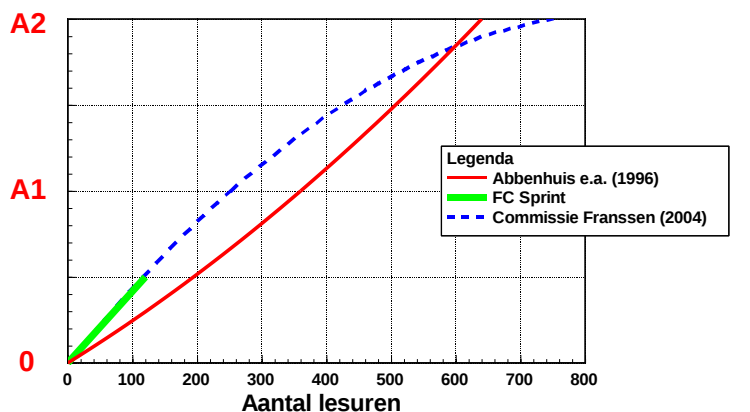
- Random toewijzing van in innovatie geïnteresseerde docenten/scholen aan een pilot- en een controlegroep, zoals Webbink en Van der Steeg (2008) voorstellen, betekent dat de onderzoeker vernieuwingsgezinde docenten en onderwijsaanbieders zal moeten teleurstellen. Velen vinden dit om begrijpelijke redenen niet acceptabel.
- In een situatie waarin er vele pilots naar potentiële succesfactoren worden uitgezet is het voorzien van elke pilot van een controlegroep (onnodig) arbeidsintensief en kostbaar.
- In ontwikkelingsgericht onderzoek gaat het doorgaans om kleine steekproeven. Mocht de pilot onverhoopt ineffectief blijken, is het aantal gedupeerde cursisten klein. De “stakes” zijn veel minder hoog dan bij bijvoorbeeld een grootschalig veldexperiment (zie hoofdstuk 5) waar er bijvoorbeeld een beslissing genomen moet worden over het al dan niet landelijk invoeren van een interventie met grote logistieke en financiële consequenties. In praktijkgerichte pilots naar het verbeteren van innovaties is het met andere woorden wat minder “erg” wanneer de vergelijking met de benchmark in een aantal gevallen tot een fout-positieve of fout-negatieve conclusie zou leiden.

Een voorbeeld van het gebruik van een benchmark voor onderzoeksdoeleinden is gepresenteerd op de Cito-werkconferentie “Op zoek naar het succes van inburgering” (Kuhlemeier, 2008). In het kader van het project FC Sprint² volgde een heterogeen samengestelde groep van 140 beginners een traject van acht weken met in totaal 120 contacturen. In deze 120 uur boekten laag, midden en hoog opgeleiden een leerwinst van

respectievelijk .5, 1.1 en 1.3 niveausprong op de taalvaardigheidladder van het CEFR (Deutekom, 2007, 2008). Figuur 3 en 4 laten zien hoe de leerwinst in de groepen hoog en laag opgeleiden zich verhoudt tot de leerlastindicatoren van Abbenhuis e.a. (1996) en de Commissie Franssen (2004).



Figuur 3
Vergelijking van de leerwinst van de hoog opgeleide cursisten van FC Sprint² met de beide benchmarks



Figuur 4
Vergelijking van de leerwinst van de laag opgeleide cursisten van FC Sprint² met de beide benchmarks

Met de nodige slagen om de arm zou men hieruit voorzichtig kunnen concluderen dat de aanpak van FC Sprint² effectief is voor beginnende hoog opgeleiden (zie Figuur 3), terwijl de leerwinst in de groep laag opgeleide beginners gemiddeld is of daar iets boven zit, afhankelijk van welke leerlastindicator men als criterium zou hanteren (zie Figuur 4).

In de situatie waarin vele kleinschalige ontwikkelingsgerichte pilots tegelijkertijd uitgevoerd worden kan een benchmark een efficiënt alternatief zijn voor onafhankelijke controlegroepen. De gemiddelde leerwinst in de benchmarkgroep vormt dan het criterium waartegen de leerwinst in de pilots kan worden afgezet. In statistisch opzicht dient te worden nagegaan in hoeverre de op basis van de benchmark voorspelde leerwinst significant afwijkt van de schatting van de daadwerkelijke leerwinst in de pilotgroep. Volgens een eenvoudige vuistregel is het verschil tussen beide parameterschattingen statistisch significant als de beide betrouwbaarheidsintervallen elkaar niet overlappen. Uiteraard moet daarbij ook de absolute

grootte en praktische relevantie van het verschil in aanmerking worden genomen zodat dit bijvoorbeeld zichtbaar wordt in de effectgrootte (Cohen, 1969). Een benchmark alleen kan geen antwoord geven op de vraag of een eventueel verschil tussen de leerwinst in de benchmarkgroep en de pilotgroep aan de innovatie dan wel aan andere factoren kan worden toegeschreven. Daartoe moet aan een groot aantal voorwaarden worden voldaan. Voor een verhandeling over de onderzoekstechnische mogelijkheden, beperkingen en valkuilen van een benchmark bij het evalueren van ontwikkelingsgerichte pilots wordt verwezen naar Bijlage 3.

4.3 Plan van aanpak voor het ontwikkelen van een benchmark NT2

Op basis van de beslissingen die aan het ontwerp van een benchmark ten grondslag liggen (zie paragraaf 4.2) schetst deze paragraaf de contouren van ontwikkelingsgericht leerlastonderzoek.

Doel

De benchmark NT2 heeft tot doel een nauwkeurige beschrijving te geven van de leertrajecten van cursisten die op weg zijn naar het inburgeringexamen of het Staatsexamen NT2 I of II. De benchmark moet bruikbaar zijn voor het evalueren van het succes van ontwikkelingsgerichte pilots naar het effect van potentiële succesfactoren uit het theoretisch kader op de leerlast en het taalniveau van groepen inburgeraars.

Populatie en steekproef

Populatie

De populatie voor het ontwikkelen van de benchmark omvat alle inburgeraars die vanaf een nader te bepalen datum zijn begonnen met inburgeringonderwijs. Degenen die zich geheel zelfstandig op het inburgeringexamen voorbereiden, behoren vooralsnog niet tot de onderzoekspopulatie.

Steekproefkader

Het steekproefkader omvat alle instellingen die NT2-onderwijs aan de populatie aanbieden (dus niet alleen de ROC's).

Steekproef

De steekproef moet het mogelijk maken om landelijk representatieve en nauwkeurige gegevens te verzamelen over de leerlast, het bereikte taalniveau en de relatie tussen beide. Het opleidingsniveau en het startniveau zijn de belangrijkste stratificatievariabelen. Er zijn vier opleidingsniveaus: analfabeten, laag, midden en hoog opgeleiden. De taalniveaus zijn de CEFR-niveaus 0, A1, A2, B1 en B2. De startniveaus verschillen zowel tussen als binnen de onderscheiden groepen en overlappen elkaar gedeeltelijk. Een mogelijke invulling van het steekproefontwerp ziet er als volgt uit:

Opleidingsniveau	Start-, tussen- en niveaus				
Analfabeten	0	A1	A2		
Laag opgeleiden	0	A1	A2		
Midden opgeleiden	0	A1	A2	B1	
Hoog opgeleiden	0	A1	A2	B1	B2

Per gevulde cel dient een landelijk representatieve groep cursisten van voldoende omvang beschikbaar te zijn om de leercurven van de subgroepen betrouwbaar te kunnen schatten. In de uitvoeringsfase verdient het verkrijgen van voldoende cursisten in de groep analfabeten met A2 als streefniveau bijzondere aandacht. Overwogen kan worden om de variabele oudkomer versus nieuwkomer als extra stratificatievariabele in het ontwerp onder te brengen.

De steekproef kent ten minste twee fasen. In de eerste fase wordt per opleidingsniveau een steekproef van instellingen getrokken met een trekkingskans omgekeerd evenredig aan het aantal cursisten in het desbetreffende opleidingsniveau van de desbetreffende instelling.⁶ In de tweede fase wordt binnen ieder opleidingsniveau van een instelling per startniveau een vast aantal cursisten at random getrokken.⁷ De resulterende steekproef is binnen elke combinatie van opleidingsniveau en startniveau zelfwegend (d.w.z. een representatieve afspiegeling van de desbetreffende deelpopulatie), maar de steekproef als geheel is dat niet. Als generalisatie naar de totale steekproef wenselijk is, kan eenvoudig worden teruggewogen naar een representatieve afspiegeling van de totale populatie (per opleidingsniveau). Vanwege het *design effect* verdient het de voorkeur om het aantal instellingen te maximaliseren ten opzichte van het aantal cursisten per instelling. Maximalisering van het aantal instellingen vergroot ook de kans dat voor elke cel van het steekproefontwerp voldoende cursisten kunnen worden verkregen.

Omvang van de steekproef

In de voorbereidingsfase van het steekproefontwerp moet worden nagegaan hoeveel individuen en/of groepen voor de constructie van de benchmark minimaal noodzakelijk zijn. Hiertoe dient een power analyse te worden uitgevoerd (Raudenbush & Liu, 2000, 2001; Schochet, 2005). Daartoe dient men meer te weten over de te verwachten effectgrootte (van het verband tussen leerlast en taalniveau), de grootte van de intraclasscorrelatie en het aantal subgroepen en startniveaus waarvoor leercurven geschat moeten worden. Vooralsnog wordt ervan uitgegaan dat een gerealiseerde steekproef van circa 70 cursisten per cel toereikend zal zijn om de leercurven en hun variantie betrouwbaar te schatten (d.w.z. 1200 cursisten). Rekening houdend met een uitval van maximaal 50%, zouden dan 2400 cursisten getrokken moeten worden (140 per cel).

Onderzoeksontwerp

Cohort versus cross-sectioneel

Voor de opzet van het onderzoek waarin de benchmark ontwikkeld wordt, komen grofweg twee designs in aanmerking:

- een cohortstudie waarbij een landelijk representatieve groep inburgeraars vanaf het moment dat zij met NT2-onderwijs beginnen (bijvoorbeeld vanaf 1 januari 2009) gedurende een bepaalde periode gevolgd wordt;
- een cross-sectionele studie waarbij een landelijk representatieve steekproef van inburgeraars die op dat moment NT2-onderwijs volgen eenmalig gemeten wordt.

Cohortstudies zijn doorgaans het meest geschikt om veranderingen in de tijd vast te stellen. Andere voordelen van de herhaalde meting bij dezelfde individuen zijn onder meer de hogere

⁶ Dit veronderstelt dat er een naar opleidingsniveau geordend steekproefkader voorhanden is met informatie over het aantal cursisten per opleidingsniveau per instelling. Is dat er niet, dan moet dat ten behoeve van het onderzoek geconstrueerd worden.

⁷ Dit veronderstelt dat van ieder opleidingsniveau binnen een instelling de verdeling van de cursisten naar startniveau bekend is.

statistische power om effecten aan te tonen en de betere interpreteerbaarheid van de gegevens. Een nadeel is de vaak grote uitval van cursisten en *testing artifacts* ten gevolge van herhaalde meting bij dezelfde individuen met hetzelfde instrument (tenzij met geëquivalenteerde toetsversies gewerkt wordt). Een ander nadeel van een cohortstudie is dat het relatief lang duurt voordat de resultaten beschikbaar komen (hetgeen, zoals studies van bijvoorbeeld Condelli e.a. (2007) en Watt & Lake (2004) aantonen, niet wil zeggen dat rapportage van tussentijdse gegevens na bijvoorbeeld één jaar geen beleidsrelevante informatie kan opleveren).

Een voordeel van een cross-sectionele studie is de onmiddellijke beschikbaarheid van gegevens van zowel gevorderde als beginnende cursisten. Een van de nadelen is echter dat de leercurven van de onderscheiden groepen alleen geschat kunnen worden op basis van verschillen tussen cursisten.

De voors en tegens tegen elkaar afwegend adviseren wij een cohortstudie met meer meetmomenten (in plaats van een cross-sectionele studie die slechts één meetmoment kent).⁸

Regelmatig verversen van het cohort

Zoals uiteengezet in paragraaf 3.2 van Bijlage 3, zijn er onderzoekstechnische argumenten om een cohort regelmatig te verversen, bijvoorbeeld door elk jaar een nieuw cohort van beginnende cursisten in de tijd te gaan volgen. Een praktische reden is dat inburgeraars weliswaar 3,5 jaar de tijd hebben om hun examen te halen, maar een groot deel van hen er in de praktijk veel minder lang over doet, waardoor er na verloop van tijd nog maar weinig cursisten in het cohort resteren. Andere argumenten die pleiten voor regelmatige verversing zijn de veranderingen in de samenstelling van de instroom (van nieuwkomers), veranderingen in het onderwijs (bijvoorbeeld marktwerking) en het verschijnsel van de continue instroom van inburgeraars in het onderwijs.

Hoe lang moet het onderzoek duren?

Hoe lang moeten cursisten in het cohortgedeelte van het onderzoek vanaf het moment dat zij inburgeringonderwijs ontvangen gevolgd worden? Overeenkomstig de huidige Wet Inburgering (WI) moeten inburgeringsplichtigen die het basisexamen inburgering in het buitenland hebben afgelegd het inburgeringexamen binnen een termijn van drieëneenhalf jaar halen, terwijl er alle anderen dat binnen vijf jaar moeten doen. Naar verwachting zal er bij voorspoedige parlementsbepaling vanaf 1 januari 2009 sprake zijn van een uniforme termijn van drieëneenhalf jaar (Ministerie van VROM, 2007). Vooralsnog stellen wij dan ook een totale looptijd voor van drieëneenhalf jaar. De langere doorlooptijd met meer metingen per cursist maakt het tevens mogelijk om stabielere voorspellingen over leerlast en rendement te doen.

Wanneer kan het onderzoek starten?

Een looptijd van drieëneenhalf jaar betekent niet dat met het gebruik van benchmarkgegevens voor onderzoeksdoeleinden gewacht moet worden tot alle cursisten in het cohort hun volledige inburgeringstraject doorlopen hebben. Ook na één jaar kan al beleidsrelevante informatie worden aangeleverd (vgl. Watt & Lake, 2004; Condelli e.a., 2007). Vooralsnog gaan we er vanuit dat de eerste pilots naar het effect van de succesfactoren en het onderzoek naar de constructie van de benchmark medio 2009 van start gaan. De eerste cursisten zullen hun onderwijstraject dan begin 2010 afronden. Om eerste resultaten van de pilots te kunnen

⁸ Een combinatie van beide designs kan wellicht voordelen bieden als er in het begin van de cohortstudie nog maar weinig gegevens voor analyse beschikbaar zouden zijn (waarbij het aantal cursisten in het cross-sectionele gedeelte met het stijgen van het aantal cursisten in het cohortgedeelte geleidelijk kan afnemen).

evalueren, zouden de eerste resultaten van het benchmarkonderzoek medio 2010 voor onderzoeksdoeleinden beschikbaar kunnen komen.

Hoe vaak en met welke afstanden moet er gemeten worden?

Hoe vaak moeten de vorderingen worden vastgesteld? Om de leerwinst te kunnen vaststellen als een functie van tijd en lesuren moeten de vorderingen van de cursisten op meer momenten worden gemeten. In het onderzoek van Watt en Lake (2004) is de leerwinst na telkens ongeveer 250 lesuren vastgesteld. In het onderzoek van Condelli e.a. (2007) is ongeveer een driemaandelijke cyclus gehanteerd. Zoals beargumenteerd in paragraaf 4.2, worden de vorderingen van de cursisten ten minste drie keer gemeten. De tijd tussen de metingen mag niet te kort zijn om leerwinst te kunnen aantonen, maar ook weer niet te lang omdat uitval dan zijn tol kan eisen. Voorgesteld wordt de leerlast en de ontwikkeling van de taalvaardigheid in het eerste onderzoeksjaar op drie momenten vast te stellen, te weten direct aan het begin, halverwege en tegen het einde van dat jaar. In het tweede en volgende onderzoeksjaar kan wellicht worden volstaan met halfjaarlijkse cyclus.

Instrumentatie

Dat meer tijd en lesuren voor NT2-onderwijs tot hogere taalprestaties leiden, lijkt een open deur. Wetenschappelijk onderzoek heeft echter de grootste moeite om dit verband empirisch aan te tonen. Voor laag opgeleide volwassenen concludeerden Cohen, Condelli en Garet (1996) dat het veronderstelde verband tussen leerlast en taalprestaties nog in geen enkele studie was aangetoond. Enkele jaren later toonden Condelli e.a. (2007) in een zorgvuldig opgezet en uitgevoerd onderzoek voor een groep zeer laag opgeleide cursisten een significante relatie tussen leerlast en taalniveau aan. De samenhang tussen leerlast en leersnelheid bleek echter zwak. Als een invalide taaltoets buiten beschouwing gelaten wordt, varieerde de correlatie tussen het totaal aantal uren instructie en het taalniveau afhankelijk van de taaltoets van $-.01$ tot $.16$. De correlatie met de proportie daadwerkelijk gevolgde uren instructie varieerde van $-.02$ tot $.21$. Kortom, onderzoek naar de relatie tussen leerlast en rendement luistert nauw en stelt hoge eisen aan het onderzoeksontwerp, de meetinstrumenten, de uitvoering van het onderzoek en de statistische analyse.

Leerlast

Ten aanzien van de leerlast wordt getracht nagenoeg alle relevante aspecten van leerlast vast te stellen (zie ook paragraaf 4.2).

Taaltoetsing

Zoals beargumenteerd is in paragraaf 4.2, adviseren wij voor de taalvaardigheidmeting gebruik te maken van de TOA-toetsen van Bureau ICE. Mochten dit niet mogelijk of wenselijk blijken, kan de Profieltoets als alternatief dienen, met als nadeel dat het meetbereik van deze toets niet verder reikt dan B1 (waardoor de toets minder geschikt is voor hoog opgeleiden) en de toets vanaf 2009 niet meer ververst wordt. De TOA-toets en de Profieltoets NT2 zijn niet geschikt voor analfabeten. Voorgesteld wordt om voor analfabeten de begin- en voortgangstoetsen Alfa te gebruiken.

Statistische analyse

In statistische termen komt het ontwikkelen van de benchmark neer op het vaststellen van de empirische relatie tussen leerlast (tijd, aantal lesuren) en taalniveau. Dit wil zeggen: het schatten van de parameters – intercept, regressiegewichten en varianties - van een zo goed mogelijk bij de data passende vergelijking voor de regressie van leerwinst op leerlast. De groeitrajecten worden daarbij beschreven met een polynoom waarbij het intercept en de

lineaire term respectievelijk het beginniveau en de lineaire toename van de groei beschrijven en de kwadratische term de verandering in de leersnelheid aangeeft. Voor de statistische analyse moet een techniek worden gekozen die past bij de aard van de gegevens. Vier kenmerken van de gegevens compliceren de statistische analyse:

- Uitval van cursisten (waardoor de gemiddelden per meetmoment op verschillende cursisten gebaseerd zijn);
- Ongelijke intervallen tussen de meetmomenten (het is logistiek immers niet mogelijk om alle cursisten op dezelfde momenten te toetsen);
- Variatie in leerwinst tussen cursisten binnen dezelfde klas (de startniveaus van cursisten binnen een klas verschillen en de ene cursist binnen een klas gaat sneller vooruit dan de andere cursist binnen diezelfde klas);
- De hiërarchische structuur van de gegevens (met meetmomenten genest binnen cursisten genest binnen klassen).

De gebruikelijke analysetechnieken, zoals variantie-analyse van herhaalde metingen, laten het in deze complexe situatie afweten. De leercurven van de cursisten worden daarom gemodelleerd met een meerniveau model voor het meten van leerwinst met drie niveaus: metingen genest binnen cursisten die op hun beurt weer genest zijn binnen klassen of instellingen (Bryk & Raudenbush, 1992; Goldstein, 1992; Van den Bergh & Kuhlemeier, 1997). Deze techniek maakt het mogelijk de groeitrajecten van de cursisten te schatten rekening houdend met a) uitval, b) ongelijke intervallen tussen de meetmomenten, c) verschillen tussen de cursisten in startniveau, de snelheid en richting van de groei, en d) de hiërarchische structuur van de gegevens.

Een deel van de onderwijsaanbieders heeft het onderwijs geflexibiliseerd in de zin dat cursisten de keuze geboden wordt om een zelfde les op verschillende tijdstippen in een andere groep en/of bij een andere docent te volgen. In analysetechnisch opzicht is er dan sprake van kruisclassificatie. Ook voor dit probleem biedt multiniveau analyse een oplossing. Zie Goldstein (1992) voor de modellering van kruisclassificaties en Kuhlemeier, Kleintjes & Van den Bergh (2001) voor een toepassing.

