Hoe goed denk je dat je bent?

Een meta-analyse over zelfbeoordelingen van vreemde- en tweedetaalleerders.

Immy Niemeijer
S4232895
Onder begeleiding van: Prof. Dr. J.J.M. Schoonen

Inhoudsopgave

Abstract

Introduction/ Self-assessment (SA) in foreign and second language learning aides learning development and offers multiple advantages over regular performative tests. The main research question in the current study is as follows: To what extent is SA in foreign and second language accquisition a valid instrument to measure actual language skills? Mulitple studies have been conducted in this field. However, the vastness of the literature warrants a systematical approach.
Method/ In this meta-analysis 25 studies (n = 2138) have been included following an extensive database search in ERIC, LLBA and Google Scholar. SA quality was assesed independently to account for interobserver bias. SA scores were compared to linguistic performative test results to evaluate construct validity. The primary outcome measure was effect size (*r*).
Results/ The main effect size was 0.52. Analysis of standard deviations and effect sizes yielded no visual indication of publication bias. Homogeneity analysis demonstrated heterogeneity in the included studies. Implementing 'can do' statements significantly increased SA validity.
Conclusion/ The findings indicate limited usefulness of SA as a summative test. However, SA might be implemented to augment learning responsibility and hence learning accquistion.

Inleiding

Wij zijn als mensen niet goed in het inschatten van ons eigen gedrag. Gemiddeld beoordelen we onze vaardigheden met 'bovengemiddeld' (Dunning, Heath, & Suls, 2004). Onze zelfbeoordelingen hangen vaak maar amper samen met objectieve vaardigheidsmetingen (hierna simpelweg 'vaardigheidsmetingen'). We overschatten onze eigen gezondheid, kunnen ons eigen werk maar moeilijk beoordelen en weten niet of we wel de juiste opleiding volgen (Carter & Dunning, 2008; Dunning, Heath & Suls, 2004). Als we zelfbeoordelingen (of self-assessment, SA) en beoordelingen door anderen met elkaar vergelijken, is de gemiddelde correlatie 0.29, waarbij de hoogste correlaties gehaald worden bij sport en de laagste in het sociale domein (Mabe & West, 1982). Dit betekent dat de twee soorten beoordelingen voor ongeveer acht procent samenvallen: een opvallend kleine overlap.

Toch vinden we het wel belangrijk om onze eigen vaardigheden goed in te kunnen schatten. We baseren namelijk veel van onze keuzes op deze inschattingen. Daarbij kun je bijvoorbeeld denken aan een inschrijving voor de marathon, maar ook aan de studiekeuze van aanstaande studenten. Om hierin goede keuzes te kunnen maken, moeten we een goede inschatting maken van onze capaciteiten. Foute zelfbeoordelingen kunnen niet alleen tot onpraktische situaties leiden (wanneer we een te moeilijk recept willen koken), of zelfs gevaarlijke (als we bijvoorbeeld kijken naar onze verkeersvaardigheden), maar ze kunnen ook op andere punten in de weg zitten. Op de werkvloer nemen werknemers bijvoorbeeld met tegenzin advies aan wanneer ze zichzelf overschatten en starten CEO's te snel met projecten die ze niet aankunnen (Dunning, Heath, & Suls, 2004). Ook het onderschatten van de eigen vaardigheden is problematisch omdat je je talenten dan niet ten volle benut. Het is daarom belangrijk om een correcte zelfbeoordeling te kunnen doen.

Ook binnen de vreemde- en tweedetaalverwerving is SA een veelvuldig gebruikt instrument. Zelfbeoordelingen worden voornamelijk ingezet in combinatie met een alternatieve toets. SA wordt onder andere gebruikt in de diagnostische online toets DIALANG, in het European Language Portfolio (ELP) en het Bergen 'Can Do' project. Zelfbeoordelingen worden in deze taaltoetsen ingezet omdat ze een taalleerder helpen doelen te stellen en stappen te zetten in het leerproces. In DIALANG wordt SA tevens gebruikt om het niveau van de daaropvolgende vaardigheidsmeting te bepalen. In geen van deze toetsen wordt het resultaat van de zelfbeoordeling summatief gebruikt om het taalvaardigheidsniveau (mede) vast te stellen.

Zelfbeoordelingen worden daarnaast steeds meer in lesmethodes opgenomen (Butler, 2018). In deze methodes wordt SA ingezet als metacognitief middel om de taalleerder te helpen over haar eigen vaardigheden na te denken. SA zorgt op die manier namelijk voor onafhankelijke taalleerders (Bullock, 2011). Door SA zouden taalleerders beter doelen kunnen stellen in hun leerproces en daar specifieke acties aan kunnen koppelen (Chen, 2008; Ross, 2006; Paris & Paris, 2001; Kohonen, 2000). Het doel van SA in lesmethodes is dus formatief: de resultaten worden door de docent en de cursisten gebruikt om het leerproces bij te sturen.

Tot dusver worden zelfbeoordelingen dus overwegend formatief ingezet. Toch zou het ook voordelen kunnen bieden wanneer een SA summatief kan worden gebruikt, in plaats van een andere soort vaardigheidsmeting. SA biedt onder meer een aantal voordelen die praktisch van aard zijn. De vorm van een zelfbeoordeling is namelijk die van een simpele questionnaire (in plaats van een luistertoets met vooraf opgestelde luisterfragmenten) die makkelijk online kan worden afgenomen. Dat scheelt bijvoorbeeld bij plaatsingstoetsen veel tijd; een cursist kan zichzelf aan de hand van verschillende vragen in een groep plaatsen, zonder dat een docent of andere beoordelaar daaraan te pas komt. Het loont dus de moeite om te kijken of een zelfbeoordeling accuraat genoeg is om als vaardigheidsmeting te worden ingezet.