4.4. Enkele aanbevelingen voor benchmarkconstructie

NT2-onderwijs geldt als een uitermate weerbarstig onderzoeksterrein. Tot de belangrijkste problemen waarmee de onderzoeker bij het construeren van een benchmark NT2 te kampen hebben, behoren onder meer (Abbenhuis e.a., 1999; Beder, 1999; Comings, Soricone & Santos, 2006; Comings & Soricone, 2007):

- hoge uitval van respondenten, hoge non-response en inconsistente gegevens;
- onbetrouwbare en invalide meting van duur en intensiteit;
- onbetrouwbare en invalide toetsing van taalvaardigheid (o.a. toetsen gekalibreerd in niveaus; te lage sensitiviteit om leerwinst te meten; te moeilijk of te gemakkelijk; te weinig of te veel tijd tussen metingen);
- hoge diversiteit van de cursisten binnen de klassen/groepen van een instelling;
- instabiliteit van de populatie van jaar tot jaar;
- wisseling van de samenstelling van de groepen cursisten binnen opleidingen;
- meerdere NT2-docenten per groep;
- te korte looptijd van het onderzoek;
- ontoereikend onderzoeksbudget.

Op grond van de bestudeerde literatuur kunnen enkele aanbevelingen voor leerlastonderzoek naar de ontwikkeling van de benchmark NT2 gedaan worden.

Marktconforme beloning

Het NT2-onderwijs is tegenwoordig sterk vercommercialiseerd. Anders dan bij ontwikkelingsgericht leerlastonderzoek het geval is, levert deelname aan het benchmarkonderzoek de instellingen geen praktisch bruikbare informatie op. Een van de middelen om de kwantiteit en kwaliteit van de gegevens te verhogen is het betalen van een marktconforme vergoeding aan de medewerkers van de instellingen die bij het onderzoek betrokken zijn (o.a. Comings & Soricone, 2007).

Harde afspraken met onderwijsaanbieders

Met de onderwijsaanbieders kan een contract worden gesloten met harde afspraken over onder meer de looptijd van het onderzoek, het toewijzen van een toetsruimte aan onderzoekers, het verlenen van toegang van de onderzoekers tot de klaslokalen en de administratie, de precieze uitvoering van de taalmetingen en het verstrekken van betrouwbare aanwezigheidsgegevens van de cursisten op de verschillende meetmomenten.

Beperking van de onderzoeksbelasting voor de instelling

Uitval, non-response en inconsistente gegevens kunnen verder worden bestreden door de belasting voor de instelling tot een minimum te beperken. Eerder onderzoek naar het ontwikkelen van een NT2-benchmark heeft laten zien dat het verzamelen van aanvullende informatie over de cursisten, de docenten, de instructie, de instelling en de randvoorwaarden zeer veel tijd, moeite en soms ook frustratie kost (o.a. Abbenhuis e.a., 1996) terwijl de kwaliteit van deze gegevens vaak sterk te wensen over laat. Dit pleit eerder voor het construeren van een smalle dan voor een brede benchmark (zie paragraaf 4.2). Voorgesteld wordt de inspanningen te concentreren op het maximaliseren van de volledigheid, betrouwbaarheid en validiteit van de meting van de leerlast (tijd/lesuren) en het taalniveau van de cursisten en bij het verzamelen van aanvullende kenmerken van de cursist, de docent, de groep, het onderwijs en de instelling een zo groot mogelijke terughoudendheid te betrachten.

Docent als intermediair tussen de instelling en de onderzoekers

De gegevens over de leerlast en de taalvaardigheid van de cursisten op de verschillende meetmomenten dienen nauwgezet gecheckt te worden op volledigheid en consistentie. Het verzamelen van de gegevens over de leerlast kan gedeeltelijk worden uitbesteed aan een medewerker die aan de onderwijsinstelling verbonden is. Een docent met goede contacten met de administratie verdient daarbij de voorkeur (Condelli e.a., 2007). Deze intermediair op de instelling dient voor zijn of haar inspanningen een marktconforme beloning uit het projectbudget te ontvangen. Deze medewerker ontvangt van de onderzoekers een intensieve training.

Check op de volledigheid en consistentie van de gegevens

Zo snel mogelijk na een meetmoment dient de kwaliteit van de ingestuurde gegevens over de leerlast van de cursisten door de onderzoekers gecheckt te worden op volledigheid en consistentie. Als een extra check zou elke instelling met enige regelmaat door een zogeheten *monitor* bezocht kunnen worden.

Multi-methoden aanpak

De kwaliteit van de gegevens kan verder worden verhoogd door eenzelfde variabele met verschillende onderzoeksmethoden te onderzoeken (Tashakkori & Teddlie, 1998). Een voorbeeld is het vergelijken van via bevraging van docenten verkregen leerlastgegevens met zelfrapportages van cursisten en gegevens uit de schooladministratie.

Afname taaltoetsen onder gestandaardiseerde omstandigheden

De afname van de taaltoetsen kan het beste worden uitgevoerd door speciaal getrainde proefleiders die voor de gegevensverzameling per meetmoment een aantal dagen op de opleiding aanwezig zijn (en daarvoor een passende beloning ontvangen).

Voldoende tijd en budget

De bruikbaarheid van onderzoek naar volwassen tweedetaal leerdere wordt vaak sterk beperkt door de korte looptijd van het project en het ontoereikende budget (o.a. Abbenhuis e.a., 1996; Beder, 1999). Het verdient aanbeveling de onderzoekers voldoende tijd en budget te geven voor de ontwikkeling van een kwalitatief hoogwaardige benchmark NT2.

5 (Quasi-)experimenteel leerlastonderzoek

In hoofdstuk 1 is een tweefasen model geschetst van leerlastonderzoek dat moet leiden tot verbetering van het NT2-onderwijs (Raudenbush, 2003). Betoogd is dat een fasering van ontwikkelingsgericht naar (quasi-)experimenteel leerlastonderzoek het NT2-onderwijs meer in overeenstemming kan brengen met de succesfactoren uit het theoretisch kader. Het ministerie van VROM is voornemens een veelvoud aan kleinschalige pilots te stimuleren naar het effect van de potentiële kenmerken van effectief NT2-onderwijs uit het theoretisch kader. Dit ontwikkelingsgerichte onderzoek moet leiden tot evidence-based kennis over effectief NT2-onderwijs en bijdragen aan de vernieuwing en verbetering van het NT2-onderwijs. Vanwege de grote logistieke en financiële implicaties kunnen in relatief kleinschalige settings ontwikkelde innovaties pas grootschalig worden “uitgerold” in het NT2-onderwijs als de effectiviteit onomstotelijk is vastgesteld. Daartoe lijkt een gerandomiseerd en gecontroleerd veldexperiment het meest geëigende middel (Shadish, Cook & Campbell, 2002; Baron, 2005, 2007; Coalition for evidence-based policy, 2005; Towne & Hilton, 2004; Torgerson e.a., 2004; Onderwijsraad, 2006). In dit hoofdstuk worden de contouren van (quasi-)experimenteel leerlastonderzoek beschreven.

5.1 Inleiding

Een onderzoeksontwerp is gerandomiseerd als de eenheden (d.w.z. cursisten, docenten, klassen of instellingen) volgens een aselekt trekkingsmechanisme aan een interventie- of controlegroep worden toegewezen. Het grote voordeel van random toewijzing is dat de groepen bij de start van de interventie in beginsel volledig identiek zijn op alle factoren, zowel bekende en onbekende factoren als gemeten en niet-gemeten factoren. Verschillen tussen beide groepen op de eindmeting kunnen daardoor relatief eenduidig aan de interventie worden toegeschreven, en niet aan allerlei andere verklarende factoren. Met de term gecontroleerd wordt bedoeld dat de onderzoeker de volledige controle heeft op het aselekt toewijzingsmechanisme. Een belangrijk argument voor een gerandomiseerd en gecontroleerd veldexperiment is de hardheid van de conclusies over het succes van de interventie en de geloofwaardigheid en bruikbaarheid van de onderzoeksresultaten voor onder meer beleidsmakers.

Diverse auteurs hebben aangegeven wanneer een gerandomiseerd en gecontroleerd veldexperiment al dan niet wenselijk of noodzakelijk is (Shadish, Cook & Campbell, 2002; Baron, 2005, 2007; Raudenbush, 2003; Towne & Hilton, 2004; Torgerson e.a., 2004). Een veldexperiment is vooral geïndiceerd als het een ingrijpende en kostbare interventie betreft, als het om logistieke of financiële redenen niet mogelijk is om de innovatie in dezelfde periode aan de hele populatie aan te bieden, als er een onwenselijke situatie bestaat met grote onenigheid en onzekerheid over de effectiviteit van meerdere wenselijk geachte interventies, als contaminatie van interventiegroep en controlegroep kan worden voorkomen, als het effect van de interventie naar verwachting klein is, als de onderzoeker de toewijzing (mede) kan bepalen en als random toewijzing uitvoerbaar en sociaal aanvaardbaar is.

Een veldexperiment is vaak niet mogelijk of niet wenselijk als een snel en/of goedkoop antwoord gewenst is, als een onnauwkeurig antwoord ook goed is, als het effect van de interventie naar verwachting zeer groot is, als de interventie in het verleden ligt (bijv. de moord op J.F. Kennedy), als de interventie al massaal is ingevoerd (en er dus geen vergelijkbare controlegroep voorhanden is), als het aanbieden (bijv. roken) of onthouden (bijv. een beloftevol medicijn tegen kanker) van de interventie aan personen op ethische bezwaren stuit, als er te weinig eenheden aan het experiment kunnen of willen meedoen, als

de interventie niet goed doordacht of uitvoerbaar is en als er nog te weinig theoretisch en empirisch voorwerk is gedaan. Wat dit laatste betreft moet onder meer de bestaande wetenschappelijke literatuur zijn geïnventariseerd op toetsbare hypothesen, er moeten kleinschalige pilots zijn uitgevoerd om de uitvoerbaarheid en aanvaardbaarheid van de interventie te exploreren en er dient al enige empirische evidentie aanwezig te zijn voor het effect van de interventie.

Een veldexperiment hoeft overigens niet altijd lang te duren en veel te kosten. Dat is mede afhankelijk van het type interventie. In een studie van het Cultureel Planbureau noemen Webbink & Van der Steeg (2008) enkele voorbeelden van experimenten die weinig middelen vergden en waarvan de resultaten binnen een jaar beschikbaar waren. Ook aan de ethische bezwaren hoeft volgens genoemde auteurs niet al te zwaar getild te worden. In de medische wetenschap waar er voor de deelnemer aan het onderzoek meestal veel op het spel staat, is experimenteren volledig geaccepteerd. Evenmin is het op voorhand zeker of de interventie inderdaad succes zal hebben (en deelnemers in de controlegroep een voordeel onthouden wordt). Ten slotte wordt het onthouden van de interventie aan een deel van de onderzoeksgroep ruimschoots gecompenseerd door het voordeel dat toekomstige generaties cursisten hiervan kunnen hebben (mits de interventie uiteraard succesvol blijkt).

5.2 Plan van aanpak voor (quasi-)experimenteel leerlastonderzoek

In deze paragraaf worden de contouren van leerlastonderzoek naar de effectiviteit van een ontwikkelde NT2-interventie volgens een (quasi-)experimenteel ontwerp geschetst.

Doel van het onderzoek

Het veldexperiment heeft tot doel zo hard mogelijk bewijs te verzamelen over de effectiviteit van de interventie. Daarnaast kan de differentiële effectiviteit voor verschillende groepen cursisten worden vastgesteld. Ook kan worden nagegaan in hoeverre verschillen in de uitvoering van de interventie van invloed zijn op de resultaten.

Methode van onderzoek

Populatie

De populatie voor het onderzoek naar het effect van de NT2-interventie omvat alle inburgeraars die NT2-onderwijs bij een onderwijsaanbieder volgen.

Steekproef

In de voorbereidingsfase van het steekproefontwerp moet worden nagegaan hoeveel individuen en/of groepen minimaal noodzakelijk zijn om te kunnen vaststellen of de interventie een betekenisvol effect heeft mocht dat effect zich daadwerkelijk voordoen. Hiertoe dient een zogeheten power analyse te worden uitgevoerd (Raudenbush & Liu, 2000, 2001; Schochet, 2005). Daarbij dient men onder meer te weten te komen wat de te verwachten effectgrootte is, hoe groot de intraclasscorrelatie is en hoe groot het effect van de interventie zou moeten zijn om de kosten van de interventie en de grootschalige verspreiding ervan te rechtvaardigen.

Volgens Whitehurst (2003) vereist het kunnen aantonen van een middelmatig effectieve interventie een steekproef van minimaal 300 individuen (gelijk verdeeld over de interventie- of controlegroep). Is de klas of de school de eenheid, zijn daartoe in totaal op zijn minst vijftig à zestig klassen of scholen noodzakelijk. Raudenbush (2003) komt tot vergelijkbare aantallen.

Onderzoeksontwerp

Om het succes te kunnen vaststellen, moet een interventie aan verschillende eisen voldoen. Zo moet de onderzoeker de cursisten lang genoeg volgen en vaak genoeg meten om het veronderstelde verschil in leerwinst tussen de interventiegroep en de controlegroep met een grote mate van waarschijnlijkheid te kunnen vaststellen. Daarnaast moet de interventie zijn afgestemd op het startniveau en opleidingsniveau van de cursisten. Verder moeten de docenten de interventie implementeren overeenkomstig de bedoelingen van de ontwikkelaars (zodat implementatieverschillen binnen programma's in vergelijking tot die tussen de programma's relatief klein zijn).

Indien gekozen wordt voor een gerandomiseerd en gecontroleerd veldexperiment, zijn er in beginsel ten minste drie mogelijkheden om cursisten, docenten en/of opleidingen aan de interventie- en controleconditie toe te wijzen:

- Random toewijzing van *cursisten* aan condities binnen klassen of opleidingen (waarbij als het ware binnen elke deelnemende klas of opleiding een “mini-experiment” plaatsvindt);
- Random toewijzing van *intacte klassen of docenten* aan condities binnen opleidingen (dit wil zeggen: een mini-experiment binnen elke deelnemende opleiding);
- Random toewijzing van *opleidingen* aan condities.

Indien de keuze op een experimenteel ontwerp zou vallen, moet worden besloten of men individuen (d.w.z. cursisten) dan wel groepen (d.w.z. klassen of opleidingen) aan condities toewijst. Random toewijzing van cursisten aan condities is het meest efficiënt en goedkoop, maar kan vanwege de verstoring van de normale gang van zaken op de scholen op weerstand stuiten. Random toewijzing van cursisten is verder alleen zinvol als cross-over of andere vormen van contaminatie tussen de interventie- en controleconditie binnen de klas of school kan worden voorkomen. Contaminatie van condities kan tot een onderschatting van het effect van de interventie leiden.

Een keuze voor randomisatie van groepen in plaats van individuen is geïndiceerd als de onderzoeker wil nagaan in hoeverre het effect van de interventie verschilt voor verschillende docenten, klassen of scholen en als men het effect van de interventie wil onderscheiden van dat van de kenmerken van de docent, de klas en de school (bijvoorbeeld motivatie en kwaliteit van de docenten).

De keuze van het best passende ontwerp hangt onder meer af van het beschikbare budget en de aard van de interventie. Zo lijkt random toewijzing van cursisten aan condities binnen klassen of opleidingen wellicht eerder haalbaar in een veldexperiment naar het effect van ‘inburgeringcoaches’ of ‘faciliteiten in de randvoorwaardelijke sfeer’ dan in een veldexperiment naar het effect van ‘eigen taal’ waar random toewijzing van intacte klassen of opleidingen meer voor de hand ligt.

Bij het opzetten van een veldexperiment moeten doorgaans grote hindernissen worden overwonnen. Een sprekend voorbeeld is het interventieonderzoek naar het effect van *direct literacy instruction* (Cronen, Silver-Pucuilla & Condelli, 2006). Mocht een veldexperiment naar het effect van NT2-interventies niet haalbaar of betaalbaar blijken, kan worden uitgeweken naar een zo sterk mogelijk quasi-experimenteel ontwerp.

Instrumentatie

Het verband tussen leerlast en rendement blijkt in onderzoek vaak moeilijk aantoonbaar; de vastgestelde correlaties tussen tijd/lesuren en de groei van taalvaardigheid zijn op cursistniveau doorgaans zwak. Het aantonen van een zwakke relatie luistert nauw. Onder meer moeten de instrumenten voor het meten van de taalvaardigheid in psychometrisch

opzicht van hoge kwaliteit zijn, geschikt zijn voor afname door getrainde proefleiders onder gestandaardiseerde omstandigheden, beschikbaar zijn in verschillende parallelle toetsversies, qua moeilijkheid zijn afgestemd op verschillen tussen de cursisten in vaardigheidsniveau (ook binnen de groep) en niet bekend zijn bij de cursisten (geheimhouding!).

Statistische analyse

De gegevens van het interventieonderzoek NT2 worden geanalyseerd met een meerniveau model voor het modelleren van leerwinst (Bryk & Raudenbush, 1992; Goldstein, 1992; Van den Bergh & Kuhlemeier, 1997).

Gebruik van de benchmark NT2 in (quasi-)experimenteel leerlastonderzoek

Indien een gerandomiseerd en gecontroleerd veldexperiment goed wordt opgezet en uitgevoerd, lijkt een benchmark NT2 niet noodzakelijk. Als de vergelijking van de leerwinst in de benchmark- en de interventiegroep tot dezelfde uitkomst leidt als die tussen de interventie- en controlegroep verleent de benchmark extra geloofwaardigheid aan de onderzoeksresultaten. Wijzen de uitkomsten in een verschillende richting, zullen onderzoekers en opdrachtgever geneigd zijn de vergelijking met benchmark minder te vertrouwen dan die met de onafhankelijke controlegroep. Naarmate het onderzoek echter een minder sterke opzet kent of minder zorgvuldig is uitgevoerd, wordt het gewicht van de benchmark NT2 belangrijker. Voor een verhandeling over de mogelijkheden, beperkingen en valkuilen van het gebruik van een benchmark in het kader van ‘zachte’ ontwikkelingsgerichte pilots en ‘harde’ (quasi-)veldexperimenten wordt verwezen naar Bijlage 3.

5.3 Enkele aanbevelingen voor (quasi-)experimenteel leerlastonderzoek

Aanvullend ten opzichte van de aanbevelingen voor het opbouwen van partnerschap in ontwikkelingsgericht onderzoek (zie paragraaf 3.2) kunnen op grond van de nog beperkte ervaringen met (quasi-)experimenteel interventieonderzoek in het onderwijs enkele aanbevelingen voor een succesvolle uitvoering worden gedaan (Towne & Hilton, 2004; Torgerson e.a., 2004).

Financiële compensatie voor deelname aan het onderzoek

Het opbouwen van partnerschap met de scholen betekent in de praktijk van veldexperimenten vaak dat de deelnemers aan het onderzoek voor hun inspanningen een passende financiële vergoeding ontvangen. Een andere mogelijkheid tot compensatie voor geleverde prestaties is het gratis ter beschikking stellen van lesmateriaal en scholing aan onderwijsgeevenden. Gemeenten kunnen de deelname aan onderzoek desgewenst opnemen in de bekostigvoorwaarden.

Contract met de onderwijsaanbieder

De onderzoekspraktijk (in de Verenigde Staten) leert dat het vertrek van een directielid soms tevens het einde betekent van de interventie op de school. Het opstellen van een contract met de onderwijsaanbieder kan dit voorkomen.

Voldoende grote steekproeven

De praktijk van gerandomiseerde en gecontroleerde veldexperimenten kent vele voorbeelden van onderzoek waarin het de onderzoekers niet gelukt is om voldoende deelnemers te werven en/of deze gedurende langere tijd voor het onderzoek te behouden. Dit onderstreept het belang van een voldoende grote steekproef en flankerende maatregelen om uitval van

cursisten tegen te gaan. Betaling van cursisten is vaak de beste remedie. Wat ook helpt, is het inhuren en scholen van een intermediair, bij voorkeur een docent afkomstig van de school zelf. De inzet van een intermediair is vooral van belang voor de uitvoering van eventuele random toewijzing van cursisten aan condities en het in stand houden van het verschil tussen de interventie- en controleconditie gedurende de looptijd van het onderzoek.

Tegengaan van contaminatie van condities

Goede docenten in de interventiegroep zullen hun kennis van wat werkt in het onderwijs willen delen met andere docenten. Ook komt het voor dat cursisten of docenten “wisselen van conditie”. De onderzoeker dient maatregelen te nemen teneinde contaminatie van condities zoveel mogelijk te voorkomen.

Interne versus externe validiteit

Het realiseren van een zo hoog mogelijke interne validiteit van het onderzoek vergt veel voorbereiding en overleg met de scholen voordat de interventie kan starten. Te strikte protocollen voor de uitvoering kunnen evenwel weer ten koste gaan van de externe validiteit (d.w.z. de generaliseerbaarheid van de onderzoeksresultaten naar andere populaties, tijdstippen en settings).