De onderzoeksvraag in het huidige onderzoek is als volgt: In hoeverre zijn zelfbeoordelingen binnen vreemde- en tweedetaalverwerving een valide methode om taalvaardigheden te meten? Daarbij gaat het niet om een formatief gebruik van SA, maar om summatief gebruik. Het huidige onderzoek richt zich dus primair op de accuraatheid van zelfbeoordelingen binnen de vreemde- en tweedetaalverwerving. Deze accuraatheid wordt hier opgevat als validiteit: de samenhang tussen zelfbeoordelingen en gelijktijdig beschikbare gegevens over de taalvaardigheid, verkregen uit valide vaardigheidsmetingen. Deze samenhang vertelt ons in hoeverre een SA te gebruiken is in plaats van een vaardigheidsmeting.

Bij deze onderzoeksvraag zijn twee subvragen geformuleerd:

1. Wat zijn de kenmerken van een valide SA?
2. Spelen eventuele confounders een rol bij de validiteit van een SA?

Allereerst zal SA binnen vreemde- en tweedetaalverwerving nader worden toegelicht in een theoretisch kader. Daarin wordt dit veld besproken aan de hand van onderzoek dat al in deze richting is gedaan. De focus zal in dit theoretisch kader voornamelijk liggen op variabelen die eventueel mee kunnen spelen bij de accuraatheid van een zelfbeoordeling.

Daarna zal een meta-analyse worden gedaan om de onderzoeksvraag te beantwoorden, waarin de vele experimenten naar dit onderwerp worden gebundeld. Het huidige onderzoek is de derde meta-analyse die is uitgevoerd naar de validiteit van zelfbeoordelingen binnen vreemde- en tweedetaalverwerving. Het eerste onderzoek is uitgevoerd door Blanche en Merino (1989), en negen jaar later is een tweede meta-analyse gedaan door Steve Ross (1998) dat nog steeds veel wordt geciteerd. In deze twee meta-analyses waren echter niet veel onderzoeken opgenomen (7 opgenomen onderzoeken in Blanche & Merino, 1989; 11 opgenomen onderzoeken in Ross, 1998), waardoor het beeld over de validiteit van SA nog niet duidelijk is. Sindsdien zijn er veel nieuwe experimenten gedaan naar dit onderwerp. Het was daarom interessant om een geüpdatete versie te maken, waarin onder andere nieuwere vormen van toetsing mee konden worden genomen. Bovendien worden in het huidige onderzoek meerdere factoren onderzocht die een rol kunnen spelen bij de accuraatheid van zelfbeoordelingen.

Theoretisch kader

In dit theoretisch kader worden achtereenvolgens de voordelen van SA besproken, de zelfbeoordelingen die tot nu toe in gestandaardiseerde toetsen zijn gebruikt, de kwaliteiten waar een goede SA aan zou moeten voldoen, en tot slot de validiteit van SA en de variabelen die daar invloed op hebben.

De definitie van SA die in dit theoretisch kader wordt aangehouden is die van Klenowski (1995): "the evaluation or judgment of 'the worth' of one's performance and the identification of one's strengths and weaknesses". Deze definitie laat zowel ruimte voor een meting waarmee de accuraatheid van een zelfbeoordeling kan worden getest, als voor een formatieve lezing die gefocust is op het effect van SA op het leerproces. In het huidige onderzoek is vooral de accuraatheid van een zelfbeoordeling van belang, omdat deze als argument dient voor het wel of niet inzetten van SA als vaardigheidsmeting.

Voordelen van SA

Zelfbeoordelingen bieden veel positieve gevolgen voor leerders binnen hun leerproces, zoals een verhoogde motivatie, meer eigen verantwoordelijkheid en het vermogen om realistische doelen te stellen. Daarnaast biedt een SA ook voordelen ten opzichte van alternatieve vaardigheidsmetingen. In deze paragraaf wordt eerst een aantal voordelen van SA op zichzelf behandeld en daarna de voordelen van SA ten opzichte van vaardigheidsmetingen.

Een groot voordeel van SA is dat het een beroep doet op de eigen verantwoordelijkheid van de taalleerder. De verantwoordelijkheid voor de evaluatie en de bewijslast daarvan zijn niet meer extern (zoals door een docent of door daartoe ontwikkelde software), maar komen bij de leerder zelf te liggen. Hoewel dit bij sommige taalleerders voor problemen zorgt omdat ze vinden dat deze verantwoordelijkheid bij iemand met autoriteit zou moeten liggen, is dit aantal maar klein en zijn de problemen meestal op te lossen door het proces van SA toe te lichten (LeBlanc & Painchaud, 1985). Door zich de metalinguïstische houding aan te meten die nodig is voor zelfbeoordeling, wordt een tweede- of vreemdetaalleerder zich bewust van meerdere zaken die essentieel zijn voor een effectief leerproces (Kohonen, 2000): 'task awareness', 'personality awareness', en 'process and context awareness'. Deze begrippen betekenen respectievelijk dat een taalleerder kennis heeft over subvaardigheden die binnen een taal van belang zijn (zoals het kunnen schrijven van brieven of het lezen van krantenartikelen), weet in hoeverre zij die zelf machtig is, en welke stappen er moeten worden genomen om verder te komen in het leerproces. Met andere woorden: met SA worden taalleerders gedwongen om met een kritische blik naar hun taalleerproces te kijken, waardoor ze realistische doelen kunnen stellen en weten welke stappen ze moeten ondernemen om daar te komen, uitgaande van een accurate zelfbeoordeling. Met een goede zelfbeoordeling krijgt een taalleerder dus kennis om in het taalleerproces toe te passen.