Ruime looptijd en budget

De bruikbaarheid van onderzoek naar volwassen tweedetaal leerders wordt vaak sterk beperkt door de korte looptijd van het project en het ontoereikende budget (o.a. Abbenhuis e.a., 1996; Beder, 1999). Het verdient aanbeveling de onderzoekers voldoende tijd en budget te geven. Om een gevonden effect aan de interventie te kunnen toeschrijven, is een gerandomiseerd en gecontroleerd veldexperiment met een voormeting en een interventie- en een controlegroep het meest geschikt (Onderwijsraad, 2006; Webbink & Van der Steeg, 2008). Een veldexperiment is echter niet altijd goedkoop. In de Verenigde Staten variëren de kosten van 50 duizend tot 50 miljoen dollar (U.S. Office of Management and Budget, 2007). Wel zijn er aanwijzingen dat de hogere kosten voor de voorbereiding en uitvoering van een veldexperiment weer worden terugverdiend in de fase van statistische analyse, rapportage en verspreiding van de onderzoeksgegevens (Towne & Hilton, 2004).

Bijlage 1

Mogelijke gebruiksdoelen en –situaties van een benchmark NT2

In het kader van leerlastonderzoek hebben wij geadviseerd de benchmark de eerste jaren alleen voor onderzoeksdoeleinden te gebruiken (zie paragraaf 2.3). De ontwikkelingsgerichte pilots en interventies (zie hoofdstuk 3) zullen laten zien hoeveel sneller cursisten kunnen leren als NT2-onderwijs in overeenstemming met de succesfactoren uit het theoretisch kader wordt vormgegeven (door competente docenten onder gunstige omstandigheden). Het is dan ook zaak om eerst de uitkomsten van de ontwikkelingsgerichte pilots en de interventies af te wachten alvorens aanbevelingen te doen aan cursisten, onderwijsaanbieders of de overheid. Indachtig deze waarschuwing kunnen ten aanzien van het toekomstige gebruik van een benchmark voor andere doeleinden dan onderzoek ten minste vier functies worden onderscheiden:

- Ondersteuning overheidsbeleid;
- Advisering van individuele cursisten bij de intake;
- Begeleiding tijdens het inburgeringsprogramma;
- Evaluatie van onderwijsaanbieders.

Ondersteuning van het overheidsbeleid

Om beleidskeuzen te kunnen onderbouwen en verantwoorden wil de overheid graag weten hoeveel tijd en lesuren verschillende groepen inburgeraars gemiddeld gebruiken om het voor het inburgeringexamen of Staatsexamen NT2 vereiste taalniveau te behalen. De benchmark en de flankerende pilots en interventies kunnen de overheid voorzien van realistische informatie over de tijd en lesuren die relevante groepen inburgeraars gebruiken voor het bereiken van het taalniveau dat voor het inburgeringexamen of het Staatsexamen NT2 I of II vereist is (zowel onder ongunstige, gemiddelde als gunstige omstandigheden). De overheid kan deze informatie benutten bij het plannen van onderwijstrajecten voor inburgeraars en het toewijzen van voldoende middelen aan gemeenten.

Advisering van individuele cursisten bij de intake

Tijdens de intakefase krijgen inburgeraars een advies over het streefniveau en de tijdsduur en het aantal lesuren dat voor het bereiken van dat streefniveau nodig geacht wordt. Dit advies met betrekking tot een passend onderwijsprogramma is vaak minder realistisch in de zin dat het streefniveau systematisch te hoog wordt ingeschat en de benodigde tijdsduur en lesuren te laag. Zo bleek een derde van het cohort nieuwkomers dat in 1999 instroomde niet op het juiste niveau te zijn geplaatst (Brink, Does & Odé, 2002). Een belangrijke gebruiksfunctie van de benchmark (in combinatie met onderzoek naar de leerlast onder gunstige omstandigheden) is het leveren van een bijdrage aan een verbetering van de advisering en plaatsing van individuele inburgeraars. Hierbij zij opgemerkt dat de benchmark uitsluitend een beschrijvende en geen vóórschrijvende functie heeft. De benchmark beschrijft hoeveel tijd/lesuren inburgeraars gemiddeld nodig hebben, en niet in hoeveel tijd of lesuren zij dat niveau zouden kunnen of moeten behalen. De resultaten van de benchmark en het leerlastonderzoek zouden op termijn aan intakers ter beschikking gesteld kunnen worden in de vorm van een computerapplicatie. De intaker hoeft dan slechts de relevante kenmerken van de inburgeraar in te voeren (bijv. opleidingsniveau, oud- versus nieuwkomer, startniveau), waarna het programma een passend advies verstrekt over onder meer het streefniveau en over de tijdsduur en het aantal lesuren die normaliter voor het bereiken van dat niveau nodig zijn.

Begeleiding tijdens het inburgeringprogramma

Aanbieders van NT2-onderwijs moeten regelmatig nagaan of het onderwijsprogramma binnen de geformuleerde randvoorwaarden nog steeds passend is. Het Deltaplan Inburgering spreekt over de noodzaak van periodieke herijking en evaluatie (Ministerie van VROM, 2007). Op basis van leerlastonderzoek met de benchmark als criterium kan onder gunstige omstandigheden worden nagegaan hoeveel tijd en lesuren groepen cursisten gegeven bepaalde kenmerken nodig hebben voor het behalen van het inburgeringexamen. Daarnaast kan het een instrument zijn aan de hand waarvan docenten en begeleiders op de instelling bepalen in hoeverre een specifieke klas of een jaarlaag van cursisten “op schema ligt” (en waarmee de instelling kan evalueren of het onderwijsprogramma nog wel voldoende is afgestemd op de groep cursisten). De docenten en begeleiders kunnen dan bepalen hoeveel langzamer of sneller dan de gemiddelde groep in Nederland een bepaalde groep cursisten “gegroeid” is (gegeven het startniveau, de duur en de intensiteit van het gevolgde onderwijsprogramma). Op dit moment ontbreken harde gegevens over de tussenniveaus die cursisten gegeven de gebruikte tijd en lesuren zouden kunnen bereiken (gegeven goed onderwijs door een competente docent onder gunstige omstandigheden). De combinatie van leerlastonderzoek en benchmark kan op termijn een bijdrage leveren aan de verbetering van de tussentijdse advisering van groepen inburgeraars die al gevorderd zijn op de weg naar hun diploma. Daarnaast kunnen onderwijsaanbieders worden ondersteund bij de evaluatie en verbetering van hun NT2-onderwijsprogramma.

Evaluatie onderwijsaanbieders

De kwaliteit van de onderwijsaanbieders is van groot belang voor het rendement van inburgeringonderwijs. Volgens het Deltaplan inburgering moet het niveau en het rendement van de inburgeringprogramma's omhoog (Ministerie van VROM, 2007). Rendement zou men in dit verband kunnen definiëren als het behalen van een zo hoog mogelijk percentage met succes afgeronde trajecten in een zo kort mogelijke tijd met zo weinig mogelijk kosten. Met de invoering van de Wet Inburgering is marktwerking in het NT2-onderwijs integraal ingevoerd. De aanbesteding van NT2-onderwijs ligt nu in handen van de gemeenten die een keuze maken uit verschillende onderwijsaanbieders. De benchmark kan op termijn een instrument worden voor gemeenten bij het ontwikkelen van hun aanbestedingsbeleid, het maken van prestatieafspraken met aanbieders, het evalueren van het rendement van de aanbieders en het bereiken van een goed evenwicht tussen kwaliteit en prijs.

Bijlage 2

Kwaliteit en bruikbaarheid van bestaande bronnen voor benchmarkconstructie

Inleiding

Een benchmark NT2 kan in beginsel op twee manieren worden ontwikkeld die ieder hun eigen mogelijkheden en beperkingen hebben:

- Secundaire analyse van bestaande bestanden of rapportages;
- Primaire gegevensverzameling via empirisch onderzoek.

In een secundaire analyse heranalyseren onderzoekers door anderen verzamelde bestanden of rapportages en bij primaire analyse verzamelen de onderzoekers de gegevens zelf. In deze bijlage wordt verslag gedaan van een haalbaarheidsstudie naar de bruikbaarheid van bestaande bronnen voor benchmarkconstructie. Daarbij is een onderscheid gemaakt in de kwaliteit en bruikbaarheid van de rapportages die gemeenten jaarlijks aan de overheid verstrekken (paragraaf 2.1) en van de cliëntsysteem waarover gemeenten, intakers en onderwijsaanbieders beschikken (paragraaf 2.2). Net als secundaire analyse kent ook primaire gegevensverzameling de nodige problemen. Een analyse van enkele veel voorkomende onderzoekstechnische problemen waarmee primaire gegevensverzameling veelvuldig te kampen heeft en die bij de constructie van een benchmark moeten worden opgelost is opgenomen in paragraaf 2.3. Paragraaf 2.4 bevat de belangrijkste conclusies en aanbevelingen van deze haalbaarheidsstudie.

Om bestaande rapportages, bestanden en systemen te kunnen evalueren op kwaliteit en bruikbaarheid is een criterium noodzakelijk. Eerder in deze publicatie is een aantal kwaliteitseisen geformuleerd waaraan een bruikbare benchmark moet voldoen. De benchmark moet niet alleen voldoende gedifferentieerd, landelijk representatief en efficiënt zijn, maar ook betrouwbaar en valide. Een van de belangrijkste eisen is dat de benchmark de tijd/lesuren die nodig zijn voor het bereiken van het overeengekomen taalniveau nauwkeurig moet kunnen voorspellen (d.w.z. met een zo klein mogelijke standaardfout/betrouwbaarheidsinterval). Daartoe moeten leerlast, taalniveau en opleidingsniveau betrouwbaar en valide worden gemeten. Meetfouten verlagen immers de correlatie tussen leerlast en rendement en hiermee de voorspelbaarheid van de tijd/lesuren die nodig zijn om het overeengekomen eindniveau te bereiken.

2.1 Kwaliteit en bruikbaarheid van bestaande rapportages

In Nederland zijn verschillende pogingen ondernomen om door middel van secundaire analyse van bestaande bestanden of rapportages gegevens over het rendement van inburgeringsprogramma's te verzamelen (o.a. Brink, Does, Odé & Tromp, 2002; Significant, 2004; Tuijl, Braun & Weteling, 2004; Significant, 2007). De secundaire analyses zijn uitgevoerd op bestanden en standaard rapportages van onder meer gemeenten, intakers en onderwijsinstellingen (zoals jaarverslagen en cursistendossiers). In deze sectie worden enkele van deze onderzoeken exemplarisch beschreven (paragraaf 2.1.1) waarbij tevens wordt ingegaan op de bruikbaarheid van dergelijke secundaire analyses als basis voor het construeren van een NT2-benchmark (paragraaf 2.1.2). Het accent ligt hierbij niet zozeer op het beschrijven van de onderzoeksresultaten, maar meer op het inventariseren van de onderzoekstechnische mogelijkheden en problemen die de ontwikkeling van een benchmark NT2 kunnen faciliteren dan wel bemoeilijken.

2.1.1 Recentelijk uitgevoerde secundaire analyses

Het onderzoek van Brink, Does, Odé en Tromp (2002)

Een voorbeeld van een goed opgezette en zorgvuldige uitgevoerde retrospectieve cohortstudie is het onderzoek van Regioplan naar de effectiviteit van de WIN (Brink, Does, Odé & Tromp, 2002). Zij “volgden” het cohort nieuwkomers dat in het jaar 1999 instroomde in een inburgeringprogramma vanaf de aanmelding en de start van het inburgeringonderzoek tot en met de uitreiking van het inburgeringcertificaat. De cohortanalyse werd uitgevoerd in het tweede en derde kwartaal van 2001 over een periode van 22 maanden. Voor de analyse van het cohort 1999 bestudeerden zij de *individuele dossiers* van 970 nieuwkomers die in 1999 waren begonnen met hun inburgeringprogramma. Brink, Does, Odé en Tromp (2002) kozen voor analyse van individuele dossiers van inburgeraars omdat de jaarverslagen die gemeenten aan het Rijk verstrekten naar hun mening “slechts een algemeen beeld (gaven) van de uitval tijdens de inburgering, het aantal afgeronde programma’s en de behaalde niveaus door nieuwkomers” (p. 10). In het onderzoek bleken de gemeenten een groot aantal gegevens van het inburgeringprogramma per individuele nieuwkomer bij te houden. Er bleek echter een groot verschil in de kwaliteit van de registraties. Lang niet van alle nieuwkomers waren alle gegevens bekend. In sommige gemeenten waren er bijvoorbeeld zeer uitgebreide verslagen van intakegesprekken in het dossier aanwezig, maar vaak was de informatie erg summier. De individuele dossiers waarover de gemeenten beschikten, gaven ook informatie over de wijze waarop intakers gegevens registreren. Dat verschilde sterk van intaker tot intaker. Zo bleken intakers verschillende registratiesystemen te gebruiken, er waren nog intakers die hun registratie voornamelijk schriftelijk bijhielden, er werden lang niet altijd dezelfde aspecten van het inburgeringprogramma geregistreerd, voor de geregistreerde gegevens werden soms verschillende definities gebruikt, de uitwisselbaarheid van de gegevens tussen verschillende organisaties bleek beperkt, jaarverslagen werden lang niet altijd volledig ingevuld en de periode waarover geregistreerd werd verschilde van intaker tot intaker. Met betrekking tot de start van het inburgeringonderzoek was van een groot aantal nieuwkomers de startdatum onbekend. De auteurs concludeerden dat er geen sprake was van een landelijk, eenduidig registratiesysteem voor het registreren van de individuele gegevens van nieuwkomers.

Met betrekking tot het geregistreerde opleidingsniveau van de nieuwkomers bleek uit de dossiers dat ROC’s met verschillende cursistenprofielen werken en dat de profielen op verschillende wijze werden gedefinieerd. De meeste ROC’s gingen uit van de profielen analfabeten, laagopgeleiden, middenopgeleiden en hoogopgeleiden. Sommige ROC’s gingen echter uit van KSE-niveaus, anderen van sociale, professionele en educatieve redzaamheid en weer andere van de uitstroomniveaus van de staatsexamens NT2 I en II.

Met betrekking tot de inhoud van het inburgeringprogramma bevatten de dossiers doorgaans alleen de aanvangsdatum, type cursus en de locatie van het ROC. Uit de dossiers kon niet worden opgemaakt hoeveel uren de nieuwkomer daadwerkelijk had gevolgd. In een aantal dossiers bevonden zich wel presentielijsten, maar vaak niet over de gehele periode van twaalf maanden. Nieuwkomers bleken nogal eens te wisselen van programma, bijvoorbeeld als zij betaald werk vonden of stopten met werken, maar de onderzoekers konden deze informatie naar eigen zeggen niet op systematische wijze onderzoeken.

Met betrekking tot de afronding van het inburgeringprogramma bleek uit de dossiers dat de gehanteerde toetsen om het eindniveau vast te stellen sterk van opleiding tot opleiding verschilden. De onderwijsinstellingen hanteerden verschillende toetsen om het begin-, tussen- en eindniveau van de cursisten te bepalen. Ook werden er voor verschillende opleidingsniveaus – analfabeten, laag, midden en hoog opgeleiden – veelal verschillende toetsen gebruikt. Omdat deze toetsen niet geëquivalet waren, was het niet goed mogelijk om scores tot elkaar te herleiden (en het taalniveau over opleidingsniveaus te kunnen

vergelijken). De veel gebruikte NT2-profieltoets leverde in beginsel wel landelijk vergelijkbare scores, maar deze toets werd vanwege het gebrek aan civiel effect door opleidingen en inburgeraars vaak niet serieus genomen. Voor bijna een kwart van de nieuwkomers gaven de dossiers geen uitsluitsel over de vraag of de NT2-profieltoets al dan niet was afgelegd. Als dat wel bekend was, bleek het afnamemoment sterk van cursist tot cursist te verschillen. In veel gevallen lukte het niet om de nieuwkomers binnen de wettelijke termijn van één jaar na inschrijving bij de onderwijsaanbieder te toetsen en andere gevallen gaven inburgeraars helemaal geen gehoor aan de oproep om te worden getoetst. Met betrekking tot de vooruitgang in taalniveau die de inburgeraars boekten en het uiteindelijk bereikte eindniveau bleken er lang niet altijd van alle nieuwkomers gegevens in de dossiers te zijn vastgelegd. Deze tekortkomingen doen sterk afbreuk aan de landelijke vergelijkbaarheid van de gegevens over het taalniveau en het opleidingsniveau.

Met betrekking tot de leerlast van nieuwkomers zijn de dossiers ook geanalyseerd op verzuim (af een toe een les missen), tijdelijke uitval (gedurende een langere periode geen lessen volgen en vervolgens weer terugkeren op school) en definitieve uitval (voortijdig stoppen en niet weer opnieuw beginnen). In de gemeenten met een bonussysteem bleek de presentie goed te worden bijgehouden, maar in andere gemeenten werd dat veel minder goed gedaan. Afhankelijk van het type nieuwkomer dat het inburgeringprogramma had afgerond, bleek er van 15% tot 36% geen gegevens over het verzuim beschikbaar. Van 7% bleek niet bekend of zij het programma hadden afgemaakt. De algemene indruk was dat er veel verzuimd werd, maar dat definitieve uitval beperkt bleef.

Het onderzoek van Tuijl, Braun en Weteling (2004)

Een ander voorbeeld van een retrospectieve cohortstudie is het onderzoek van de PIERS GROEP (Tuijl, Braun & Weteling, 2004). De onderzoekers analyseerden de door 496 gemeenten ingevulde ‘vragenlijst inburgering nieuwkomers’ over het verslagjaar 2002. De vragenlijst was zo opgesteld dat het mogelijk was om één jaar terug te kijken waardoor de cohortanalyse een periode van twee jaar kon bestrijken. De onderzoekers presenteren gegevens over het rendement van NT2-onderwijs in het jaar 2002. Met rendement bedoelden de onderzoekers de toegevoegde waarde in beheersing van de Nederlandse taal tussen de start van het programma (moment waarop de beschikking wordt opgesteld) en de verklaring zoals door het ROC werd afgegeven. In één jaar tijd steeg het gemiddelde taalniveau met ongeveer één taalniveau (in 2000 van .47 naar 1.50 en in 2001 van .41 naar 1.38). Het daadwerkelijk bereikte taalniveau bleek echter gemiddeld sterk achter te blijven bij het beoogde eindniveau (van 2.23 in 2000 en 2.20 in 2001). De toegevoegde waarde werd ook per startniveau gerapporteerd. Van de inburgeraars die waren gestart op niveau 0 bevond 30% zich na één jaar nog steeds op datzelfde niveau, 37% had niveau 1 bereikt, 25% niveau 2, 7% niveau 3, 1% niveau 4 en eveneens 1% niveau 5. Ook bij de “hogere” startniveaus boekten lang niet alle inburgeraars progressie, terwijl een groot deel merkwaardigerwijs in taalvaardigheid achteruit ging. Zo bevond van de groep die in 2002 op niveau 3 instroomde zich aan het eind van dat jaar 43% nog steeds op niveau 3, van 23% was de taalvaardigheid gestegen naar niveau 4 of 5 en van de overige 43% was de taalvaardigheid gedaald naar niveau 2, 1 of 0. Van degenen die geïndiceerd waren op niveau 4 ging 29% qua taalvaardigheid achteruit en bij degenen die startten op niveau 5 bedroeg dit percentage zelfs 94%. Als verklaring noemen Tuijl, Braun en Weteling (2004) dat “bij het begin en eind verschillende toetsen worden gebruikt die niet op eenzelfde schaal kunnen worden ondergebracht” (p. 54).

In hun secundaire analyse bestudeerden de onderzoekers ook gegevens over het vastgestelde en het daadwerkelijk gevolgde aantal uren NT2-onderwijs. Volgens de opgave van de gemeenten bleek 1% van de nieuwkomers meer dan 800 uur gevolgd te hebben, 13% volgde 600 tot 800 uur, 49% volgde 400 tot 600 uur, 35% volgde 1 tot 400 uur en 2% volgde

geen onderwijs. Uit de gegevens bleek ook dat cursisten gemiddeld 13% minder uren NT2-onderwijs volgden dan tijdens de vaststellende beschikking was overeengekomen.

Het onderzoek van Significant (2004)

In een cross-sectionele studie analyseerde Significant (2004) de geaggregeerde gegevens die 472 gemeenten in het peiljaar 2003 aan het Rijk aanleverden over de voortgang van de 44.643 deelnemers aan inburgeringstrajecten (in het kader van de financiële verantwoording aan de overheid). De analyse van de taalvaardigheid van de 20.188 afronders liet zien dat de taalvaardigheid gemiddeld met één NT2-niveau was gestegen. Hierbij merken de onderzoekers op dat het beginniveau van de afronders doorgaans door de ambtenaar van het inburgeringsonderzoek en niet met een NT2-toets was vastgesteld. Gegevens over de vragen naar de taalvaardigheid van de inburgeraar bleken in 14% tot 20% van de gevallen te ontbreken, terwijl ook de vraag naar de uitsplitsing van de inburgeraars naar cohort bij veel gemeenten slecht was ingevuld en inconsistenties te zien gaf. De vragen naar de verdeling van de inburgeraars naar het afgesproken aantal uren educatie, het opleidingsniveau, het land van herkomst en de leeftijd werden daarentegen wel goed ingevuld (met hooguit 6% missings).