Een tweede positief resultaat van SA is een verhoogde motivatie bij taalleerders. Doordat een zelfbeoordeling doet reflecteren en evalueren, krijgen taalleerders een verhoogd gevoel van controle in hun eigen taalproces. Dit zou ertoe leiden dat ze een positieve houding krijgen tegenover leren (Paris & Paris, 2001). Waar reguliere vaardigheidsmetingen negatieve gevolgen kunnen hebben voor sommige leerders - zoals de uitsluitend extrinsieke motivatie voor het halen van goede cijfers en de tegenslag wanneer dat niet lukt - richt een SA zich op een intrinsieke manier op de prestaties. De controle op het eigen proces en deze onafhankelijke manier van evalueren leiden samen tot een verhoogde motivatie:

> Self-evaluation of schoolwork is linked to affective characteristics such as attitudes, interests, feelings of success at school, and enjoyment of reading and writing at home. These are positive motivational characteristics of achievement-striving students and suggest that students who are more effective at self-appraisal have more positive attitudes about school and enjoy reading and writing.
> (Paris & Paris, 2001, p. 96).

Deze twee gevolgen van SA - eigen verantwoordelijkheid en een verhoogde motivatie- zorgen voor het onafhankelijk leren van een taal (self-regulated learning, SRL*)*. SRL is een thema dat veelvuldig is behandeld door wetenschappers en docenten omdat het een belangrijk doel is voor studenten in alle disciplines. Een onafhankelijke taalleerder weet zijn eigen vaardigheden juist te interpreteren en verbindt daar zelf gepaste handelingen aan door doelen te stellen en planningen te maken. De assumptie die hieraan ten grondslag ligt is dat een docent zijn cursisten niet alles kan leren en dat een groot gedeelte van het taalleerproces buiten het klaslokaal gebeurt (Bullock, 2011). Door van studenten zelfstandige taalleerders te maken kunnen ze daar bewust en kritisch mee omgaan. Dit alles zorgt voor een positief effect op het leerproces (Paris & Paris, 2001; Zimmerman, 2000).

De vraag is natuurlijk of het inzetten van een SA ook daadwerkelijk invloed heeft op de resultaten van taalleerders. Er zijn dan ook verschillende experimenten gedaan naar het effect van SA. In deze experimenten wordt het effect van SA getest door twee groepen van taalleerders met elkaar te vergelijken: een groep volgt een traject waaraan zelfbeoordelingen te pas komen, de andere dient ter controle. De gemiddelde vooruitgang van de twee groepen wordt dan met elkaar vergeleken. In het onderzoek door Butler en Lee (2010) deed de experimentele groep vier maanden lang, elke twee weken een SA. Aan het einde bleken de zelfbeoordelingen een positief effect te hebben op de leeropbrengst, te zien aan de resultaten van de Cambridge Young Learners' English Test, wanneer ze werden vergeleken met de controlegroep. Daarnaast had de experimentele groep een groter zelfvertrouwen over de taalvaardigheden en waren ze nauwkeuriger geworden in hun inschattingen. Het literatuuronderzoek van Jamrus en Razali (2019), waarin tien van dit soort experimenten zijn opgenomen, laat een vergelijkbaar resultaat zien. Het inzetten van zelfbeoordelingen in taallessen zou dus een positieve invloed hebben op de leeropbrengst.

Zelfbeoordelingen kunnen, zoals we hierboven hebben gelezen, ingezet worden als hulpmiddel bij het leren van een tweede of vreemde taal. De vraag die in het huidige onderzoek centraal staat, is of ze ook als toets ingezet zouden kunnen worden. Omdat

taalleerders in een SA zelf aan kunnen geven of en in welke mate ze een bepaalde vaardigheid beheersen en dat niet hoeven te bewijzen, bieden zelfbeoordelingen namelijk veel voordelen ten opzichte van een vaardigheidsmeting. SA heeft bijvoorbeeld een aantal voordelen die praktisch van aard zijn. Omdat er geen rekening hoeft te worden gehouden met afkijken, kan een SA makkelijk thuis worden gemaakt. Er hoeven geen afspraken te worden gemaakt of lokalen te worden vrijgehouden. Ook komen in een SA doorgaans geen luisterfragmenten of spreekopdrachten voor, waardoor een simpele online questionnaire volstaat. Technisch materiaal zoals microfoons, beeldschermen en speakers is dan niet nodig. Bovendien is een SA doorgaans korter dan een vaardigheidsmeting: volgens LeBlanc en Painchaud (1985) kost een zelfbeoordeling zelfs maar een vijfde van de tijd die een vaardigheidsmeting zou duren.

Bij het toetsen van receptieve vaardigheden biedt SA ook een voordeel ten opzichte van de vaardigheidsmetingen met betrekking tot washback. Washback is het effect dat een toets heeft op de lessen voorafgaand aan de toets, doordat een docent toetsvragen oefent met haar studenten (Andringa, 2015). Bij receptieve vaardigheden is vaak sprake van negatieve washback. Dat komt doordat lezen en luisteren niet direct getoetst kunnen worden: het 'product' dat deze activiteiten opleveren zit in het hoofd van de lezer (het begrip van de tekst). Het taalvaardigheidsniveau is voor een toetsafnemer niet direct te zien zonder begripsvragen te stellen. Deze vragen komen in de vorm van (semi-)gesloten vragen, zodat de taalleerder geen productieve vaardigheiden hoeft te gebruiken in de lees- of luistertoets (Hughes, 2002). Als een toets bijvoorbeeld veel meerkeuzevragen bevat, zal de docent daar waarschijnlijk op inspelen in de voorafgaande lessen. Zij kan haar cursisten trucs aanleren voor meerkeuzevragen (zoals 'kies bij twijfel altijd voor het langste antwoord' en 'kies voor het genuanceerde antwoord', 'kies nooit nooit', verkregen van *5 grandioze geheimen om elke meerkeuzetoets te halen!*). Deze kennis heeft natuurlijk niets te maken met het leren van de daadwerkelijke taal. SA heeft ook een washback effect, maar hierbij is het effect juist positief. Zoals we in deze paragraaf hebben kunnen lezen, heeft het oefenen van zelfbeoordelingen namelijk een positief effect op het taalleerproces. Het zorgt voor onafhankelijke taalleerders die betere resultaten halen.

Natuurlijk is een zelfbeoordeling niet in alle situaties in te zetten. Wanneer een toekomstige werkgever of universiteit vraagt om een bepaald taalvaardigheidsniveau, volstaat het niet om een eigen inschatting te maken. Daarvoor zijn de consequenties van beslissingen die gemaakt worden op basis van een taalniveau soms te hoog. Het gaat hier vaak om diploma's of certificaten die de kandidaat grote academische of aan werk gerelateerde voordelen geven. In dat geval spreken we van high stakes (Hughes, 2002). Er zijn ook toetsen waarvan de consequenties aanzienlijk minder groot zijn. De gevolgen van het wel of niet halen van zo'n toets kunnen makkelijk worden rechtgezet. Het kan dan bijvoorbeeld gaan om een plaatsingstoets, waarbij de uitkomst bepaalt in welke groep een cursist wordt gezet. In dit geval spreken we van low stakes. Kandidaten ondervinden bij een plaatsingstoets vooral hinder wanneer ze in de verkeerde groep terecht komen (doordat ze boven of juist onder hun niveau moeten werken), dus is het in hun eigen belang om de toets naar waarheid in te vullen. Bij een high stakes toets, waarbij het behalen voor aanzienlijke voordelen zorgt, kan de

zelfbeoordeling door de kandidaten iets rooskleuriger worden ingevuld dan hun daadwerkelijke kijk op het eigen leerproces. Deze aanname werd bevestigd in het onderzoek van Djiwandono (2017) waarin participanten zowel een high stakes als een low stakes toets maakten. De resultaten lagen dicht bij de docentbeoordeling wanneer ze een low stakes SA maakten, maar lagen steevast hoger dan de docentbeoordeling bij de high stakes SA.

Uit de voorgaande alinea's kunnen we concluderen dat het veel voordelen zou kunnen bieden wanneer we SA inzetten als low stakes toets. Het heeft niet alleen een positief effect op de leeropbrengst, ook zou het voordelen bieden ten opzichte van reguliere vaardigheidsmetingen. In de volgende paragraaf staan daarom de zelfbeoordelingen centraal die op dit ogenblik in gebruik zijn.

SA in de praktijk

In de inleiding staan enkele gestandaardiseerde zelfbeoordelingen al kort genoemd. Op het moment zijn er binnen vreemde- en tweedetaalverwerving drie grote projecten waar SA aan te pas is gekomen: DIALANG, ELP en het Bergen 'Can Do' project. Om een idee te geven van SA in de praktijk, worden deze drie projecten hier nader toegelicht.

DIALANG is een diagnostisch taalbeoordelingssysteem dat is gebaseerd op het Common European Framework of Reference (CEFR). Het is opgezet als een grootschalig project waar veertien Europese landen bij betrokken waren, elk met hun eigen toetsonwikkelingsteam (Alderson & Huhta, 2005). Er worden vijf vaardigheden in getoetst, namelijk lezen, schrijven, luisteren, grammatica en woordenschat. De toets bestaat uit drie delen: eerst doen de kandidaten een lexicale decisietaak, dan vullen ze een zelfbeoordeling in, om vervolgens een vaardigheidsmeting te doen waarin gatenteksten en meerkeuzevragen staan. De zelfbeoordeling bestaat uit achttien stellingen die als can do statements zijn geformuleerd, zoals 'Ik kan helder geschreven, ongecompliceerde instructies bij een apparaat goed begrijpen'. Hierop kan met 'ja' of 'nee' geantwoord worden. Deze vragen worden niet in de doeltaal gesteld, maar in een taal die de kandidaat zelf kiest. De zelfbeoordeling dient in DIALANG twee doelen. Hiermee wordt het niveau ingeschat waarop de vaardigheidsmeting zal worden aangeboden (laag, middel, of hoog). Daarnaast wordt de zelfbeoordeling diagnostisch ingezet, omdat de gegeven antwoorden worden vergeleken met de toetsresultaten om te kijken of de kandidaat een realistische inschatting heeft gemaakt van haar eigen niveau. Een foutieve inschatting zou bijvoorbeeld tot de volgende feedback leiden: "U heeft de neiging om uw eigen vaardigheid te overschatten. Dit zou u kunnen verhinderen om zinvolle en realistische doelen te formuleren voor het leren van een taal" (Verkregen uit https://dialangweb.lancaster.ac.uk/). In deze feedback wordt waarde gehecht aan een realistische zelfbeoordeling omdat het doelen helpt te formuleren in het leerproces.