Het onderzoek van Significant (2007)

Significant (2007) analyseerde geaggregeerde gegevens die gemeenten over het jaar 2006 aan het Rijk verstrekten over de in-, door- en uitstroom van oud- en nieuwkomers. In de Monitor oud- en nieuwkomers is ook gekeken naar kenmerken van de starters, de aard, duur en intensiteit van hun onderwijsprogramma's en het resultaat, waaronder het aantal dat voortijdig uitvalt, het aantal dat een programma afrondt en de toename van de taalvaardigheid.

Met betrekking tot de leerlast bleek bijna tweederde van de nieuwe programma's voor oudkomers (64%) een intensiteit te hebben van tussen de 9 en 14 uur per week. De meeste programma's voor oudkomers (54%) duurden tussen een half jaar en een jaar. Met betrekking tot de nieuwkomers werd in 2006 bij ruim de helft (55%) van de nieuwe trajecten minder dan 600 uur educatie aangeboden. Bij 20% van de trajecten werd tussen de 200 en 399 uur educatie aangeboden en in 32% van de trajecten tussen de 400 en 599 uur. Bij 45% van de trajecten werd meer dan 600 uur educatie aangeboden. Een overgrote meerderheid hiervan (44%) had tussen de 600 en 799 uur educatie. Gezien het relatief grote aandeel van trajecten met tussen de 200 en 399 uur educatie en het zeer kleine aandeel van trajecten meer dan 800 uur educatie (1%), vermoedden de onderzoekers dat het gemiddelde aantal uur educatie onder de (toenmalige) richtlijn van 600 uur lag.

Voor 45% van de afrondende oudkomers bleken consistente gegevens beschikbaar over de gemiddelde niveaustijging in één jaar tijd. Van deze afrondende oudkomers bleef 52% in één jaar tijd op hetzelfde niveau, 40% steeg één niveau en de overige 8% steeg twee of meer niveaus. Anders dan in eerder onderzoek het geval was, ging het taalniveau van de oud- en nieuwkomers in geen enkel geval achteruit (hetgeen in werkelijkheid niet plausibel lijkt, maar gezien de beperkte vergelijkbaarheid van de gebruikte toetsen een reële mogelijkheid is). Een nadere bestudering van de gebruikte vragenlijst brengt aan het licht dat dit resultaat waarschijnlijk niet te danken is aan een verbeterde meting van het taalniveau bij de start en de afronding. De aard van de vraagstelling maakte rapportage van niveaudaling namelijk onmogelijk. Dit wil zeggen: in de Vragenlijst oudkomers konden gemeenten voor starters op niveau 1 en 2 wel aangegeven of het niveau hetzelfde of hoger was geworden, maar niet of het achteruit gegaan was. Bij niveau 3 konden gemeenten voor oudkomers alleen het aantal cursisten invoeren dat op niveau gebleven was (en dus niet of het niveau gedaald of gestegen was).

Bijna de helft (47%) van de 35.455 deelnemende nieuwkomers die in 2006 of in de jaren daarvoor startten, rondde in 2006 het inburgeringstraject af. Het gemiddelde startniveau

bedroeg .33 en het gemiddelde eindniveau 1.33. Het niveau per taalvaardigheden steeg voor lezen gemiddeld van .30 naar 1.40, voor schrijven van .30 naar 1.25, voor spreken van .32 naar 1.34 en voor luisteren van .35 naar 1.33. Voor de nieuwkomers waarvoor gegevens beschikbaar waren, bleef 34% in één jaar tijd op hetzelfde niveau, 37% steeg één niveau en 29% steeg twee of meer niveaus. De onderzoekers rapporteerden ook groeicijfers per startniveau. Zo stegen nieuwkomers die bij aanvang van hun inburgeringstraject een NT2-niveau hadden van 1 relatief vaker een niveau dan nieuwkomers die op NT2-niveau 0 startten. Ongeveer een derde van de nieuwkomers die op niveau 0 startten, eindigde ook op dat niveau. Van de starters die op niveau 1 startten, bleef slechts een kwart op niveau 1.

De door de gemeente verstrekte informatie over de verdeling van de oud- en nieuwkomers naar de aard, duur en intensiteit van de programma's en naar het geslacht, land van herkomst en leeftijd bleek nagenoeg compleet te zijn. Voor het door oudkomers in 2006 bereikte eindniveau in lezen, schrijven, spreken en luisteren lagen de percentages ontbrekende waarnemingen tussen de 31% en 33% (en voor de nieuwkomers tussen 12% en 14%). Het percentage afronders in 2006 waarvoor betrouwbare taalgegevens beschikbaar waren, bedroeg 45%. De vragen over het taalniveau van de vier taalvaardigheden lezen, schrijven, spreken en luisteren bleken voor veel gemeenten lastig in te vullen. De onderzoekers constateerden een aantal grote, onverklaarbare afwijkingen tussen de stroomaantallen in het meest recente en voorgaande monitor waardoor vergelijkingen over tijd met de nodige voorzichtigheid geïnterpreteerd moeten worden. Bij de groep nieuwkomers kozen de onderzoekers er vanwege deze inconsistenties in de data voor om ontwikkeling in de aantallen nieuwkomers over meerdere jaren uit te drukken in percentages en niet in absolute aantallen.

California Department of Education (2004) en Connecticut State Department of Education (2007)

Op basis van een omvangrijke dataset onderzocht het California Department of Education (2004) de relatie tussen de hoeveelheid instructie en de gemiddelde leerwinst op een gestandaardiseerde leestest bij volwassen *English as a second language* (ESL) cursisten. Het Connecticut State Department of Education (2007) voerde een vergelijkbaar onderzoek uit naar de relatie tussen het aantal lesuren en de basale luister- en leesvaardigheid van volwassen ESL cursisten. In beide onderzoeken werd een monotoon stijgende relatie tussen het aantal lesuren en de gemiddelde leerwinst in taalvaardigheid gevonden. De analyse is uitgevoerd op geaggregeerde gegevens en de onderzoekers rapporteren geen correlaties of percentages verklaarde variantie, zodat de resultaten moeilijk interpreteerbaar zijn.

De overzichtsstudie van Beder (1999)

De ervaringen met secundaire analyse in de Verenigde Staten lijken sterk op die in Nederland. Op basis van een overzichtsstudie concludeerde Beder (1999) dat secundaire analyse van voor andere doeleinden dan onderzoek verzamelde gegevens over *adult literacy programs* in de VS gekenmerkt wordt door “inaccurate learner records, failure to pre and posttest at specified intervals, administration of inappropriate levels of tests, failure to test, high attrition of subjects between pre and post data collection, programs' withdrawal from the study before program data collection was complete, and failure to forward data to researchers in a timely fashion” (p. 116). Enkele jaren later trekken Comings en Soricone (2007) een soortgelijke conclusie.

2.1.3 Enkele specifieke problemen van secundaire analyse van gegevens over leerlast, taalniveau en rendement

In deze paragraaf bespreken we nog enkele specifieke problemen van secundaire analyse waar het de compleetheid en kwaliteit van variabelen betreft die voor de constructie van een (smalle) benchmark NT2 essentieel zijn.

Benodigde tijdduur

Ten aanzien van de leerlast is een onderscheid gemaakt in de duur van het traject tussen intake en afronding en de intensiteit van het onderwijsprogramma. De tot nu toe uitgevoerde secundaire analyses geven weinig aanleiding om aan de betrouwbaarheid van de gegevens over de duur van het inburgeringstraject van start tot finish te twijfelen. Ten aanzien van de gemiddelde tijdsduur van de afzonderlijke niveaustappen voor de verschillende groepen inburgeraars (bijvoorbeeld van niveau 0 tot 1, van 1 tot 2, et cetera) lijken de gegevens veel minder nauwkeurig te zijn. Daarnaast ontbreken nauwkeurige metingen van de verschillen tussen cursisten binnen de groepen in de per niveaustap benodigde tijd.

Intensiteit van het programma

In de secundaire analyses zijn de variabelen met betrekking tot de intensiteit van het inburgeringprogramma grofmazig en hooguit op ordinaal niveau gemeten. Zo mat Significant (2007) de verdeling van de aantallen oudkomers naar de duur van het programma op een schaal met slechts vijf categorieën, te weten korter dan zes maanden, zes tot twaalf maanden, één tot anderhalf jaar, anderhalf tot twee jaar en langer dan twee jaar. De verdeling van de aantallen oudkomers naar de intensiteit van het programma werd gemeten op een schaal met slechts vier categorieën: 0-4 uur les, 5-8 uur les, 9-14 uur les en 15 of meer uur les. De verdeling van aantallen nieuwkomers naar de intensiteit van het overeengekomen en daadwerkelijk gerealiseerde programma werd gemeten op een schaal met de categorieën 0, 1 – 199, 200 – 399, 400 – 599, 600 – 799 en 800 of meer lesuren. Een dergelijk grofmazige meting van de leerlast verlaagt de correlatie tussen leerlast en rendement en hiermee de voorspelbaarheid van de tijd/lesuren die nodig zijn om het overeengekomen eindniveau te bereiken.

Het taalniveau bij de start, tussentijds en aan het eind

In de bestudeerde secundaire analyses is het taalniveau van de inburgeraars net als de gegevens over de leerlast gebaseerd op geaggregeerde gegevens (met het gevaar van aggregatiebias en het onderschatten van de varianties). Tevens zijn de resultaten op de taaltoetsen niet als fijnmazige continue toetsscores beschikbaar, maar alleen in de vorm van de ingedikte categorisering in de vier taalniveaus NT2 0, 1, 2 en 3 (hetgeen de correlatie tussen leerlast en rendement attenuëert en de voorspelbaarheid beperkt). Uit onderzoek is gebleken dat een groot deel van de inburgeraars zelfs na vele honderden lesuren op een dergelijk grofmazige schaal geen enkele leerwinst weet te realiseren. Bij de ontwikkeling van een nieuwe benchmark NT2 verdient het derhalve aanbeveling de vooruitgang van de cursisten op een minder grove manier vast te stellen en te rapporteren (d.w.z. niet alleen als een categoriale variabele in termen van de CEFR-niveaus, maar ook op de onderliggende continue schaal van het meetinstrument).

Daarnaast lijkt het startniveau nogal eens gebaseerd te zijn op een subjectieve inschatting van de intaker in plaats van op een objectieve toets. Informatie over de aard van de toetsen lijkt in geen van de geanalyseerde bestanden en rapportages voorhanden. Een ander probleem is dat er voor de vaststelling van het taalniveau verschillende niet-geëquivalente NT2-toetsen in omloop zijn waarvan de scores beperkt vergelijkbaar zijn. Veel van deze

toetsen zijn bovendien sterk verouderd, worden niet regelmatig ververs, worden onder onvoldoende gestandaardiseerde condities afgenomen en beoordeeld en/of worden niet erg serieus genomen omdat ze leergangonafhankelijk zijn en weinig civiel effect hebben. Het is mogelijk om deze redenen dat er in de besproken secundaire analyses geen direct verband gelegd kon worden tussen het aantal gevolgde uren en het behaalde resultaat.

Verder zijn de beschikbare secundaire gegevens in de vorm van bijvoorbeeld jaarverslagen inflexibel omdat ze geen analyse van frequentere tijdintervallen toelaten. Vaak zijn alleen gegevens beschikbaar over de leerlast en het rendement aan het begin en einde van het onderwijstraject of aan het begin en einde van de verslagperiode (terwijl voor het beschrijven van de leercurven van de cursisten meer dan twee metingen noodzakelijk zijn).

Het opleidingsniveau

In de aangehaalde secundaire analyse is de verdeling van de aantallen inburgeraars naar opleidingsniveau doorgaans beschreven in de vijf categorieën analfabeet, lees- en schrijfvaa, laag opgeleid, middelbaar opgeleid, hoog opgeleid (en een restcategorie onbekend). Onduidelijk is in hoeverre deze informatie het resultaat is van een gedegen internationale diplomawaardering dan wel berust op mondelinge informatie van de inburgeraars zelf of een subjectieve inschatting van de intaker. Fijnmazige informatie over het exacte aantal jaren dat de inburgeraars in het land van herkomst onderwijs hebben gevolgd lijkt te ontbreken. Een bijkomend probleem is dat het opleidingstype in de meest recente versie van de Monitor oud- en nieuwkomers (Significant, 2007) wel van nieuwkomers maar niet van oudkomers bekend lijkt te zijn. Hiermee lijkt de meting van het opleidingsniveau voor doeleinden van leerlastonderzoek onvolledig volledig en onvoldoende precies.

2.1.4 Bruikbaarheid secundaire bronnen voor benchmarkconstructie

Secundaire analyse van bestaande gegevens heeft onmiskenbare voordelen. Het is vaak snel, efficiënt en goedkoop omdat de fase van primaire dataverzameling wordt overgeslagen. De jaarverslagen en andere rapportages die gemeenten in het kader van de financiële verantwoording standaard aan het Rijk verstrekken bieden een grote hoeveelheid gegevens over de inburgering. Daarnaast zijn de gegevens makkelijk toegankelijk en snel te analyseren en te rapporteren. Verder zijn er gegevens over een langere periode beschikbaar. Aan het construeren van een NT2-benchmark op basis van deze secundaire gegevens zijn echter ook potentiële nadelen verbonden:

- De secundaire gegevens zijn niet verzameld met het oog op de constructie van een benchmark NT2, hetgeen de bruikbaarheid beperkt. Belangrijke variabelen ontbreken of zijn onvoldoende betrouwbaar of valide gemeten.
- Omdat gebruik wordt gemaakt van eerder verzamelde gegevens, zijn de gegevens minder actueel (van vóór de WI) en daardoor minder bruikbaar.
- Secundaire gegevens zijn meestal niet onder gecontroleerde omstandigheden verzameld waardoor de betrouwbaarheid en validiteit van de gegevens niet gegarandeerd is.
- De secundaire analyses lijken op geaggregeerd niveau relatief complete informatie op te kunnen leveren over onder meer de doelgroep (bijv. oudkomer/nieuwkomer, gezinsvormer/-hereniger, asielzoeker, geestelijke bedienaar), het NT2-streefniveau, het geslacht, de leeftijd, het land van herkomst, de start en afronding van het inburgeringstraject en de uitval. De percentages ontbrekende gegevens over het taalniveau zijn doorgaans hoog. De gegevens over het taalniveau zijn gebaseerd op verschillende, niet-geëquivalente toetsen van sterk wisselende kwaliteit. De toetsresultaten lijken ook minder goed vergelijkbaar omdat er niet onder gestandaardiseerde condities is afgenomen

en beoordeeld. Daarnaast zijn de verzamelde toetsgegevens te grofmazig (van ordinaal niveau) om als basis voor benchmarkconstructie te kunnen dienen (waarvoor gegevens op ten minste intervalniveau nodig zijn).

- Makkelijk toegankelijke en ogenschijnlijk direct bruikbare gegevens zijn beschikbaar in de vorm van overzichten, stroomdiagrammen en tabellen (denk bijvoorbeeld aan de verdeling van de aantallen inburgeraars naar onder meer startniveau, aantal lessen en eindniveau). Bij analyse van geaggregeerde gegevens is rekening te houden met aggregatiebias. Een zelfde variabele kan op geaggregeerd niveau een geheel andere betekenis hebben dan op individueel niveau. De op het hogere niveau vastgestelde samenhang is vaak anders (doorgaans sterker) dan op individueel niveau. Daarnaast is de variantie van geaggregeerde gegevens meestal kleiner dan van gegevens op individueel niveau (*range restriction*).
- Als niet-geanonimiseerde gegevens op individueel niveau van belang zijn (bijvoorbeeld om bestanden te kunnen koppelen, het succes van de match te kunnen checken of om personen op naam te kunnen benaderen), moeten zij de desbetreffende personen vanwege de wet op de bescherming persoonsgegevens achteraf om toestemming kunnen vragen. Omdat het eerder verzamelde gegevens betreft, kan het verkrijgen van toestemming erg lastig zo niet onmogelijk zijn en gepaard gaan met aanzienlijke experimentele uitval (denk aan de hoge mobiliteit van bepaalde groepen inburgeraars).
- Het aanvullend opvragen van gegevens op individueel niveau kan in de praktijk veel tijd, moeite en geld kosten. Denk bijvoorbeeld aan het doorspitten van cursistendossiers waarvan de inhoud en structuur sterk van instelling tot instelling kunnen verschillen.

Concluderend kan men stellen dat de bruikbaarheid van de bestaande rapportages als basis voor de constructie van een benchmark NT2 zeer beperkt is.

2.2 Kwaliteit en bruikbaarheid van cliëntvolgsystemen

Een andere voor de hand liggende bron zijn de cliëntvolgsystemen van gemeenten, intake- en onderwijsaanbieders. Hieruit zouden immers vergelijkbare gegevens van inburgeraars te verifiëren kunnen zijn. Belangrijk is echter wel dat de registratie van die gegevens zorgvuldig en op een zelfde manier gebeurt.

In de haalbaarheidsstudie is ook gekeken of het systeem ISI (Informatiesysteem Inburgering) de benodigde gegevens bevat. Dit systeem is van de Informatie Beheer Groep en hierin kan per cliënt informatie worden geregistreerd, met het oog op wat vanuit de Wet Inburgering (WI) vereist is. Hierna is een oriënterend onderzoek uitgevoerd naar cliëntsystemen die naast ISI gebruikt worden door gemeentes. Hieruit kwamen in eerste instantie drie cliëntvolgsystemen naar voren: GWS4all, IW3 en Stratech. Er is informatie opgevraagd over de systemen, hetgeen enigszins een beeld gaf van wat er al dan niet in geregistreerd kan worden. Echter, het is ook voor een aanzienlijk deel aan gemeentes wat ze registreren en hoe ze dat doen. Daarom is ervoor gekozen om in eerste instantie bij ieder cliëntstelsel een gemeente te benaderen die dat systeem gebruikt om na te gaan hoe men te werk gaat. Naar aanleiding van die resultaten wordt besloten of het zinvol is om bij meer gemeentes navraag te doen. Daarnaast is bij een sample van veertien gemeentes (verschillend in grootte en ligging in het land) nagegaan of zij een cliëntvolgsysteem gebruiken en zo ja, welke. Dit om te voorkomen dat er nog een belangrijker speler in deze markt over het hoofd gezien wordt.

Registratiesystemen

ISI

Via de website van de Informatie Beheer Groep is een zogenoemde legenda beschikbaar waarin te lezen staat welke gegevens in ISI geregistreerd worden. De financiering van inburgeringstrajecten aan gemeentes is mede gebaseerd op wat er in ISI ingevoerd wordt. De invulvelden hebben dan ook met name betrekking op zaken die met het financiële aspect te maken hebben: doelgroep (oudkomer / nieuwkomer), het aanbod, handhaving van inburgeringsplichtigen, vrijstellingen, resultaten van het inburgerings- of staatsexamen et cetera. Enkele gegevens zijn bruikbaar, maar informatie over vrijstellingen en handhaving zijn niet relevant voor dit onderzoek. Reden te meer om na te gaan wat gemeentes naast ISI registreren en hoe ze dat doen.

Stratech

Van Stratech is het marktaandeel niet geheel helder, ook niet na telefonisch contact met de softwareontwikkelaar. Wel is verteld dat als het om gemeentes gaat Den Haag en Almelo de grootste klanten zijn. Stratech is volledig gebaseerd op het Berichtenboek WI 7.2 zoals opgesteld door de IB-Groep. Dit Berichtenboek bevat de functionele beschrijving van de Wet Inburgering. Stratech richt zich dus op de gegevens die in het kader van de naleving van de WI van belang zijn. Dat blijkt ook bij bestudering van het Berichtenboek: het zijn voornamelijk gegevens zoals ook in ISI te vinden zijn. Informatie over leerlast en rendement zal dus niet in voldoende mate geregistreerd zijn in Stratech. Enkel gegevens over het eindexamen, zoals niveau en of de kandidaat geslaagd dan wel gezakt is, kunnen worden opgenomen.

IW3

IW3 wordt gebruikt door Rotterdam, Utrecht, Nijmegen, Den Bosch en nog een aantal kleinere gemeentes. In dit systeem lijkt het meeste geregistreerd te kunnen worden; het is in ieder geval niet zo aan ISI en / of het Berichtenboek gebonden als Stratech. Gegevens over de inburgeraar kunnen worden geregistreerd, leertrajecten, overige activiteiten zoals kinderopvang en meer. Naast standaardoverzichten zoals vereist voor verantwoording naar de overheid is het systeem ook in te richten door de applicatiebeheerder van een gemeente. Er is dus een aantal velden dat door elke gemeente op een andere manier zal worden ingevuld. Het is mogelijk dat ook bijvoorbeeld taalaanbieders informatie in het systeem zetten. Zo kunnen voortgangsrapportages als documenten worden opgeslagen. Concluderend: er kan redelijk veel worden geregistreerd in het systeem, maar juist de gegevens die voor de benchmark NT2 van belang zijn (leerlast, rendement) zullen vermoedelijk niet door alle gemeentes op dezelfde manier geregistreerd worden.

GWS4all

Door de ontwikkelaars van GWS4all werd een wat concreter antwoord gegeven op de vraag naar het marktaandeel: de module inburgering is verkocht aan 112 gemeentes.