Het European Language Portfolio (ELP) is een document waarin taalleerders hun progressie weer kunnen geven. De hoofddoelen van het portfolio zijn dat leerders worden geholpen vorm te geven aan hun leerervaringen, worden gemotiveerd tot verder leren en dat het portfolio kan worden gebruikt als bewijs voor het huidige taalniveau (What is the ELP?, 2020). Het

portfolio bestaat uit drie delen: een taalpaspoort, waarin de taalvaardigheden globaal staan weergegeven, een biografie, waarin informatie staat over het leergedrag, en een dossier, waarin certificaten en bewijzen zoals schrijfopdrachten zijn opgenomen. De zelfbeoordeling van de taalleerder staat in het taalpaspoort. Daarbij wordt het CEFR-grid gebruikt (Counsil of Europe, 2001) zonder aanvullende vragen; er wordt enkel gevraagd om per taalvaardigheid het niveau in te schatten. Het gaat hier om vijf taalvaardigheden in plaats van vier: de vaardigheid spreken is in 'spraakproductie' en 'interactie' opgedeeld. De SA is om meerdere redenen aan het ELP toegevoegd, die aansluiten op de hoofddoelen van het ELP (Little, 2005). Ten eerste moet de leerder op principiële gronden verantwoordelijkheid kunnen nemen in de beslissingen over haar eigen prestaties. Ten tweede is het belangrijk dat taalleerders hun eigen niveau goed in kunnen schatten, omdat beslissingen in hun leerproces anders willekeurig zijn. Ten slotte is er in het ELP gekozen voor een SA omdat een accurate zelfbeoordeling de leeropbrengst van spreeksituaties in de doeltaal verhoogt, doordat de taalleerder op die manier realistische doelen kan stellen binnen haar leerproces.

Het Bergen 'Can Do' project is ter aanvulling van het ELP opgezet. Het is specifiek voor jonge tweede- en vreemdetaalleerders ontwikkeld en geeft hun materiaal om aan het ELP toe te voegen. Het project heeft acht primaire doelen, waaronder "It incorporates self-assessment (including reflection on learning processes)" (Hasselgreen, 2003). Dit is een belangrijk onderdeel in het project, omdat de makers de autonomie van de jonge leerders willen vergroten. Hieronder wordt verstaan dat de leerders zich bewust zijn van hun eigen leergedrag, van hun eigen taalvaardigheden en van de stappen die ze ter verbetering kunnen zetten (Grahn-Saarinen, 2003). Dit zou ertoe leiden dat het leerproces van de jonge leerders wordt versneld. De SA bestaat in dit project uit twee delen, waarin alle items in de moedertaal worden aangeboden. Er is een zelfbeoordeling op basis van can do statements die aan de hand van het CEFR zijn opgesteld. Een leerder kan daarop antwoorden met 'very well', 'well', 'quite well', of 'badly'. Dit geeft antwoord op de vraag in hoeverre een vaardigheid ze goed afgaat. Daarnaast worden er reflectievragen gesteld. Die kunnen gaan over de leerder zelf ('what is important for you as a person?'), de sociale vaardigheden ('what do you feel about working in a group?'), de houding tegenover school ('why do you go to school?'), de motivatie ('do you study in order to get good marks?'), en haar vermogen zichzelf te begeleiden ('who is responsible for your studies?'). Per onderwerp worden vijf vragen gesteld, die samen antwoord geven op de vraag hoe de leerders hun doelen behaald hebben. Samengenomen vertelt deze SA ons welk niveau een taalleerder denkt te hebben en op welke manier zij dat niveau heeft behaald.

Er is een aantal duidelijke overeenkomsten tussen de zelfbeoordelingen in de drie projecten. Allereerst wordt bij alle drie als reden voor het gebruik van SA gerapporteerd dat SA een leerder helpt doelen te stellen en stappen te zetten in haar leerproces. Realistische doelen zouden op hun beurt tot een efficiënter leerproces leiden. Daarnaast zijn alle zelfbeoordelingen in deze projecten in ieder geval deels geformuleerd als can do statements die aan de hand van het CEFR zijn opgesteld. De antwoordmogelijkheden bij deze stellingen verschillen wel van elkaar. In het ELP kan een taalleerder per vaardigheid kiezen welk can do statement het beste bij diens taalsituatie past. In DIALANG moet op elk can do statement

antwoord worden gegeven met ‘ja’ of ‘nee’ en in het Bergen ‘Can Do’ project wordt ook gevraagd in hoeverre de taalleerder het met een can do statement eens is. Ook zien we dat de motivatie om een SA toe te voegen verschilt: in DIALANG wordt SA vooral inbegrepen om een leerder informatie te verschaffen over de eigen inschattingen en het juiste niveau toets aan te bieden, terwijl het ELP en Bergen ‘Can Do’ project zelfbeoordelingen ook gebruiken om de eigen verantwoordelijkheid van de leerder aan te spreken. De taalleerder staat dan niet passief in de evaluatie van haar vaardigheden, maar denkt daar actief over mee. Ook dit draagt bij aan een efficiënt leerproces.