Er zit een wijziging aan te komen, maar op dit moment is het enkel mogelijk om begin- en eindniveau bij een traject te registreren (5 taalvaardigheden: luisteren, lezen, spreken, schrijven en gespreksvaardigheid), gerelateerd aan de niveaus van het Common European Framework of Reference (CEF). Er wordt echter geen aantal uren aan gekoppeld. Alleen de intensiteit van het traject kan worden ingevoerd. Niveaustappen (met toetsgegevens) en besteed aantal uren kunnen niet in het systeem worden opgenomen, maar dat zou wellicht een optie zijn voor de nieuwste release. Men gaat dan de mogelijkheid aanbieden om de maatschappelijke positie van een inburgeraar te registreren. Gemeentes kunnen die

‘maatschappelijke posities’ zelf definiëren. Ook hier is dus geen sprake van een standaard. Op het moment dat er contact was met de softwareontwikkelaar waren de velden van de nieuwste release nog niet in definitieve vorm beschreven. Ook GWS4all lijkt sterk gebaseerd te zijn op ISI en / of het Berichtenboek.

Conclusie over de bruikbaarheid van de cliëntvolgsystemen

Gegevens zoals die door de overheid gevraagd worden kunnen in alle drie de systemen worden opgenomen. Dit betekent dat er veel overeenkomsten zijn met ISI. Voor een groot deel van de voor benchmarkconstructie benodigde informatie geldt dat gemeentes zelf keuzes kunnen maken wat betreft de manier van registreren (IW3 en GWS4all). Daarom lijkt het nuttiger en informatiever om bij een paar gemeentes te informeren hoe zij de systemen gebruiken. Waar het gaat om aanvullende gegevens zoals specifiekere kenmerken van een inburgeraar (bijvoorbeeld motivatie, gezondheid), en faciliterende en beperkende voorwaarden zal er, als er wat over vastgelegd wordt, gewerkt worden met documenten die in een systeem ‘gehangen’ worden of met papieren dossiers, aangezien de bestudeerde cliëntvolgsystemen hier geen ruimte voor bieden. Er zal dus bij de te interviewen gemeentes nagegaan moeten worden of zij hierover gegevens registreren en zo ja, hoe ze dat doen. Het moge duidelijk zijn dat er voor het merendeel van de gewenste informatie geen sprake lijkt te zijn van beschikbaarheid ‘met één druk op de knop’. Hierdoor ontstaan naar verwachting problemen met het oog op de kwaliteitseis ‘efficiëntie’.

Welke cliëntvolgsystemen worden gebruikt? Navraag bij 14 gemeentes.

Er is naar 14 gemeentes (verschillend in grootte en ligging) gebeld met de vraag of er een cliëntvolgsysteem gebruikt wordt en zo ja welke. Dit is gedaan om na te gaan of de indruk die tot nu toe ontstaan was, namelijk dat er drie grote spelers zijn op de markt (GWS4all, IW3 en Stratech) juist is.

De aanname dat er sprake is van drie grote (in marktaandeel vergelijkbare) cliëntvolgsystemen die door gemeentes voor inburgering gebruikt worden lijkt op grond van de uitgevoerde belronde niet juist. GWS4all wordt het meest genoemd: door 7 gemeentes. Stratech wordt één keer genoemd en IW3 helemaal niet. Binnen deze (kleine) steekproef is het voornamelijk één cliëntvolgsysteem dat aangekocht is of wordt (als er wordt aangekocht): GWS4all. Dit verklaart mogelijk ook de eerdere bevinding dat alleen GWS4all een concreet aantal gemeentes kon geven waaraan de inburgeringmodule verkocht is. Overigens, het feit dat GWS4all genoemd is betekent niet per definitie dat dit systeem ook (al) gebruikt wordt. Vier van de zeven gemeentes die dit systeem bezitten hebben het systeem nog niet of slechts gedeeltelijk in gebruik (“we zijn nog bezig met het inrichten van het systeem”). Eenmaal wordt een ander systeem genoemd: EBB (van Horlings en Eerbeek), gebruikt door een kleinere gemeente. Dit systeem lijkt geen speciale inburgeringmodule te bevatten; de website vermeldt niets over inburgering, terwijl de onderdelen van het cliëntvolgsysteem wel opgesomd worden. De kans dat via dat systeem de voor de benchmark benodigde informatie te vinden is, is daarom klein.

Een aantal gemeentes gebruikt geen cliëntvolgsysteem, maar registreert met behulp van het Windows Officeprogramma Excel. Bij zes gemeentes werd Excel genoemd, al dan niet in combinatie met (een nog in te richten) cliëntvolgsysteem. Het zelf (provisorisch?) registreren in Excelsheets lijkt dus ook vaak aan de orde te zijn. Deze bevinding vergroot de kans op onderlinge vergelijkbaarheid van door gemeentes geregistreerde gegevens niet bepaald.

Tot slot zijn er enkele gemeentes die de registratie hebben uitbesteed. Harderwijk heeft daarover afspraken gemaakt met IGSD (Intergemeentelijke Sociale Dienst). De gemeente Zoetermeer heeft veel van de registratie uitbesteed aan een scholingsinstituut en houdt zelf

alleen ISI bij. De gemeente Veendam heeft een deel van de registratie uitbesteed aan Welzijnsorganisatie Kompaan (reïntegratiebedrijf) en houdt zelf zaken bij op een eigen excelsheet.

Al met al lijkt gezien de verschillende manieren van handelen de kans dat er vergelijkbare informatie te vinden is via wat gemeentes registreren niet heel groot.

Interviews met drie gemeentes over de registratie van gegevens van inburgeraars

Om een beter beeld te krijgen van de manier waarop gemeentes gegevens registreren en hoe ze met de cliëntvolgsystemen omgaan is bij drie gemeentes (Utrecht, Arnhem en Ede) een interview afgenomen. De gemeente Utrecht gebruikt het systeem IW3, Arnhem Stratech en Ede GWS4all. Een algemeen beeld is dat gemeentes veel moeten registreren en dat dat behoorlijk ingewikkeld kan zijn onder meer doordat het beleid nogal eens verandert. Er wordt vooral geregistreerd wat de overheid wil weten en wat de gemeente zelf van belang vindt. Dat wisselt per gemeente, maar vast staat dat lang niet alles uit de lijst van gegevens die nodig zouden zijn voor het construeren van de benchmark geregistreerd is in het cliëntvolgsysteem. Er zal, zoals het er nu naar uitziet, dossieronderzoek uitgevoerd moeten worden om alle gewenste gegevens te vinden en dat zou zeer intensief zijn, nog los van de vraag of de informatie die daaruit naar voren komt wel vergelijkbaar is. En ook in de dossiers is niet alles te vinden. Voor inhoudelijke onderwijskenmerken en docentkenmerken (van belang voor een zogenaamde brede benchmark) wordt door de drie gemeentes verwezen naar de taalaanbieder.

Rendement: Niveaumetingen

Bij het onderwerp niveaumetingen is onderscheid gemaakt tussen beginniveau, streefniveau, tussenniveau(s) en eindniveau. Alleen bij het eindniveau geldt voor de drie gemeentes dat ze het vastleggen in het cliëntstelsel. De niveaus die de gemeente Utrecht hanteert zijn echter volgens een oud systeem (0 t/m 5), terwijl Ede en Arnhem de Raamwerkniveaus gebruiken. De manier van registreren is niet hetzelfde, het beleid erachter ook niet helemaal.

Het beginniveau wordt niet op een vergelijkbare manier geregistreerd, maar het wordt wel door alle drie gemeten, door middel van een toets. Ede en Arnhem gebruiken hiervoor de TIWI, een intake-toets van Bureau ICE. Ede laat de toets door een ROC afnemen en beoordelen, Arnhem doet dat zelf. In Utrecht wordt de intake-toets ook binnen het eigen toetsbureau afgenomen, maar daar wordt de Intake-toets NT2 (Bureau ICE) gebruikt, of voor analfabeten de Begintoets Alfa van Cito. De resultaten van de Begintoets Alfa worden gegeven in speciale alfabetiseringsniveaus. Deze worden door Utrecht niet in het systeem geregistreerd (ze willen gaan uitzoeken of dat wel mogelijk is); een analfabeet heeft in IW3 niveau 0. De indicatie van het alfaniveau wordt wel opgenomen in het dossier. Ede neemt de uitkomst van de toets op in het systeem. Ze hebben daarvoor wel zelf de niveaus boven A2 moeten toevoegen. In Arnhem wordt het resultaat van de toets niet in het systeem opgenomen. Men vindt deze informatie belangrijker voor de school dan voor de eigen administratie, al wordt één en ander wel genoteerd in het dossier.

Streefniveau wordt ook door alle drie bepaald, maar niet vergelijkbaar geregistreerd. In Ede wordt het in GWS4all opgenomen en bepaald aan de hand van de TIWI, in Utrecht bepaalt men het streefniveau naar aanleiding van het intakegesprek en het profiel (oudkomer / nieuwkomer) van de inburgeraar en zet men dit ook in IW3 ('beoogd niveau', maar hierbij geen alfaniveaus) en in Arnhem is het profiel leidend voor de bepaling van het streefniveau. Dit is dus standaard. De gemeente neemt het streefniveau niet op in Stratech.

Tussenniveaus worden alleen in Utrecht echt gemeten en bijgehouden. Er wordt door de taalaanbieder in principe één keer per vier maanden een trajecttoets afgenomen (ook niveau 0 t/m 5) en de taalaanbieder registreert de uitslag dan in het toetsregistratie-onderdeel van IW3.

De informatie wordt ook opgenomen in het dossier. In Arnhem is er misschien iets over tussenniveaus terug te vinden in voortgangsrapportage van de taalaanbieders, maar dat kan niet met zekerheid worden gezegd. Het is voor de gemeente dus niet echt van belang om het te registreren. In Ede wordt tussentijds niet getoetst. Er is enkel een voortgangsrapportage waarin vermeld wordt of het goed gaat.

Het eindniveau wordt dan wel door alle drie geregistreerd, maar van een vergelijkbaar beleid is geen sprake. In Ede stelt men het eindniveau (ook voor schrijfvaardigheid) vast op A2, omdat men dat een minimaal te beheersen niveau vindt voor inburgeraars. Als het niet lukt worden oudkomers in een niet-leerbaarheidstraject geplaatst. In Arnhem hanteert men de niveaus zoals in de WI vastgesteld en verlengt men het traject tot iemand het haalt. Overigens wordt door deze gemeente opgemerkt dat het nog te kort dag is om daadwerkelijk uitspraken te kunnen doen over de trajecten: ze zijn in oktober 2007 gestart in deze hoedanigheid en aangezien een traject een jaar duurt zijn er nog nauwelijks gegevens over bekend op het moment van het interview. In Ede wordt iets dergelijks gezegd. In Utrecht wordt aan het einde van een traject (nog) een profieltoets afgenomen (met niveauaanduiding 0 t/m 5) door het eigen toetsbureau. Een traject duurt in Utrecht in principe een jaar, en als het nodig en mogelijk is met het oog op leerbaarheid, motivatie en dergelijke zaken verlengt men het traject met een half jaar (plustraject).

Concluderend: zowel met betrekking tot het gevoerde beleid, als tot niveauaanduiding en registratie daarvan is het feitelijk onmogelijk vergelijkbare informatie te verzamelen, als de informatie er al is. Registratie van niveaustappen zijn voor de gemeentes niet belangrijk, al wordt in Utrecht na het stellen van de vragen hierover wel even nagedacht over de voordelen ervan, bijvoorbeeld als het gaat om beoordeling van trajecten.

Leerlast: tijdmetingen

De vaste meetmomenten van de drie gemeentes zijn bij de start en bij de afronding van het traject. Alleen in Utrecht wordt tussentijds gemeten. Gegevens over niveaustappen en de tijd die een inburgeraar daarvoor nodig had zijn dus zeker niet bij de gemeentes Arnhem en Ede te vinden. Wel wordt bij alle drie de gemeentes een nagestreefd aantal uren vastgesteld. In Ede is er een aantal trajecten met een standaardlengte in maanden (en uren). Een cursist wordt op grond van de uitslag van de TIWI (daarin wordt ook leerbaarheid getoetst) geplaatst in één van die trajecten. Van die trajecten staat het aantal maanden en uren geregistreerd in GWS4all. Wel wordt erbij opgemerkt dat de nagestreefde trajectperiode vaak wordt overschreden omdat men te maken heeft met examentrajecten die op vaste tijden plaatsvinden. In Arnhem wordt naar aanleiding van de zogenaamde 'diagnose' (een uitgebreid rapport naar aanleiding van de intake) een scholingsplan opgesteld met een traject op maat. Gegevens hierover staan in dat scholingsplan en niet in Stratech. In Utrecht maakt men gebruik van een 'stroommatrix'. Hier staan de standaardpakketten per doelgroep in. Aan de hand van de stroommatrix wordt bepaald welk traject een cursist in gaat. Dit wordt opgenomen in het trajectplan en ook in IW3, al wil men met dat laatste stoppen omdat het geen informatie is die door de overheid vereist wordt.

Het aantal gerealiseerde uren wordt door Ede in GWS4all opgenomen naar aanleiding van de werkelijke einddatum. In Utrecht en Arnhem wordt er niet duidelijk iets over vastgelegd. De discrepantie tussen het aantal geprogrammeerde uren en het aantal geregistreerde uren wordt niet geregistreerd. Wel wordt er enigszins verzuim bijgehouden, maar dit gaat met name via de taalaanbieders die dat in rapportages doorgeven en door de gemeentes wordt voornamelijk nagegaan of het verzuim niet te veel oploopt. In Utrecht wordt er aan het einde van het traject wel een percentage in de rapportage opgenomen, maar dit komt dus niet in een invulveld in IW3 terecht. Er wordt met betrekking tot het gebruik van IW3 door

taalaanbieders opgemerkt dat het de bedoeling is dat ze rapportages erin zetten, maar dat nog niet iedereen goed is overgestapt. Deze informatie zal dus vermoedelijk niet compleet zijn. Specifiekere kenmerken van lessen zoals lestijd en het aandeel les in de groep versus individuele les worden alleen door Ede gegeven: een les duurt drie uur en er zijn drie dagdelen waarvan één dagdeel werken met de computer. Er wordt vrijwel altijd in de groep gewerkt, maar ook wel individueel, hetgeen ook op maat gebeurt: bijvoorbeeld hulp bij het opbouwen van het portfolio en begeleiding aan de computer. Ook hier lopen de trajecten lopen nog maar kort. In Utrecht wordt verwezen naar contracten zoals afgesloten met taalaanbieders. Deze gemeente registreert wel iets over het aantal uur per week per periode, maar daarbij wordt opgemerkt dat er nogal wat marge in zit. In Arnhem wordt verwezen naar het scholingsplan. Wel wordt verteld dat er een Werk- en een OGO-traject is. Onder het Werktraject vallen ook Staatsexamenkandidaten.

Al met al lijkt er over periodes en tijdmetingen ook geen eenduidige en vergelijkbare informatie beschikbaar te zijn. Ook dit aspect behoort niet tot de registratieprioriteiten van de gemeentes.

Kenmerken inburgeraars: opleidingsachtergrond

Het aantal jaren onderwijs in land van herkomst wordt alleen door Utrecht in het cliëntvolgsysteem geregistreerd. Hierbij wordt iedere school apart geregistreerd met het aantal jaren erbij, gerelateerd aan het Nederlandse onderwijssysteem. Restcategorie is 'geen onderwijs'. Informatie wordt aangenomen zoals beschreven door de inburgeraar.

In Ede registreert men alleen het opleidingsniveau. Wel wordt met betrekking tot dit onderwerp in de reïntegratiemodule van GWS4all meer informatie opgenomen waarbij onderwijstype wordt aangevinkt. Men vraagt naar diploma's, maar als die er niet zijn wordt maar aangenomen wat de inburgeraar zegt. In Arnhem neemt men één en ander (ook arbeidsverleden) op in het dossier, dit valt samen met de 'diagnose' en wordt vrij uitgebreid uiteengezet. Het wordt als een soort curriculum vitae vastgelegd. Ook hier vraagt men naar diploma's, maar meer in het belang van de inburgeraar met het oog op vrijstellingen et cetera. Men vraagt wel door over het soort onderwijs.

Het lijkt erop dat er in principe tamelijk eenduidige informatie te verkrijgen zou moeten zijn over dit onderwerp, maar men registreert het wel op verschillende plaatsen. Ook zijn er vraagtekens te plaatsen bij de betrouwbaarheid van de informatie wanneer simpelweg wordt aangenomen wat de inburgeraar zegt. Soms lijkt er echter geen keus: bijvoorbeeld in het geval van een vluchteling die zijn diploma kwijt is. In Ede wordt bij dit onderwerp nog opgemerkt dat er ook één en ander op te maken valt uit de uitkomst van de TIWI.

Kenmerken inburgeraars: overig

De volgende gegevens zijn bij de drie gemeentes in ieder geval in de cliëntvolgsystemen te vinden:

- oudkomer / nieuwkomer
- leeftijd
- geslacht
- land van herkomst

Utrecht neemt de gesproken talen ook op in IW3, waarbij aangegeven kan worden of het de moedertaal betreft en of de kandidaat de taal ook kan lezen en schrijven.

Arnhem en Ede hebben dit in het dossier geregistreerd.

Brede benchmark: onderwijskenmerken

Voor het opstellen van een brede benchmark zoals toegelicht in de inleiding zijn onder meer aanvullende gegevens over het aangeboden onderwijs nodig. Het blijkt dat gemeentes die informatie niet of slechts in beperkte mate beschikbaar hebben. Er wordt verwezen naar taalaanbieders en / of contracten. In Ede is er een aantal standaardafspraken gemaakt over het onderwijs, meer informatie moet gehaald worden bij de taalaanbieders. In Utrecht weet men heel globaal wat algemene dingen die men ook communiceert naar een potentiële cursist. Ze noemen het ook beleid zich er niet te veel mee te bemoeien. Er worden vanuit de gemeente voornamelijk eisen gesteld aan intensiteit en prestaties. Ook in Arnhem kan er weinig over worden gezegd, behalve dat er OGO- en Werktrajecten zijn. Voor meer informatie zouden de taalaanbieders dus benaderd moeten worden.

Brede benchmark: specifieke cursistkenmerken

Hierover wordt niets in de cliëntvolgsystemen geregistreerd (voor zover dat al mogelijk zou zijn). Er wordt verwezen naar dossiers. Arnhem lijkt daarvan het meeste werk te hebben gemaakt.

Hier is dossieronderzoek dus vereist, hetgeen niet tegemoet lijkt te komen aan de kwaliteitseis efficiëntie.

Brede benchmark: docentkenmerken

Hierover wordt niets vastgelegd. In Utrecht wordt gezegd dat het vroeger wel meer werd gedaan toen de trajecten nog alleen bij het ROC liepen. Tegenwoordig is het hoe dan ook lastiger, zo wordt verteld, aangezien cursisten vaak in één jaar al meerdere docenten hebben. In Ede wordt nog wel verwezen naar de offerte, waarin een aantal algemene eisen wordt gesteld aan docenten. Vanuit de gemeente wordt hier niet op gecontroleerd, dus men neemt maar aan dat deze eisen worden nagekomen.

Het gebruik van de cliëntvolgsystemen

Voorwaarde voor het gebruik van gegevens uit een cliëntvolgsysteem is dat de gegevens compleet en zorgvuldig geregistreerd zijn. Een algemeen probleem bij de gemeentes is dat het beleid zoals door de overheid bepaald vaak verandert, en daarmee tevens de trajecten en wat er al dan niet geregistreerd moet worden. In Utrecht wordt ook gezegd dat IW3 vaak gewijzigd wordt. Utrecht lijkt relatief het meeste te registreren in het cliëntvolgsysteem, maar het is soms onduidelijk waar gegevens staan, zoals toetsgegevens en –resultaten, die op twee plaatsen kunnen staan. Daar bestaan dan nogal eens misverstanden over, zo wordt verteld: wie wat waar invult.

Wie de gegevens invoeren in het systeem verschilt per gemeente. In Utrecht zijn dat trajectbegeleiders, secretariaatmedewerkers en ook taalaanbieders. Mensen worden op het systeem ingewerkt, waarbij de vraag wie degene is die inwerkt een grote rol blijkt te spelen (werkt die persoon al lang met het systeem, hoe scherp is hij/zij op wijzigingen?) en wie degene is die ingewerkt wordt (worden er vragen gesteld als het niet duidelijk is, of wordt er gewoon wat geprobeerd?). Inhoudelijke kennis van bijvoorbeeld de WI is niet bij iedereen aanwezig. Na een volgende wijziging in IW3 wil men trainingen gaan geven aan de trajectbegeleiders. Eén en ander lijkt wel te verklaren dat er soms misverstanden ontstaan. De informatie wordt er niet betrouwbaarder op. Er is overigens wel een protocol waarmee gewerkt wordt. In Ede wordt door de applicatiebeheerders een training GWS4all gegeven en worden nieuwe medewerkers door collega's ingewerkt op het inhoudelijke aspect. De personen die in Ede met het systeem werken zijn van de administratie. Daarnaast registreren consultants trajectplannen en dergelijke en financieel consultants houden zich bezig met geldzaken. In Arnhem werken alleen de casemanagers met het systeem, dus diegenen die ook

het dichtst bij ‘het vuur’ zitten. Men wordt ingewerkt en er wordt scholing aangeboden over het systeem, de Wi, beschikkingen, et cetera.