Zelfbeoordelingen worden al in de praktijk ingezet, maar ze worden nog altijd vergezeld door een extra toets. De redenen om SA te incorporeren in een toets zijn formatief van aard: SA wordt ingezet als diagnostische toets die van invloed is op het leerproces. Een gestandaardiseerde SA die op zichzelf tot bewijs dient voor het taalvaardigheidsniveau van een taalleerder, is er niet. De vraag in het huidige onderzoek is of een zelfbeoordeling valide genoeg zou kunnen zijn om op zichzelf te staan in een low stakes toets. In de volgende paragrafen wordt uitgelegd wat precies onder validiteit wordt verstaan en welke variabelen daar invloed op hebben.

Validiteit van SA

Validiteit houdt kort gezegd in dat een toets meet wat hij beoogt te meten. De uitkomstmaat van een spreektoets moet dus informatie geven over het spreekvaardigheidsniveau van een kandidaat. Als een kandidaat tijdens deze toets veel informatie moet lezen om over te spreken, is er een kans dat de toets niet valide is. De leesvaardigheid van de kandidaat kan bijvoorbeeld van een dusdanig laag niveau zijn, dat de kandidaat tijdens een dergelijke toets niet over voldoende informatie beschikt om te kunnen spreken. In zo’n geval beïnvloedt het toetsen van leesvaardigheid dus de toetsscore van de spreektoets. De toets is dan als spreekvaardigheidstoets minder valide.
Om te meten of een zelfbeoordeling valide is als taalvaardigheidstoets, wordt die vergeleken met een criterium: “validity in self-assessment typically means agreement with teacher judgments” (Ross, 2006). Een docent geeft bijvoorbeeld een algehele inschatting van het niveau van een student, of kijkt een essay of een spreekbeurt na. Naast docentbeoordelingen worden ook reeds gevalideerde toetsen als criterum ingezet (Hughes, 2002), zoals het C2 Proficiency Exam (CPE) of de Test of English as a Foreign Language (TOEFL). Aangezien deze toetsen uitgebreid getest zijn, kunnen we daarbij uitgaan van een hoge validiteit. Alle toetsen die als criterium kunnen worden ingezet worden in het huidige onderzoek eenvoudigerwijs ‘vaardigheidsmetingen’ genoemd. In hoeverre een zelfbeoordeling correleert met zo’n vaardigheidsmeting laat zien hoe valide de zelfbeoordeling is. Hiervoor moeten de participanten beide toetsen vrijwel gelijktijdig maken. Hoe meer tijd er namelijk tussen de vaardigheidsmeting en de zelfbeoordeling zit, hoe minder valide de resultaten zijn. Wanneer er veel tijd verstreken is kan de kandidaat al vaardiger zijn geworden in de betreffende taal, of juist woordenschat of grammatica verleerd zijn. Het is daarom van belang dat de twee metingen zo gelijk mogelijk worden afgenomen.

Uit de vergelijking tussen de zelfbeoordeling en een vaardigheidsmeting komt vervolgens een correlatiecoëfficiënt die tussen de 0 en de 1 ligt. Deze coëfficiënt geeft aan hoe sterk het verband is tussen de SA en de vaardigheidsmeting waarmee die vergeleken wordt (de effectgrootte). Hoe dichter de correlatiecoëfficiënt bij de 0 ligt, hoe minder groot de correlatie is tussen de zelfbeoordeling en de gevalideerde toets. Als de correlatiecoëfficiënt dicht bij de 1 ligt, is de correlatie tussen de twee toetsen juist groot. Hoe hoger de correlatiecoëfficiënt, hoe hoger de concurrent validity. Hoe hoog de coëfficiënt moet zijn, hangt af van de hoogte van de stakes. Een correlatiecoëfficiënt van 0.7 kan bij een selectieprocedure voor een diplomatieke baan te laag zijn, terwijl diezelfde mate van samenhang tussen de zelfbeoordeling en een vaardigheidsmeting hoog genoeg is voor een plaatsingstoets (Hughes, 2002).

Wat zijn de kenmerken van een valide SA?

Om ervoor te zorgen dat een zelfbeoordeling voldoende valide is om het taalvaardigheidsniveau van kandidaten aan te tonen, is het belangrijk dat de SA zelf van een voldoende kwaliteit is. Zelfbeoordelingen kunnen onderling nogal verschillen. De ene SA bestaat uit een paar algemene vragen die met ‘ja’ of ‘nee’ kunnen worden beantwoord, terwijl de andere uit een lange lijst stellingen bestaat waarbij de antwoorden op een schaal kunnen worden aangegeven. Het is denkbaar dat ook dit een invloed heeft op de uiteindelijke resultaten. In deze paragraaf worden vier kwaliteitskenmerken geformuleerd die aan de basis staan van een zo valide mogelijke zelfbeoordeling.

Het uitgangspunt van de eerste drie kwaliteitskenmerken is het boek van Oskarsson (1980), waarin hij een format probeert op te stellen voor een gestandaardiseerde zelfbeoordeling binnen de vreemde- en tweedetaalverwerving. Dit doet hij op basis van een literatuuronderzoek waarin hij relevante, wetenschappelijke literatuur bespreekt, om vervolgens een survey-onderzoek te houden onder verschillende taalinstituten om zo de praktijk te bespreken. Dit werk heeft globaal veel invloed gehad op de uitvoering van zelfbeoordelingen (Todd, 2002). De kwaliteitskenmerken die Oskarsson (1980) noemt, zijn dat de SA is geschreven in de moedertaal, dat de vragen worden gesteld als can do statements en dat de antwoordmogelijkheden niet dichotoom zijn.