Bij alle drie de gemeentes zijn er controles op wat ingevoerd wordt. Bijvoorbeeld door meldingen of een speciale controlemodule. Het lijkt voornamelijk te gaan om meldingen uit het systeem, dingen die fout gaan (bijvoorbeeld betalingen die niet uitgevoerd kunnen worden omdat de gegevens niet compleet zijn) en controle van dossiers die toch al voorbij komen.

Geheel waterdichte controle lijkt dan ook niet mogelijk.

Velden die zelf gedefinieerd zijn, zijn bijvoorbeeld: niveaus (uitbreiding naar boven, Ede) en traject en opdracht (soort, Arnhem). In Utrecht zegt men vrij veel te hebben aangepast, ook omdat het de bedoeling is dat het protocol voor alle doelgroepen gelijk is.

Over de eenduidigheid van de velden zijn Arnhem en Ede unaniem: ja, dat is het geval. In Utrecht vindt men dat een goede vraag. Onlangs bleek uit een rapport dat een bepaald veld (‘traject beëindigd’) op meerdere manieren te interpreteren was.

Concluderend: vooral in Utrecht lijkt het cliëntvolgsysteem in het gebruik nog wel eens lastig. Een juiste manier van invullen wordt echter wel belangrijk gevonden, gezien het feit dat men trainingen gaat verzorgen. Ook worden er in alle drie de gemeentes controles uitgevoerd op het gebruik van het systeem. De manier waarop het systeem gebruikt wordt lijkt een minder groot probleem te zijn dan de beschikbaarheid van de gewenste informatie.

Conclusie

De voorlopige conclusie die voortvloeit uit de bestudering van de cliëntsystemen is dat het niet eenvoudig, zo niet ondoenlijk zal zijn de gewenste informatie te verzamelen via de registratiesystemen van gemeentes. Het voornaamste probleem lijkt het gegeven dat de gemeentes aan het (zorgvuldig en volledig) registreren van deze informatie geen hoge prioriteit geven. Het gaat dan met name om niveaustappen en de daaraan verbonden tijdsduur. Dit zijn echter precies de fundamenteën waarop de benchmark gebaseerd zou moeten zijn, zodat het effect van de succesfactoren daartegen afgezet kan worden. Een beeld van het aantal taalniveaus en niveaustappen voor iedere doelgroep kan ook niet ontleend worden aan de registratie van gemeentes.

Simpelweg uitdraaien van informatie vanuit cliëntvolgsystemen lijkt niet tot de mogelijkheden te behoren, omdat daar weinig tot niets van de benodigde informatie te vinden zal zijn. Hiervoor zal men afhankelijk zijn van dossiers en eventueel gegevens van taalaanbieders. Ook de navraag bij veertien gemeentes naar het gebruikte cliëntvolgsysteem stemt niet echt optimistisch voor deze opzet: als men al een cliëntvolgsysteem heeft aangeschaft, is het maar de vraag of het (volledig) gebruikt wordt. Eigen systemen in Excel betekent dat bij iedere gemeente die het zo aanpakt nagegaan moet worden hoe dat systeem in elkaar zit.

Om bovengenoemde redenen (lage prioriteit voor de gewenste gegevens, weinig te vinden in cliëntvolgsystemen en allerlei gebruikte systemen om (provisorisch?) in te registreren) lijkt het niet wenselijk om in de onderzoeksfase de benodigde informatie te verzamelen via de cliëntvolgsystemen van gemeentes.

2.3 Primaire gegevensverzameling via empirisch onderzoek

2.3.1 Recentelijk uitgevoerd onderzoek

In Nederland en daarbuiten is nog weinig empirisch onderzoek gedaan waarin primaire gegevens over leerlast en rendement verzameld zijn. Nederlandse voorbeelden waarbij

onderzoekers zelf leerlast- en toetsgegevens verzamelden, zijn Abbenhuis e.a. (1996) en Huizinga en Van Kalsbeek (1996). Bij de uitvoering van dit onderzoek deden zich grote onderzoekstechnische problemen voor. Vandaar dat we ons vooral richten op buitenlands onderzoek (Watt & Lake, 2004; Condelli e.a., 2007). Het accent ligt hierbij niet op het beschrijven van de onderzoeksresultaten, maar op het inventariseren van de onderzoekstechnische mogelijkheden en problemen die bij de ontwikkeling van een benchmark NT2 kunnen faciliteren dan wel bemoeilijken.

Het onderzoek van Abbenhuis e.a. (1996)

Abbenhuis e.a. (1996) ontwikkelden een zogeheten leerlastindicator die moest beschrijven onder welke condities cursisten bepaalde taalniveaus behalen. Leerlast werd daarbij omschreven als “het aantal onderwijsuren dat gebruikt is om van het ene niveau van taalvaardigheid naar het andere niveau te gaan” (p. 9). In het onderzoek is behalve over de leerlast en de niveaustappen een grote hoeveelheid gegevens verzameld over kenmerken van de cursist, het onderwijs en het onderwijsprogramma. De onderzoekers maakte daarbij zowel gebruik van door de instellingen verzamelde toetsgegevens als van zelf afgenomen toetsinstrumenten. Voor dit laatste werd uit bestaande NT2-toetsen een nieuw toetsinstrument samengesteld. Op basis van een responsegroep van 310 cursisten bepaalden de onderzoekers het percentage cursisten dat na 150, 300, 450, 600 en meer dan 600 uur al dan niet een niveaustap maakte. Uit het onderzoek bleek onder meer dat grote groepen cursisten geen meetbare progressie boekten, zelfs niet na vele honderden uren. Zo bleek ook na meer dan 600 uur ruim een kwart van de cursisten die op taalniveau 0 instapten nog geen niveaustap gerealiseerd te hebben. Cursisten die de stap van taalniveau 0 naar 1 realiseerden, hadden daartoe gemiddeld meer uren les gebruikt dan cursisten die de stap van taalniveau 1 naar 2 (en van 2 naar 3) zetten. De grootste toename van het aantal cursisten dat een stap zette, vond plaats vóór 450 uur (basiseducatie) of 300 uur (vavo). Meer uren les leidde in verhouding tot een relatief geringe verhoging van het percentage cursisten dat een stap zette. Dit wijst er volgens Abbenhuis e.a. (1996) op dat “cursisten die na 450 resp. 300 uur nog geen stap hebben gezet, niet erg gebaat zijn bij meer van hetzelfde” (p. 2 van Bijlage 2).

Bij de uitvoering van het onderzoek deed zich een groot aantal problemen voor. Hieronder een bloemlezing uit de onderzoekstechnische problemen waarmee de onderzoekers te kampen hadden (en die in het onderzoek naar de constructie van een benchmark NT2 opgelost moeten worden):

- Veel instellingen waren niet bereid tot deelname aan het onderzoek.
- De uitval van cursisten en de hoeveelheid ontbrekende of inconsistente gegevens bij de leerlast- en taalvariabelen waren zeer hoog.
- Het ter beschikking stellen van de gegevens over het aantal contacturen kostte de instellingen veel meer werk en tijd dan de onderzoekers hadden ingeschat.
- Binnen de groepen cursisten van een instelling bestonden grote verschillen in onder meer startniveau, streefniveau en bereikt eindniveau. Instellingen toetsten cursisten over het algemeen niet op het moment dat zij daar aan toe waren, maar op het moment dat de meeste cursisten uit de groep zo ver leken te zijn of volgens het programma zouden moeten zijn. Door de grote breedte aan nagestreefde taalniveaus binnen de groepen en het gering meetbereik van de gebruikte toetsen konden niet altijd optimale schattingen van het niveau van de cursist worden verkregen.
- De instellingen toetsten niet altijd alle vier taalvaardigheden. Voor het onderdeel spreken kwamen veel minder toetsgegevens terug dan voor de onderdelen luisteren, spreken en schrijven. Een verdere complicatie was dat de instellingen de vier taalvaardigheden bij eenzelfde cursist vaak op twee of drie verschillende taalniveaus toetsten waardoor het

globale taalniveau van de cursist niet goed bepaald kon worden. Daarnaast speelde het probleem dat de door de instellingen gebruikte toetsen vaak sterk verouderd waren (en dat de antwoorden op de vragen bij de cursisten bekend waren). De onderzoekers stelden dan ook voor om voor eventueel vervolgonderzoek toetsen samen te stellen die meer niveaus bestrijken en waarvan de antwoorden niet bij de cursisten bekend zijn.

- Als gevolg van in- en uitstroom van cursisten hanteerden instellingen binnen een en dezelfde groep cursisten verschillende peilmomenten voor de toetsing en de telling van het aantal gerealiseerde contacturen. Dit beperkte de mogelijkheden om leerlast eenduidig aan rendement te relateren.
- De van de instellingen afkomstige toetsgegevens en de zelf verzamelde toetsgegevens werden op zeer gedetailleerd niveau verzameld (“ruwe scores”). Om de gegevens over de verschillende toetsen en de beide methoden van gegevensverzameling vergelijkbaar te maken, moesten deze gegevens eerst “veralgemeniseerd” worden voordat statistische analyse kon plaatsvinden. De gedetailleerde scores per vaardigheid per toetsmoment werden daartoe “ingedikt” in een beperkt aantal niveaus en niveaustappen (de genoemde vaardigheidsladder).
- Per niveaustap gingen de onderzoekers na hoeveel uren les daarvoor gemiddeld nodig waren. De analyse van de relatie tussen de snelheid waarmee cursisten niveaustappen maakten en de onderzochte kenmerken van de cursist, de docent en het programma leverde resultaten op die volgens de auteurs vaak niet eenduidig te interpreteren waren. Daardoor konden geen profielen van succesvolle cursisten en cursussen worden opgesteld.

De ontwikkelde leerlastindicator is geëvalueerd tijdens een expertmeting. Een belangrijk kritiekpunt was het ontbreken van “‘voorspellingen’ over de kans ‘wie’ ‘welk niveau’ in ‘welke periode’ onder ‘welke condities’ zou kunnen bereiken, terwijl dat juist de essentie van een leerlastindicator zou moeten zijn” (Van der Linden, 1996, p. 4).

Het onderzoek van Huizinga en Van Kalsbeek (1996)

Huizinga en Van Kalsbeek (1996) onderzochten onder meer de taalvaardigheid van cursisten na 50, 150 en 250 contacturen in een intensieve NT2-cursus. Zij namen daartoe een aantal leerstofonafhankelijke toetsen af die functionele taalvaardigheid en een aantal deelvaardigheden maten. De onderzochte groepen cursisten bleken duidelijk te verschillen in beginniveau, eindniveau en leersnelheid. Het eindniveau van de cursisten bleek sterk achter te blijven bij het streefniveau. De auteurs rapporteren dat zij tijdens het onderzoek te maken hadden met een grote uitval van cursisten. Daarnaast gaven er vaak meerdere docenten aan één groep les en werden de groepen regelmatig opnieuw ingedeeld. Daardoor kon er geen relatie worden gelegd tussen enerzijds kenmerken van de docent en het onderwijs en anderzijds de toetsresultaten van de cursisten.

Het onderzoek van Condelli e.a. (2007)

Condelli e.a. (2007) volgde twee cohorten van volwassen *English-as-a-second-language literacy (ESLL)* cursisten voor een periode van tien maanden. Een van de doelen van de studie was het identificeren van effectieve vormen van instructie voor *ESLL* cursisten. Aan het onderzoek namen 495 cursisten van 38 klassen deel.

De leerlast van het onderwijsprogramma en het bereikte taalvaardigheidsniveau van het eerste cohort cursisten werd op drie momenten gemeten: bij de start van het programma en drie en negen maanden daarna. De duur en intensiteit van het onderwijsprogramma werd vastgesteld aan de hand van het aantal tot dan toe gevolgde lessen, het totaal aantal deelgenomen weken, de proportie bijgewoonde lessen en het gemiddeld aantal lessen per week.

De taalvaardigheidstoetsing werd onder gecontroleerde omstandigheden op de scholen uitgevoerd. De gegevensverzameling werd uitgevoerd door een medewerker van de school zelf wiens activiteiten uit het onderzoeksbudget werden bekostigd, de zogeheten *study liason*. Deze ontving daartoe een tweedaagse training. Na elk meetmoment werd de kwaliteit van de ingestuurde gegevens door de onderzoekers gecheckt op volledigheid en consistentie. Als een extra check werd elke school drie à vier keer door zogeheten *monitors* bezocht.

Ondanks alle maatregelen om volledigheid en consistentie te bevorderen, had het project sterk te leiden onder uitval. Van de 38 klassen namen er 17 klassen alleen deel in het eerste jaar van de studie, 4 alleen in het tweede jaar en 17 zowel in het eerste en tweede jaar. Van de 495 cursisten die deelnamen aan de eerste meting namen er 365 deel aan de tweede meting en 263 aan de derde meting.

De groei van de luister-, spreek-, lees- en schrijfvaardigheid van de cursisten werd vastgesteld met behulp van gestandaardiseerde taaltoetsen (en daarnaast met meer informele toetsvormen). De zorgvuldige opzet en uitvoering van het onderzoek maakte het mogelijk om de relatie tussen leerlast en rendement nauwkeurig vast te stellen. Omdat ook aanvullende gegevens verzameld werden, kon worden nagegaan in hoeverre de leerwinst verschillend was voor verschillende typen cursisten, klassen en vormen van instructie. De relatie tussen leerlast, instructiekenmerken en leerwinst werd vastgesteld met een meerniveau model voor het analyseren van leerwinst. De volgende instructiekenmerken bleken daarbij positief gerelateerd aan groei in taalvaardigheid:

- Verbinding met de buitenwereld. In klassen waarin docenten een relatie legden met de wereld buiten de klas behaalden de cursisten betere leesresultaten.
- Gebruik van de taal van de cursist voor uitleg en verduidelijking. In klassen waarin leraren de eigen taal van de cursisten gebruikten om begrippen te verduidelijken en instructies te geven behaalden de cursisten meer leerwinst op het terrein van leesvaardigheid en mondelinge taalvaardigheid.
- Gevarieerde oefen- en interactiepraktijk. Als de docent concepten op verschillende wijzen (*modalities*) instrueerde en voldoende gelegenheid gaf tot gevarieerde oefening, resulteerde dat in een snellere groei van de mondelinge taalvaardigheden.
- Nadruk op mondelinge communicatie. Als docenten mondelinge communicatievaardigheden expliciet onderwezen, behaalden de studenten een grotere leerwinst.

Het onderzoek van Watt en Lake (2004)

In een tweejarige studie volgden Watt en Lake (2004) de ontwikkeling van de taalvaardigheid van 612 recentelijk gearriveerde en doorgaans hoog opgeleide immigranten. De taaltoetsing werd uitgevoerd met de *Canadian Language Benchmark Assessment Tool (CLBA)*. De CLBA wordt in Canada breed ingezet bij de start en afronding van taaltrajecten voor nieuwe immigranten. De CLBA meet de toename van de taalvaardigheid van de cursist op een schaal met acht niveaus (benchmarks genoemd). In het eerste onderzoeksjaar was de toetsing gericht op de taalniveaus 1 t/m 5 en in het tweede jaar op de taalniveaus 4 t/m 8. Het instrument is niet geschikt om de groei van de taalvaardigheid binnen deze niveaus vast te stellen.

De taaltoetsen werden onder gecontroleerde omstandigheden op de scholen afgenomen door een team van vijf gecertificeerde externe assessoren waarbij de betrouwbaarheid van de meting zorgvuldig gemonitord werd. De groei van de taalvaardigheid van de cursisten werd in intervallen van ongeveer 250 lesuren vastgesteld. Daartoe was het team van assessoren tot twee weken op de deelnemende school aanwezig. Daar hadden zij de beschikking over een eigen toetsruimte en kregen zij toegang tot de klassen. Van de CLBA waren vier parallelle versies beschikbaar zodat test artefacten ten gevolge van herhaalde afname van dezelfde toets konden worden voorkomen. De assessoren besteedden speciale aandacht aan een zorgvuldige

registratie van het aantal gevolgde lesuren. De zelfrapportages van het aantal gevolgde lesuren Engels in het land van herkomst werden als onvoldoende nauwkeurig beschouwd en buiten de analyse gelaten. Het opleidingsniveau werd gemeten als het aantal jaren onderwijs in het land van herkomst (en vervolgens alleen voor rapportagedoeleinden gecategoriseerd tot de niveaus 0-7, 8-12, 13-16 en 17 of meer). Daarnaast werden bij de cursisten achtergrondgegevens verzameld, zodat de vergelijkbaarheid van de onderzoeksgroep met de landelijke populatie van immigranten kon worden vastgesteld. Door uitval en mobiliteit (verhuizen) van cursisten waren er aan het einde van het eerste onderzoeksjaar slechts van 125 van de 612 cursisten voor drie toetsmomenten volledige gegevens beschikbaar terwijl er voor ongeveer 300 cursisten gegevens van twee metingen voorhanden waren.

De toename van de taalbeheersing van de immigranten aan het einde van het eerste onderzoeksjaar werd gerapporteerd in relatie tot het aantal lesuren dat nodig was om een niveaustap te maken op de CLBA-schaal van 1 t/m 8. De relatie tussen het aantal lesuren en de leerwinst bleek sterk afhankelijk van het startniveau (de eerste meting met de CLBA) en het opleidingsniveau van de cursist (gemeten als het aantal jaren genoten onderwijs in het land van herkomst). Niet alleen het gemiddeld startniveau verschilde van opleidingsniveau tot opleidingsniveau, maar ook de leersnelheid en het voorspelde eindniveau. Van de cursisten met een lager opleidingsniveau bleek de leersnelheid gemiddeld aanmerkelijk lager dan van degenen met een hoger opleidingsniveau.

Ander onderzoek naar leerlast en taalvaardigheid

Onvolledige en inconsistente gegevens is een van de grootste problemen van onderzoek naar (onder meer) de relatie tussen leerlast en taalniveau bij volwassen *literacy students* (Comings, Soricone & Santos, 2006; Young, Fleischman, Fitzgerald & Morgan, 1994; Bingman, Evert & Smith, 1999). Een andere uitdaging is dat de veronderstelde relatie tussen leerlast en groei van taalvaardigheid lang niet altijd kan worden aangetoond (Alamprese, Tao & Price, 2003; Balmuth, 1987; Brooks e.a., 2001; Comings, Soricone & Santos, 2006; Kruidenier, 2002; Porter, Cuban & Comings, 2005; White & Hoddinott, 1991).

2.3.2 Bruikbaarheid voor benchmarkconstructie

Primaire gegevensverzameling door middel van empirisch onderzoek is doorgaans tijdrovender, arbeidsintensiever en kostbaarder dan secundaire analyse van bestaande bestanden en rapportages (zoals de huidige Monitor inburgeraars). Een ander evident nadeel is dat het meestal veel langer duurt voordat de resultaten beschikbaar komen. Een belangrijk voordeel van primaire gegevensverzameling is dat de gegevensverzameling beter kan worden toegesneden op de specifieke onderzoeksvraag. Andere pluspunten zijn de gegevensverzameling op individueel niveau, de betrouwbaarheid van de meetinstrumenten, de controle over de afname- en beoordelingscondities, de hogere kwaliteit van de gegevens (bijvoorbeeld minder *range restriction*) et cetera. Primaire analyse levert echter lang niet altijd meer complete en minder inconsistente gegevens op. Het tot op heden uitgevoerde onderzoek naar de relatie tussen leerlast en rendement heeft sterk te leiden onder het probleem van onvolledige en inconsistente gegevens. Bij de gegevensverzameling doen zich vergelijkbare problemen voor met uitval en non-response. Echter, omdat de onderzoeker de gegevens zelf verzamelt, kan het voorkomen van non-response meer aandacht krijgen (onder meer door cursisten voor hun deelname aan het onderzoek passend te belonen). In hun overzichtsstudie van onderzoek naar *adult literacy students* melden Comings en Soricone (2007) echter dat “some well-resourced studies are able to maintain their sample” (p. 18). Mits ruime middelen voorhanden zijn, het onderzoek goed wordt uitgevoerd en deelnemers

worden betaald voor hun inspanningen (Reder, 2006), is een responsepercentage van 70% of hoger haalbaar.⁹

Daarnaast is het mogelijk de consistentie en validiteit van de gegevens te verhogen via triangulatie (waarbij secundaire analyse een belangrijke bron van informatie kan zijn).

2.4 Keuze voor primaire dan wel secundaire gegevens

De voor- en nadelen van primaire en secundaire gegevensverzameling tegen elkaar afwegend kan een duidelijke keuze voor het verzamelen van primaire gegevens worden gemaakt. Een plan van aanpak voor een mogelijke opzet van de ontwikkeling van een benchmark NT2 is beschreven in hoofdstuk 4 van deze publicatie.

⁹ Bij de voorbereiding van het interventieonderzoek naar het effect van direct literacy instruction gaan Cronen, Silver-Pucuilla & Condelli (2006) uit van 20% uitval over een periode van zestien weken.