Ten eerste is het belangrijk dat de toetsitems zijn geschreven in de moedertaal en niet in de tweede of vreemde taal van de kandidaten. Als een toetsitem in de moedertaal geschreven staat, zal die voor (zeker de beginnende) taalleerders beter worden begrepen dan wanneer hij in de tweede of vreemde taal wordt aangeboden. Eerder in dit theoretisch kader is al genoemd dat een SA duidelijk moet zijn voor de kandidaten om de validiteit te verhogen (Jamrus en Razali, 2019; Hughes, 2002; Kruger & Dunning, 1999). Om dat te bereiken is het aanbieden van de items in de moedertaal van de kandidaten dus een eerste stap. Een ander argument voor het belang van het gebruik van de moedertaal, is dat het voor (ook hier vooral de beginnende) taalleerders moeilijk is om een meta-linguïstische houding aan te nemen in een taal die niet hun moedertaal is (Hasselgreen, 2003). Het beoordelen van de eigen taalvaardigheden is daarom lastiger wanneer dat in een vreemde of tweede taal moet

gebeuren. Kortom, een kwaliteitskenmerk van SA is dat de items geschreven zijn in de moedertaal van de taalleerders.

Ten tweede moet de zelfbeoordeling bestaan uit can do statements. Can do statements beschrijven communicatief gedrag, zoals ‘ik kan eenvoudige uitdrukkingen en zinnen gebruiken om de mensen die ik ken te beschrijven’ (ERK spreken A1). Can do statements gaan dus niet over algemene vaardigheden maar duiden specifieke taalsituaties aan. Zelf gebruikt Oskarsson (1980) het volgende voorbeeld: ‘I can spell my name in English’. Over het algemeen zijn taalleerders in staat om te weten of ze een dergelijke taak uit kunnen voeren (Little, 2002). Can do statements zijn concreet en daardoor kunnen participanten preciezer inschatten wat ze wel en niet kunnen (Oskarsson, 1980). Hoe concreter de items bij een zelfbeoordeling zijn, hoe accurater de kandidaten hun eigen vaardigheden in kunnen schatten (Bradshaw, 2001). Het tweede kwaliteitskenmerk van SA is daarom dat de items zijn geformuleerd als can do statements.

Ten derde moeten de antwoordmogelijkheden zich op een schaal bevinden (Oskarsson, 1980). Dat betekent dat er niet alleen met ‘ja’ of ‘nee’ geantwoord moet kunnen worden, maar dat er minimaal een neutrale mogelijkheid is. In zijn voorstel voor een gestandaardiseerde SA bestaat de schaal uit vijf opties: *Yes, Probably, Uncertain, Probably not, Definitely not*. Dit geeft de kandidaten meer ruimte tot nauwkeurigheid, ze worden niet gedwongen om te kiezen in een ja/nee dichotomie. Deze nauwkeurigheid zorgt voor een accuratere zelfbeoordeling. In het huidige onderzoek is dit kwaliteitskenmerk opgevat als dat de antwoordmogelijkheden zich op een schaal van minstens drie moeten bevinden, omdat er dan een neutrale mogelijkheid is.

Het vierde kwaliteitskenmerk komt uit een artikel van Todd (2002), die kritiek had op de kenmerken zoals Oskarsson (1980) ze had opgesteld. Hij vond dat Oskarsson had nagelaten om het onderscheid te maken tussen on-task en off-task toetsing. On-task toetsing houdt in dat de kandidaten een SA invullen, vlak nadat ze de betreffende vaardigheid zelf hebben uitgevoerd. Dit in tegenstelling tot off-task toetsing, waarbij de zelfbeoordeling niet gepaard gaat met het daadwerkelijke uitvoeren van een taalactie als lezen of schrijven. Todd (2002) beargumenteert dat kandidaten beter in staat zouden zijn om hun vaardigheden accuraat in te schatten wanneer ze die kort geleden nog hadden uitgevoerd. Deze aanname wordt bevestigd in het onderzoek van Butler & Lee (2006). Hierin deden participanten achtereenvolgens een off-task SA over hun schrijfvaardigheid, een schrijfopdracht, en een on-task SA. Het resultaat was dat een on-task SA meer overeenkwam met de 'ware' vaardigheid van de participant dan een off-task SA. Het vierde kwaliteitskenmerk is daarom als volgt: de participanten maken de SA vlak nadat ze de betreffende vaardigheid zelf hebben uitgevoerd.

Samengevat zijn er vier eigenschappen gevonden die de basis zouden vormen voor een valide SA. De zelfbeoordeling bestaat uit can do statements die zijn geschreven in de moedertaal en waarop de kandidaat minstens drie antwoordmogelijkheden heeft. Bovendien is een goede SA vlak na de uitvoering van de betreffende vaardigheid afgenomen. In de meta-analyse zal

worden onderzocht of deze kwaliteitsvoorwaarden inderdaad een significant effect hebben op de validiteit van de zelfbeoordeling.

Variabelen die van invloed zijn op de validiteit van een SA

Om te begrijpen waarom een zelfbeoordeling minder valide zou zijn ten opzichte van een vaardigheidsmeting, is het belangrijk om te weten welke variabelen hier invloed op hebben (confounders). In bovenstaande tekst zijn de stakes van de toets al genoemd. Bij high stakes zou een rooskleuriger beeld van de taalvaardigheden verwacht kunnen worden dan bij low stakes, omdat daar voor de kandidaat meer van afhangt (Djiwandono, 2017). De context waarin de toets wordt gegeven - of het gaat om een plaatsingstoets (low stakes) of een eindexamen (high stakes) - heeft invloed op de validiteit van de SA. Het zou ook uitmaken welke taalvaardigheid getoetst wordt; kandidaten beoordelen zichzelf nauwkeuriger bij hun receptieve vaardigheden dan bij de productieve vaardigheden (Ross, 1998). Naast de hoogte van de stakes van een toets, zijn ook verschillende kandidaateigenschappen van belang voor de accuraatheid van de zelfbeoordeling. In deze paragraaf worden achtereenvolgens de leeftijd, het geslacht, de achtergrond, het opleidingsniveau en het taalvaardigheidsniveau van de taalleerder besproken. Daarnaast wordt het effect van de gemeten taalvaardigheid besproken, het verschil tussen de leerders van een tweede taal en van een vreemde taal, en de voorgaande ervaring van de kandidaten met SA.