Bijlage 3

Het gebruik van een benchmark NT2 in het kader van leerlastonderzoek

Inleiding

In hoofdstuk 1 is een gefaseerd model geschetst van leerlastonderzoek. Betoogd is dat een geleidelijke overgang van ontwikkelingsgericht naar (quasi-)experimenteel leerlastonderzoek wenselijk is om het NT2-onderwijs meer in overeenstemming te brengen met de succesfactoren uit het theoretisch kader. Het ministerie van VROM is voornemens kleinschalig ontwikkelingsgericht leerlastonderzoek te stimuleren naar het effect van potentiële kenmerken van effectief NT2-onderwijs. Dit onderzoek moet leiden tot evidence-based kennis over effectief NT2-onderwijs en bijdragen aan de vernieuwing en verbetering van het NT2-onderwijs. Vanwege de grote logistieke en financiële implicaties kunnen in relatief kleinschalige settings ontwikkelde interventies pas grootschalig worden “uitgerold” in het NT2-onderwijs als de effectiviteit onomstotelijk is vastgesteld. Daartoe is een (quasi-)experiment het meest geëigende middel. In deze bijlage wordt nader ingegaan op het gebruik van een benchmark voor onderzoeksdoeleinden. Hierbij wordt een onderscheid gemaakt tussen gebruik in het kader van “harde” (quasi-)veldexperimenten (paragraaf 3.1) en “zachte” ontwikkelingsgerichte pilots (paragraaf 3.2).

3.1 Gebruik van een benchmark in (quasi-)experimenteel leerlastonderzoek

De benchmark dient bij te dragen aan het bepalen van de mate waarin een positief, neutraal of negatief effect van een veldexperiment daadwerkelijk aan de interventie kan worden toegeschreven en niet aan andere factoren. De effectiviteit van de interventie wordt bepaald aan de hand van een vergelijking van de bereikte leerwinst in de interventiegroep met de voorspelling op basis van de benchmark. Overeenkomstig de logica van *counterfactual evidence* moet de NT2 benchmark een voorspelling doen van de leerwinst die de NT2-interventiegroep bereikt zou hebben in de hypothetische situatie dat deze groep niet aan de interventie zou hebben meegedaan. De benchmark fungeert hiermee als een proxy voor zowel een pretest als een onafhankelijke controlegroep. In zijn meest eenvoudige vorm kan dit impliciete onderzoeksontwerp op een tijdslijn als volgt worden weergegeven (waarbij we ervan uitgaan dat onderzoekers eerst een benchmark ontwikkelen en die vervolgens gebruiken ter voorspelling van de leerwinst in de interventiegroep onder de aanname dat deze groep gemiddeld onderwijs had gekregen van een gemiddelde docent op een gemiddelde onderwijsaanbieder). Hierbij representeert X de interventie en O_1 de nameting in de interventiegroep). Met de onderbroken lijn (---) is aangegeven dat de benchmark- en interventiegroep niet via random toewijzing van cursisten en/of klassen tot stand zijn gekomen.

Benchmark

X

O_1

Het onderzoeksdesign met alleen een benchmark, een interventiegroep en een nameting in deze groep heeft enkele voordelen. Daartoe behoren de eenvoud, de geringe kosten, de beperkte belasting van het onderwijs en de snelheid waarmee resultaten beschikbaar komen. Omdat de interventie aan de effectmeting voorafgaat, is er in vergelijking met een cross-sectioneel ontwerp volstrekte duidelijkheid over de temporele relatie van interventie en eventueel effect. Omdat er in de interventiegroep maar één keer gemeten wordt en de benchmarkgroep een andere is dan de interventiegroep, kunnen *pretest artefacten* als alternatieve verklaring voor het effect van de interventie (of het ontbreken ervan) worden uitgesloten. Vanwege het ontbreken van een expliciete pretest in de interventiegroep zelf en de afwezigheid van een onafhankelijke controlegroep hebben alle andere alternatieve verklaringen voor een verschil tussen X en O_1 (of het ontbreken ervan) echter vrij spel (Shadish, Cook & Campbell, 2002; Fraenkel & Wallen, 2003). Tot de belangrijkste factoren die de interne validiteit en hiermee de bewijskracht en de geloofwaardigheid van het interventieonderzoek bedreigen behoren:

- *Selection bias*. De samenstelling en het beginniveau van de benchmarkgroep kan verschillen van de interventiegroep waardoor een verschil tussen het waargenomen en voorspeld taalniveau mogelijk ten onrechte aan de interventie wordt toegeschreven. *Selection bias* kan zich bijvoorbeeld voordoen als de steekproef van de benchmarkgroep landelijk representatief is en de interventiegroep via zelfselectie tot stand is gekomen of de onderzoeker de vergelijkingsgroep onzorgvuldig heeft samengesteld. Verder kunnen er zich veranderingen voordoen in de kwalitatieve samenstelling van de instroompopulatie die mogelijk groter worden naarmate er meer tijd verstrijkt tussen de constructie en de daadwerkelijke toepassing van de benchmark. Hierbij zij opgemerkt dat de kans op *selection bias* groter wordt naarmate de populatie heterogener en instabieler is/wordt.
- *Regression artifacts*. Als vooral betere of juist slechtere docenten (of cursisten met relatief lage of juist hoge scores) zich voor de interventie opgeven (of worden geselecteerd) is de kans groot dat zij aan het einde van de interventie lagere respectievelijk hogere prestaties behalen (en de onderzoeker het verschil op de effectmaat ten onrechte interpreteert als bewijs voor de effectiviteit respectievelijk ineffectiviteit van de interventie).
- *Instrumentation*. Als in de benchmarkgroep andere meetinstrumenten worden gebruikt of dezelfde instrumenten op een andere manier of tijdstip worden afgenomen dan in de interventiegroep kan verandering van de meting als alternatieve verklaring voor het effect van de interventiegroep niet worden uitgesloten.
- *History*. De benchmark is ontwikkeld voordat de interventie van start ging (tenzij men uiteraard benchmark en interventieonderzoek parallel zou plannen). Tijdens de interventie kunnen zich gebeurtenissen voordoen die niet met de interventie samenhangen en die zich evenmin ten tijde van de constructie van de benchmark voordeden, maar die wel van invloed zijn op het geconstateerde verschil tussen benchmark en interventiegroep (of het ontbreken ervan).
- *Mortality*. De vergelijkbaarheid kan worden ondergraven door uitval of differentiële uitval (als de uitval in de interventiegroep anders is dan in de benchmarkgroep). Als er sprake is van uitval, moet de onderzoeker kunnen aantonen dat de uitval niet verschilt van de niet-uitval en dat de uitval in de interventiegroep niet verschilt van die in de benchmarkgroep. Zonder een pretest op dezelfde schaal als de nameting is dat niet goed mogelijk.
- *Maturation*: Cursisten in de interventiegroep kunnen “van nature” beter gaan presteren onafhankelijk van de interventie, bijvoorbeeld ten gevolge van simpelweg ouder worden.

In de literatuur over experimentele en quasi-experimentele designs lijkt een ontwerp met uitsluitend een benchmark en een nameting niet voor te komen. Wel heeft het ontwerp enige gelijkenis met het zogeheten “*regression-extrapolation-contrast design*” (REC; Shadish,

Cook & Campbell, 2002). Hierbij wordt de score op een posttest in een interventiegroep vergeleken met de voorspelde score op basis van een eerder in die groep geconstrueerde benchmark. Een voorbeeld is het onderzoek van Cook e.a. (1975) naar het effect van het kijken naar Sesam Straat op de leerprestaties van leerlingen. De onderzoekers namen eerst een pretest af en zes maanden later een posttest. Tijdens de pretest werd behalve de leerprestaties ook de leeftijd in maanden vastgesteld. Dit stelde de onderzoekers in staat om een schatting te verkrijgen van de verwachte leerwinst voor elke maand dat de leerling ouder werd. De schatting van de gemiddelde leerwinst per maand werd vervolgens gebruikt om te voorspellen hoeveel leerwinst de leerling zou hebben geboekt in de periode van zes maanden tussen de pretest en de posttest (als deze niet naar Sesam Straat zou hebben gekeken). Deze voorspelling (die verondersteld werd niet beïnvloed te zijn door het kijken naar Sesam Straat omdat deze alleen op de pretest gebaseerd was) werd vervolgens vergeleken met de score op de posttest na afloop van de interventie (die naar verondersteld werd wel door het kijken naar Sesam Straat beïnvloed zou zijn). Shadish, Cook en Campbell (2002) noemen het REC ontwerp in hoge mate problematisch. Niet alleen vanwege de vele bedreigingen van de interne validiteit, maar ook omdat voor stabiele voorspellingen betrouwbare instrumenten en grote steekproeven nodig zijn. Daarnaast merken zij op dat voor het REC design vaak geaggregeerde gegevens worden gebruikt zonder dat rekening gehouden wordt met de hiërarchische datastructuur (met individuen genest binnen hogere niveaus van bijvoorbeeld klas of instelling). Vanwege deze en andere problemen achten Shadish, Cook en Campbell (2002) het REC alleen de moeite waard als a) een controlegroep echt niet tot de mogelijkheden behoort of b) wanneer REC aanvullend ten opzichte van andere designelementen¹⁰ met een hogere bewijskracht kan worden ingezet (zoals een onafhankelijke controlegroep).

Een ontwerp met alleen een benchmark en een nameting verschilt fundamenteel van het REC ontwerp van Shadish, Cook en Campbell (2002) in de zin dat de voorspelling niet gebaseerd is op een pretest in de interventiegroep zelf, maar op een afzonderlijke benchmarkgroep.¹¹ Een passende benaming zou dan ook het “*posttest-only-regression-extrapolation-contrast design*” (POREC) kunnen zijn. Vanwege het ontbreken van een expliciete pretest in de interventiegroep is de bewijskracht en geloofwaardigheid van het POREC ontwerp (nog) zwakker dan van het REC ontwerp. Het lijkt dan ook niet verstandig om de effectiviteit van de NT2-interventies louter en alleen vast te stellen op basis van vergelijking met de vooraf geconstrueerde benchmark. Naast de benchmark dienen ook andere designelementen aan het NT2-interventieonderzoek te worden toegevoegd. De belangrijkste elementen die voor toevoeging in aanmerking komen zijn een expliciete pretest en een controle- of vergelijkingsgroep. Het POREC ontwerp en de drie belangrijkste in aanmerking komende alternatieven voor het POREC ontwerp zijn hieronder visueel weergegeven.¹² Met O₁, X en O₂ zijn respectievelijk de pretest, de interventie en de nameting aangeduid. Met de

¹⁰ Een design element is hetgeen de onderzoeker in een (quasi-)experiment manipuleert of onder controle houdt teneinde andere verklaringen voor het gevonden effect dan de interventie uit te sluiten of minder aannemelijk te maken.

¹¹ Tenzij voor het benchmarkonderzoek en de NT2-interventie dezelfde cursisten gebruikt zouden worden, maar daar ga ik gemakshalve maar even niet van uit. Een methodologisch sterkere optie is het onderzoek naar het ontwikkelen van de benchmark en dat van de effectiviteit van de interventie bij verschillende groepen in dezelfde periode uit te voeren waarbij men een landelijk representatieve steekproef van cursisten bij de start van het onderzoek random over beide groepen verdeelt. In deze notitie wordt ervan uitgegaan dat de benchmark vooral gebruikt wordt ter voorspelling van het “gemiddelde te verwachten” taalniveau van een andere groep op een later tijdstip.

¹² Men zou nog kunnen denken aan een uitbreiding van het POREC ontwerp met alleen een controlegroep maar hiervan is afgezien omdat een controlegroep zonder “echte” pretest weinig bewijskracht toevoegt.

onderbroken lijn (---) is aangegeven dat de benchmarkgroep, interventiegroep en/of controlegroep niet via random toewijzing van docenten, klassen en/of cursisten tot stand zijn gekomen (en met de termen R en NR voor respectievelijk random en niet-random). Desgewenst kunnen meer interventiegroepen of voor- en nametingen aan het ontwerp worden toegevoegd.

A POREC (posttest-only-regression-extrapolation-contrast design)

NR Benchmark

NR		X	O ₂

B POREC met expliciete pretest

NR Benchmark

NR	O ₁	X	O ₂

C POREC met expliciete pretest en vergelijkgroep

NR Benchmark

NR	O ₁	X	O ₂

NR	O ₁		O ₂

D POREC met expliciete pretest en controlegroep

NR Benchmark

R	O ₁	X	O ₂
R	O ₁		O ₂

De vier hierboven weergegeven onderzoeksdesigns zijn geordend naar toenemende hardheid van het bewijs dat de interventie het positieve, neutrale of negatieve resultaat heeft bewerkstelligd.

POREC met expliciete pretest

Het toevoegen van de afhankelijke pretest O₁ in het ontwerp “POREC met expliciete pretest” heeft als belangrijkste voordeel dat inzicht kan worden verkregen in eventuele *selection bias* (de mogelijkheid dat de benchmark en interventiegroep al vanaf bij de start van de interventie qua samenstelling en/of taalniveau van elkaar verschilden). Een ander voordeel is dat de power om een werkelijk bestaand effect van de interventie daadwerkelijk aan te tonen flink toeneemt (mits voor- en nameting op dezelfde schaal uitgedrukt kunnen worden).

POREC met expliciete pretest en vergelijkingsgroep

Een tweede verbetering van de bewijskracht is de toevoeging van een non-random vergelijkingsgroep in het ontwerp “POREC met expliciete pretest en vergelijkingsgroep”. Met het gebruik van de term vergelijkingsgroep in plaats van controlegroep is aangegeven dat de toewijzing van docenten en/of cursisten aan de interventie- en controlegroep niet via random toewijzing tot stand is gekomen. De samenstelling kan tot stand komen via zelfselectie. Volgens Shadish, Cook en Campbell (2002) zou zelfselectie zoveel mogelijk moeten worden vermeden. Docenten die zichzelf voor de interventie hebben aangemeld of die voor de interventie geselecteerd zijn, kunnen immers beter gekwalificeerd of meer gemotiveerd zijn dan hun collega's in de controlegroep. Als de onderzoeker te weinig met deze (zelf)selectie rekening houdt, kan dat de schatting van het effect van de interventie sterk vertekenen (Webbink & Van der Steeg, 2008). Een controlegroep biedt dan uitkomst. Vaak probeert de onderzoeker random toewijzing zo goed mogelijk te benaderen door een “natuurlijke” vergelijkingsgroep te zoeken of samen te stellen die zoveel mogelijk op de interventiegroep lijkt (en in ons geval natuurlijk ook op de benchmarkgroep). De gebruikelijke middelen om een zo groot mogelijke gelijkens te bewerkstellingen zijn matching (en stratificatie) en statistische correctie voor bias achteraf. Het bewijs van een goed uitgevoerde quasi-experimentele studie kan harder zijn dan van een slecht uitgevoerd veldexperiment (Treasury Board of Canada, 1998). Nauwkeurige matching in een quasi-experimenteel ontwerp met een pretest en een vergelijkingsgroep lijkt in de meeste gevallen tot correcte algemene conclusies te leiden over de vraag of de interventie effectief of ineffectief is, maar de grootte van het effect van de interventie wordt vaak onder- of overschat (Whitehurst, 2003). Helaas is matching op alle relevante variabelen niet altijd mogelijk. Vrijwel altijd resteren alternatieve verklaringen voor het effect van de interventie (vaak dat de groepen al bij de start van de interventie van elkaar verschilden). Als niet alle mogelijk versturende factoren bekend zijn of de relevante factoren niet allemaal in de matching (*undermatching!*) of statistische correctie betrokken kunnen worden, kunnen deze middelen meer kwaad dan goed doen (Baron, 2005, 2007; Shadish, Cook & Campbell, 2002; Whitehurst, 2003; Glazerman, Levy & Myers, 2002). Zo concluderen laatstgenoemde onderzoekers op basis van een overzichtsstudie van niet-experimentele replicaties van sociale experimenten dat designs met een via matching tot stand gekomen vergelijkingsgroep de uitkomst van het experiment slechts in uitzonderingsgevallen goed benaderen (p. 46). Desalniettemin, mits goed uitgevoerd, heeft uitbreiding met een vergelijkingsgroep als grote voordeel dat er meer informatie beschikbaar komt over hoe groot de leerwinst van de interventiegroep zou zijn geweest in de hypothetische situatie dat deze groep niet aan de interventie zou hebben meegedaan.

POREC met expliciete pretest en controlegroep

Een derde verhoging van de bewijskracht is het random toewijzen van individuen, klassen en/of scholen aan de interventie- en controlegroep zoals dat in het “POREC met expliciete pretest en controlegroep” gebeurt. Een onderzoeksontwerp is gerandomiseerd als de eenheden volgens een aselekt trekkingsmechanisme aan de interventie- of controlegroep worden toegewezen. Het grote voordeel van random toewijzing is dat de groepen bij de start van de interventie in beginsel volledig identiek zijn op alle factoren, zowel bekende en onbekende factoren als gemeten en niet-gemeten factoren. Verschillen tussen beide groepen op de eindmeting kunnen daardoor relatief eenduidig aan de interventie worden toegeschreven, en niet aan allerlei andere verklarende factoren. Met de term gecontroleerd wordt bedoeld dat de onderzoeker de volledige controle heeft op het aselekt toewijzingsmechanisme. Een belangrijk argument voor een gerandomiseerd en gecontroleerd veldexperiment is de hardheid van de conclusies over het succes van de interventie en de geloofwaardigheid en bruikbaarheid van de onderzoeksresultaten voor onder meer beleidsmakers. Randomisatie

betekent overigens niet dat de interne validiteit van een veldexperiment volledig gegarandeerd is. Zo kan differentiële uitval tijdens de onderzoeksperiode het effect van de randomisatie weer teniet doen. Ook kan er contaminatie tussen interventie- en controlegroep ontstaan, bijvoorbeeld als cursisten van de interventiegroep overstappen naar de controlegroep (en vice versa) of als docenten in de interventiegroep hun kennis en ervaringen omtrent de beloftevolle nieuwe aanpak delen met hun collega's in de controlegroep. Randomisatie betekent evenmin dat de externe validiteit - de mate waarin de resultaten van het onderzoek stand houden in andere groepen, situaties, tijdstippen - in een veldexperiment beter gegarandeerd is dan in een niet-experimenteel onderzoek. Een veldexperiment is niet alleenzaligmakend, maar levert wel het hardste en meest geloofwaardige bewijs.

Conclusie

De voor- en nadelen van de onderscheiden designs tegen elkaar afwegend, lijkt een gerandomiseerd en gecontroleerd veldexperiment voor het NT2-interventieonderzoek in methodologisch opzicht de beste optie. Indien het veldexperiment goed wordt opgezet en uitgevoerd, lijkt een benchmark NT2 niet noodzakelijk om het effect van de interventie of het ontbreken ervan aan de interventie te kunnen toeschrijven. Als de vergelijking van de leerwinst in de benchmark- en de interventiegroep tot dezelfde uitkomst leidt als die tussen de interventie- en controlegroep verleent de benchmark extra geloofwaardigheid aan de onderzoeksresultaten. Wijzen de uitkomsten in een verschillende richting, zullen onderzoekers en opdrachtgever geneigd zijn de vergelijking met de benchmark minder te vertrouwen dan die met de onafhankelijke controlegroep. Naarmate het leerlastonderzoek echter een minder sterke opzet kent of minder zorgvuldig is uitgevoerd, wordt het gewicht van de benchmark NT2 belangrijker. Mocht een veldexperiment niet uitvoerbaar en/of niet aanvaardbaar blijken, kan worden uitgeweken naar een zo sterk mogelijk quasi-experimenteel onderzoek met een expliciete pretest en een vergelijkingsgroep waarbij onderzoekers de benchmark kunnen benutten voor het verkrijgen van extra evidentie.

3.2 Gebruik van een benchmark in ontwikkelingsgericht leerlastonderzoek

Het ministerie van VROM is voornemens onderzoek uit te zetten naar de benodigde duur van inburgeringsprogramma's voor verschillende groepen inburgeraars (Ministerie van VROM, 2007). Daarnaast wil het Ministerie praktijkgericht onderzoek stimuleren naar het effect van potentiële kenmerken van effectief NT2-onderwijs op de leerlast en het taalniveau (zie hoofdstuk 3). Daartoe worden naar verwachting vele ontwikkelingsgerichte pilots uitgevoerd waarin docenten en onderzoekers samenwerken aan het verbeteren van het NT2-onderwijs in overeenstemming met de gevonden succesfactoren. De gebruikssituatie van een benchmark in (quasi-)experimenteel leerlastonderzoek verschilt wezenlijk van die in ontwikkelingsgericht onderzoek. Het verschil zit hem niet alleen in het doel van het onderzoek, de aard en omvang van de steekproef en de kwaliteit van de gebruikte instrumenten en onderzoeksprocedures, maar ook in het aantal innovaties dat met de benchmark geëvalueerd wordt en in de logistieke, financiële en ethische consequenties van eventuele fout-negatieve of fout-positieve conclusies. Webbink en Van der Steeg (2008) zijn van mening dat pilots te weinig benut worden voor het vergaren van evidence-based kennis over wat werkt en wat niet werkt in het onderwijs. Hun advies is om aan toekomstige pilots een controlegroep toe te voegen. Vóór de start van een pilot zouden scholen geworven kunnen worden voor deelname. Vervolgens zouden de onderzoekers geïnteresseerde scholen via loting aan de pilot- of controlegroep kunnen toewijzen. Hiermee wordt een eerste stap gezet van een pilot naar een gecontroleerd

experiment. Behalve voordelen heeft het voorstel van Webbink en Van der Steeg (2008) echter ook enkele innovatiestrategische nadelen:

- Random toewijzing van in innovatie geïnteresseerde personen of scholen aan een pilot- en een controlegroep, zoals Webbink en Van der Steeg (2008) voorstellen, betekent dat de onderzoeker vernieuwingsgezinde docenten en onderwijsaanbieders zal moeten teleurstellen. Velen vinden dit om begrijpelijke redenen niet acceptabel.
- In een situatie waarin er vele pilots naar potentiële succesfactoren worden uitgezet is het voorzien van elke pilot van een controlegroep (onnodig) arbeidsintensief en kostbaar.
- In ontwikkelingsgericht onderzoek gaat het doorgaans om kleine steekproeven. Mocht de pilot onverhoopt ineffectief blijken, is het aantal gedupeerde cursisten klein. De “stakes” zijn veel minder hoog dan bij bijvoorbeeld een grootschalig veldexperiment waar er bijvoorbeeld een beslissing genomen moet worden over het al dan niet landelijk invoeren van een interventie met grote logistieke en financiële consequenties. In praktijkgerichte pilots naar het verbeteren van innovaties is het met andere woorden wat minder “erg” wanneer de vergelijking met de benchmark in een aantal gevallen tot een fout-positieve of fout-negatieve conclusie zou leiden.

In de situatie waarin vele kleinschalige ontwikkelingsgerichte pilots tegelijkertijd uitgevoerd worden kan een benchmark NT2 een efficiënt alternatief zijn voor een onafhankelijke controlegroep. De gemiddelde leerwinst in de benchmarkgroep vormt dan het criterium waartegen de leerwinst in de pilots kan worden afgezet. In statistisch opzicht dient te worden nagegaan in hoeverre de op basis van de benchmark voorspelde leerwinst significant afwijkt van de schatting van de daadwerkelijk leerwinst in de pilotgroep. Volgens een eenvoudige vuistregel is het verschil tussen beide parameterschattingen statistisch significant als de beide betrouwbaarheidsintervallen elkaar niet overlappen. Een bij ontwikkelingsgericht leerlastonderzoek passend onderzoeksontwerp met een benchmark en bijvoorbeeld vijf simultaan plaatsvindende pilots is hieronder afgebeeld (waarbij we ervan uitgaan dat er eerst een benchmark ontwikkeld wordt en dat deze vervolgens wordt gebruikt ter voorspelling van de leerwinst in de pilots onder de aanname dat de cursisten niet aan de pilot hadden meegedaan, maar gemiddeld onderwijs hadden ontvangen van een gemiddelde docent bij een gemiddelde onderwijsaanbieder).

Ontwerp E

NR Benchmark

NR	Pilot 1	O ₁	X	O ₂
NR	Pilot 2	O ₁	X	O ₂
NR	Pilot 3	O ₁	X	O ₂
NR	Pilot 4	O ₁	X	O ₂
NR	Pilot 5	O ₁	X	O ₂

Een benchmark alleen levert in deze situatie geen antwoord op de vraag of een eventueel verschil tussen de leerwinst in de benchmarkgroep en de pilot aan de onderzochte succesfactoren dan wel aan andere invloeden kan worden toegeschreven. Omdat cursisten, docenten of aanbieders niet at random aan de pilot- of controlegroep zijn toegewezen, moet de

onderzoeker bijzonder alert zijn op alle mogelijke alternatieve verklaring voor het effect van de pilot (zoals *selection bias*, *regression artifacts*, *instrumentation*, *history*, *mortality* en *maturation*). Al dient te worden gezegd dat de voormeting van de taalvaardigheid van de vier groepen cursisten eventuele *selection bias* waarschijnlijk voor een deel zal afvangen (als de voormeting op dezelfde schaal is als de nameting).

In ontwerp E zijn we ervan uitgaan dat er eerst een benchmark ontwikkeld wordt en dat deze vervolgens wordt gebruikt ter voorspelling van de leerwinst in de pilots (onder de aanname dat de cursisten niet aan de pilot hadden meegedaan, maar gemiddeld onderwijs hadden ontvangen van een gemiddelde docent bij een gemiddelde onderwijsaanbieder). De bewijskracht van het onderzoeksontwerp kan verder worden verbeterd door de benchmark regelmatig te verversen, bijvoorbeeld door elk jaar een nieuw cohort cursisten in de tijd te gaan volgen. Het verbeterde ontwerp kan dan als volgt worden weergegeven (waarbij we er gemakshalve van uitgaan dat er in alle groepen slechts twee meetmomenten zijn en de voor- en nameting in de benchmark- en pilotgroepen op precies hetzelfde moment plaatsvinden).

Ontwerp F

NR	Benchmark	O ₁		O ₂
NR	Pilot 1	O ₁	X	O ₂
NR	Pilot 2	O ₁	X	O ₂
NR	Pilot 3	O ₁	X	O ₂
NR	Pilot 4	O ₁	X	O ₂
NR	Pilot 5	O ₁	X	O ₂

Als de benchmark voorafgaand aan de pilot ontwikkeld wordt, kan *history* als alternatieve verklaring voor het effect van de succesfactoren niet worden uitgesloten (zie paragraaf 3.1 van deze bijlage). Het jaarlijks actualiseren van de benchmark, zoals weergegeven in ontwerp F, heeft onder meer als voordeel dat het mogelijk versturende effect van *history* wordt tegengegaan.

Ter verdere verhoging van de bewijskracht van het onderzoeksontwerp zijn volgende maatregelen wenselijk:

- Een eerste verbetering is het constant houden van de onderzoeksinstrumenten en – procedures. Als er in de benchmarkgroep andere instrumenten worden gebruikt of dezelfde instrumenten op een andere wijze worden afgenomen dan in de pilot, kan *instrumentation* als mogelijke alternatieve verklaring voor het effect van de potentiële succesfactoren niet worden uitgesloten. Het is niet alleen wenselijk dat de onderzoeksinstrumenten voor de meting van leerlast en taalniveau dezelfde zijn. Ook zouden de procedures voor het afnemen, scoren en beoordelen van de toetsen idealiter constant gehouden moeten worden. Als de taaltoetsen in de benchmarkgroep bijvoorbeeld onder gecontroleerde omstandigheden door externe en getrainde proefleiders worden afgenomen en beoordeeld, zou dat in de pilotgroep op dezelfde wijze moeten gebeuren. Het zelfde geldt voor het geval de taalvaardigheid van de cursisten in de pilotgroep vaker of minder vaak gemeten zou zijn dan in de benchmarkgroep (zogenoeten *testing artifacts*).

- Een tweede verbetering is het uitvoeren van een post hoc robuustheidstudie. Daarin kan worden nagegaan in hoeverre de regressiegewichten zoals bepaald in de benchmarkgroep vergelijkbaar zijn met de regressiegewichten zoals geschat in de pilotgroep. Houden de regressiegewichten uit het benchmarkonderzoek stand in de pilotgroep, zou men het verschil in leerwinst met een aantal slagen om de arm aan de succesfactoren kunnen toeschrijven. Zijn de regressiegewichten onvergelijkbaar, dan kan geen uitspraak over het effect van de pilot worden gedaan. Een voorwaarde voor een zinvolle toepassing is dat er ten minste enige zekerheid moet zijn dat alle belangrijke voorspellers van leerwinst in de robuustheidanalyse verdisconteerd zijn (en de situatie van random toewijzing aan een benchmark- en pilotgroep zo dicht mogelijk benaderd kan worden). Daarbij dienen er bij voorkeur behalve de leerlast een groot aantal andere voorspellers van de leerwinst beschikbaar te zijn, niet alleen in de benchmarkgroep, maar ook in de pilotgroep (denk bijvoorbeeld aan kenmerken van de cursisten, de NT2-docenten, de instructie, de instelling en de beperkende randvoorwaarden). Voor stabiele schattingen van regressiegewichten zijn overigens grote steekproeven en betrouwbare meetinstrumenten nodig, zowel in de benchmarkgroep als de pilotgroep.

Conclusie

Al met al lijkt een ontwerp met een benchmark en een pre-posttest ontwerp in het kader van ontwikkelingsgerichte pilots een efficiënt alternatief voor onafhankelijke controlegroepen of vergelijkingsgroepen. Wel moet de onderzoeker bijzonder alert zijn op alternatieve verklaringen voor het effect van de pilot (zoals *selection bias*, *mortality*, *regression artifacts*, *instrumentation* en *testing artifacts*) en bij het rapporteren van conclusies over het effect van de pilot de grootst mogelijke terughoudendheid betrachten.

Literatuur

Abbenhuis, T., & Linden, T. van der, & Duijm, K. (1996). *Leerlast-indicator 1996. Deel 1: Het maken van een leerlast-indicator: verslag en resultaten*. Amersfoort: Bureau ICE.

Alamprese, J., Tao, F., & Price, C. (2003). *Study of reading instruction for low-level learners in adult basic education programs: Volume 1: Study findings* (Draft Report). Bethesda, MD: Abt Associates, Inc.

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

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

Akker, J. van den, & Voogt, J. (1994). The use of innovation and practice profiles in the evaluation of curriculum implementation. *Studies in Educational Evaluation*, 20, 503-512.

Arvodi (2008). *Algemene Rijksvoorwaarden voor het verstrekken van Opdrachten tot het verrichten van Diensten 2008 (ARVODI-2008)*. Vastgesteld bij besluit van de Minister-President, Minister van Algemene Zaken, van 26 februari 2008, nr. 3313667 (Stcrt. 2008, 52, p. 7).

Balmuth, M. (1987). *Essential characteristics of effective adult literacy programs: A review and analysis of the research*. Albany, NY: University of the State of New York.

Baron, J. (2005). *Key items to get right when conducting a randomised controlled trial in education*.

Online beschikbaar via: http://ies.ed.gov/ncee/wwc/pdf/guide_RCT.pdf

Baron, J. (2007). *Checklist for reviewing a randomized controlled trial of a social program or project, to assess whether it produced valid evidence*. Online beschikbaar via: <http://www.evidencebasedpolicy.org/docs/Checklist-For-Reviewing-RCTs-Final.pdf>

Beder, H. (1999). *The outcomes and impacts of adult literacy education in the United States* (NCSALL Report No. 6). Cambridge, MA: National Center for the Study of Adult Learning and Literacy.

Bekkers, L., Bossers, B., Jetten, M., & Soeting, J. (1999). *De NT2 Profieltoets in de praktijk*. Amersfoort/Arnhem: ICE/Cito.

Bergh, H. van den, & Kuhlemeier, H. (1997). Multi-level modellen voor de analyse van leerwinst vergeleken. *Tijdschrift voor Onderwijsresearch*, 22, 2, 54-75.

Bingman, M. B., Ebert, O., & Smith, M. (1999). *Changes in learners' lives one year after enrollment in literacy programs: An analysis from the Longitudinal Study of Adult Literacy Participants in Tennessee* (NCSALL Report No. 11). Cambridge, MA: National Center for the Study of Adult Learning and Literacy.

Brink, M., B. Does, A. Odé & E. Tromp (2002). *Verscheidenheid in integratie. Evaluatie van de effectiviteit van de Wet inburgering nieuwkomers*. Amsterdam: Regioplan Beleidsonderzoek.

Brooks, G., Giles, K., Harman, J., Kendall, S., Rees, F., & Whittaker, S. (2001). *Assembling the fragments: A review of research on adult basic skills* (Research Report No. 220). London: Department for Education and Employment.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Bureau ICE (2008). *TOA Inhoudelijke handleiding Toolkit Onderwijs en arbeidsmarkt Versie 5.0*. Culemborg: ICE.

California Department of Education (2004). *The relationship of adult ESL reading performance to instructional time*. Sacramento, CA: CDE.

Campbell, D.T., & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Coalition for evidence-based policy (2005). *Key items to get right when conducting a randomised controlled trial in education*. Online beschikbaar via:
http://ies.ed.gov/ncee/wwc/pdf/guide_RCT.pdf

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cohen, J., Condelli, L., & Garet, M. (1996). *Reanalysis of the National Evaluation of Adult Education Programs: Methods and results*. Washington, DC: U.S. Department of Education, Planning and Evaluation Services, Office of the Under Secretary.

Comings, J.P., & Soricone, L. (2007). *Adult literacy research: Opportunities and challenges*. Washington, DC: National Institute for Literacy.

Comings, J., Soricone, L., & Santos, M. (2006). *An evidence-based adult education program model appropriate for research* (NCSALL Occasional Paper). Cambridge, MA: National Center for the Study of Adult Learning and Literacy.

Commissie Franssen (2004). *Normering inburgeringsexamen: Advies over het niveau van het nieuwe inburgeringsexamen*. Den Haag: Ministerie van VROM.

Condelli, L., & Wrigley, H. (2004b). *Identifying promising literacy interventions for adult ESL literacy students: A review of the literature*. Washington, DC: U.S. Department of Education, Institute for Educational Sciences.

Condelli, L., Wrigley, H., Yoon, K., Cronen, S., & Seburn, M. (2007). *What works study for adult ESL literacy students: Final report*. Washington, DC: U.S. Department of Education.

Connecticut State Department of Education (2007). *Learning gains and attendance*. Online beschikbaar via:

http://www.sde.ct.gov/sde/lib/sde/pdf/deps/adult/accountability/esl_learning_gain_final.pdf

Cook, T. D., Appleton, H., Conner, R., Shaffer, A., Tamkin, G., & Weber, S. J. (1975). *"Sesame Street" revisited*. New York: Russel Sage Foundation

Corte, E. de, Geerligs, C., Lagerweij, N., Peters, J., & Vandenberghe, R. (1976). *Beknopte didaxologie*. Groningen: Wolters-Noordhoff.

Corte, E. de, & Verschaffel, L. (2001). Construerend onderwijskundig onderzoek: hefboom voor theorievorming en innovatie". *Pedagogisch Tijdschrift*, 26, 2, 167-188.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University press.

Cronen, S., Silver-Pucuilla, H., & Condelli, L. (2006). *Conducting large-scale research in adult ELS : Challenges and approaches for the explicit literacy impact study*. Washington, DC: American Institutes for Research.

Dalderop, K., Kuhlemeier, H., Kuijper, H., Speijers, A-M., Stevens, C., & Treep, J. (2008). *Theoretisch kader voor leerlastonderzoek: Op zoek naar het succes van inburgering*. Arnhem: Cito.

Design-Based Research Collective (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, 32, 1, 5-8.

Deutekom, J. (2007). *Passief zijn geen optie*. Inburgeren bij het Friesland College. *Les*, 148, 6-9.

Deutekom, J. (2008). *FC Sprint²: grenzeloos leren*. Meppel: Boom.

Duijm, K. (red.) (1996). *Leerlast-indicator 1996. Deel 2: Een eerste aanzet. Hoeveel lesuren zijn nodig voor het bereiken van NT2-niveaus?*. Amersfoort: Bureau ICE.

Fraenkel, J.R., & Wallen, N.E. (2003). *How to design and evaluate research in education* (5th ed.). Boston: McGraw-Hill.

Gelder, L. van, Oudkerk Pool, T., Peters, J., & Sixma, J. (1973). *Didactische analyse*. Groningen: Wolters-Noordhoff.

Glazerman, S., Levy, D.N., & Myers, D. (2002). *Nonexperimental replications of social experiments: A systematic review*. Interim Report/Discussion Paper, Mathematica. MPR Reference No.: 8813-300.

Goldstein, H. (1995). *Multilevel models in educational and social research*. London: Griffin.

Huizinga, M., Kalsbeek, A. van (1996). *Kijk op de klas*. Amsterdam: VU Uitgeverij.

Kok-Damave, M. de (1980). *De Curvo-strategie: handboek voor leergangontwikkeling*. 's-Gravenhage: Stichting voor Onderzoek van het Onderwijs.

Kruidenier, J. (2002). *Research-based principles for adult basic education: Reading instruction*. Washington, DC: The Partnership for Reading.

Kuhlemeier, H. (2008). *De bijdrage van een benchmark NT2 aan onderwijsvernieuwing*. Presentatie gehouden tijdens de Cito-werkconferentie "Op zoek naar het succes van inburgering" op 14 november 2008 te Utrecht.

Kuhlemeier, H., Kleintjes, F., & Bergh, H. van den (2001). Effect van toetsvorm en vraagtype op de moeilijkheid van de afsluitingstoetsen basisvorming: Een toepassing van multiniveau analyse met random kruisclassificatie. *Pedagogische Studiën*, 78, 3, 197-211.

Leithwood, K.A., & Montgomery, D. J. (1987). *Improving classroom practice: Using innovation profiles*. Toronto: The Ontario Institute for Studies in Education.

Linden, T. van der (1996). *Leerlast NT2 nader bekeken: verslag van een expertmeting naar aanleiding van het onderzoek ten behoeve van de leerlast-indicator 1996 op 12 september 1996*. Amersfoort: Bureau ICE.

Ministerie van VROM (2007). *Deltaplan inburgering: Vaste voet in Nederland*. Den Haag: Ministerie van VROM.

OCW-SenterNovem (2008). *Leidraad deename prijsvraag actieprogramma Onderwijs Bewijs*. Den Haag: Ministerie van OCW.

Onderwijsraad (2006). *Naar meer evidence-based onderwijs*. Den Haag: Drukkerij Artoos.

Porter, K. E., Cuban, S., & Comings, J. (2005). "One day I will make it." *A study of adult student persistence in library literacy programs*. New York: MDRC.

Raudenbush, S.W. (2003). *Designing field trials of educational educations. Paper prepared for the national invitational conference 'Conceptualizing scale-up: Multidisciplinary perspectives'*. Online beschikbaar via: <http://www-personal.umich.edu/~rauden/DRDC.pdf>

Raudenbush, S.W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5, 2, 199-213.

Raudenbush, S. W., & Liu, X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, 6, 4, 387-401.

Reder, S., & Bynner, J. (Eds, 2008). *Tracking adult literacy and numeracy skills: Findings from longitudinal research (Series Routledge Research in Education)*. New York, NY: Routledge.

Schochet, P.Z. (2005). *Statistical power for random assignment evaluations of education programs*. Washington, DC: Institute of Education Sciences.

Significant (2004). *Inburgering nieuwkomers. Kwalitatieve rapportage 2003*. Barneveld: Ministerie van justitie.

Significant (2007). *Monitors inburgering 2006*. Barneveld: Ministerie van VROM.

Snoeken, H. (2008). *Inleiding op de Cito-werkconferentie “Op zoek naar het succes van inburgering” (tentatieve titel)*. Presentatie gehouden tijdens de Cito-werkconferentie “Op zoek naar het succes van inburgering” op 14 november 2008 te Utrecht.

Tashakkori, A., & Teddlie, C. (1998). Mixed methodology: Combining qualitative and quantitative approaches. *Applied Social Research Methods Series*, 46. Thousand Oaks, CA: Sage Publications.

Tuijl, P.van., Braun, C., & Weteling, R.A. (2004). *Inburgering nieuwkomers 2002. Inhoudelijk verslag van resultaten en uitvoering*. Capelle aan de IJssel: Piers groep.

Torgerson, C., Brooks, G., Porthouse, J., Burton, M., Robinson, A., Wright, K. & Watt, I. (2004). *Adult literacy and numeracy interventions and outcomes: A review of controlled trials*. London: Institute of Education.

Towne, L., & Hilton, M. (2004). *Implementing randomised field trials in education: Report of a workshop* (Washington, DC, September 24, 2003). Washington, DC: The National Academies Press.

Treasury Board of Canada, Secretariat (1998). *Program evaluation methods: Measurement and attribution of program results (Third edition)*. Ottawa: Public Affairs Branch.

Tyler, R.W. (1949). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.

U.S. Office of Management and Budget (2007). *What constitutes strong evidence of a program's effectiveness?* Online beschikbaar via:
www.whitehouse.gov/omb/part/2004_program_eval.pdf

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development*, 53, 4, 5-23.

Watt, D.L.E. & Lake, D.M. (2004). *Benchmarking adult rates of second language acquisition and integration. How long and how fast? Final report for phase 1. 2001/2001*. Calgary: University of Calgary.

Webbink, D., & Van der Steeg, M. (2008). *Essay voor de Parlementaire Onderzoekscommissie Onderwijshervormingen (CPB Memorandum 201)*. Den Haag: Cultureel Planbureau.

White, J., & Hoddinott, S. (1991). *Organizing adult literacy and basic education in Canada—A policy and practice discussion document*. Ottawa: Movement for Canadian Literacy.

Whitehurst, W.J. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: U.S. Department of Education Institute of Education Sciences.

Young, M., Fleischman, H., Fitzgerald, N., & Morgan, S. (1995). *National evaluation of adult education programs: Executive summary*. Arlington, VA: Development Associates.