Leeftijd is een van de kandidaateigenschappen die een effect hebben op de accuraatheid van SA. Mensen worden naarmate ze ouder worden namelijk beter in zelfbeoordelingen. Dit zien we niet alleen bij zelfbeoordelingen over taal - zoals in het onderzoek van Bradshaw (2001), waarin kinderen van 10 tot 11 jaar oud hun leesvaardigheid nauwkeuriger konden inschatten dan kinderen van 8 tot 9 jaar - maar in zelfbeoordelingen binnen meerdere velden. Butler (1990) vond dat kinderen van 10 tot 11 jaar hun tekeningen beter konden beoordelen dan kleuters, Hewitt (2005) zag dat kinderen van 15 tot 18 jaar hun muzikale vaardigheden accurater in konden schatten dan kinderen van 11 tot 14 jaar oud, en Blatchfords (1997) longitudinale onderzoek toonde aan dat ook sociale vaardigheden beter worden ingeschat naarmate men ouder was. Leeftijd heeft dus invloed op de validiteit van zelfbeoordelingen.

Ook het geslacht van de kandidaten lijkt een rol te spelen in de accuraatheid van hun zelfbeoordelingen. Wat voor een invloed dat heeft, hangt af van het domein waarin de SA wordt afgenomen. In een meta-analyse van Huang (2013) over 247 onafhankelijke studies kwam naar voren dat mannen zichzelf hoger inschatten dan vrouwen wat betreft wiskunde en natuurkunde, maar dat dit binnen taal- en artistieke vaardigheden juist andersom was. Binnen de lees- en luistervaardigheid schatten vrouwen zich bijvoorbeeld accurater in dan mannen (Denies & Janssen, 2016). Deze verschillen tussen man en vrouw zijn leeftijdsafhankelijk; boven de 21 jaar zijn ze het grootst (Huang, 2013). Een verklaring hiervoor is dat van mannen en vrouwen al van jongs af aan verschillende dingen worden verwacht, en dat dit weerklinkt in hun opvoeding en educatie. Bij vrouwen zou de focus meer liggen op de sociale vaardigheden, bij mannen juist op de exacte wetenschappen (Denies & Janssen, 2016).

Binnen de vreemde- en tweedetaalverwerving verwachten we dus dat vrouwen zichzelf accurater beoordelen dan mannen.

Ook de achtergrond van de taalleerders heeft invloed op de resultaten van de SA. Binnen sommige culturen zou de neiging bestaan om zichzelf te overschatten, terwijl dat bij andere culturen juist andersom zou zijn (Blue, 1994). Bij wiskundige sommen zijn Taiwanese studenten significant beter in zelfbeoordelingen dan studenten uit de VS (Chen & Zimmerman, 2007) en kunnen Aziatisch-Amerikaanse studenten zichzelf strenger beoordelen dan andere Amerikaanse studenten (Whang & Hancock, 1994). Binnen taaleducatie is ditzelfde fenomeen gevonden: Japanse studenten zijn significant beter in zelfbeoordelingen over hun Engelse vocabulaire dan studenten uit Israël (Laufer & Yano, 2001). Janssen-van Dieten (1989) vond in haar analyse van 73 verschillende achtergronden maar een klein effect op de SA van lezen, schrijven, luisteren en lezen. Dat effect werd groter wanneer ze haar participanten verdeelde in westers en niet-westers. De reden daarvoor was waarschijnlijk dat de westerse participanten hoger opgeleid waren dan de niet-westerse participanten.

Opleidingsniveau is namelijk ook van invloed op de zelfbeoordelingen. Het zojuist genoemde onderzoek van Janssen-van Dieten (1989) vond onder de 730 participanten een significante, positieve correlatie tussen opleidingsniveau en accuraatheid van de SA. Deze invloed van opleidingsniveau is ook in andere experimenten aangetoond (Barnett & Hixon, 1997; Claes & Salame, 1975). Een verklaring hiervoor wordt aangedragen in het onderzoek van Laveault en Miles (2002). Daarin scoren de studenten die een schematische weergave van een beoordelingsvoorschrift konden gebruiken beter op de zelfbeoordeling van schrijven dan studenten die dat niet goed konden: “Students who are more ‘severe’ (…) are usually the most competent users of rubrics” (Leveault & Miles, 2002). De gedachte hierachter is dat abstract denken een belangrijke rol speelt in het begrijpen van evaluatiecriteria en dat mensen met een hoog opleidingsniveau daar beter in zijn dan die met een laag, meer praktijkgericht opleidingsniveau.

Het taalvaardigheidsniveau van de kandidaten heeft ook een effect op hoe accuraat de zelfbeoordelingen zijn. Zoals in de inleiding al genoemd is, lijken mensen over het algemeen niet goed te zijn in het inschatten van hun eigen vaardigheden. Hier speelt het ‘Dunning-Kruger effect’ een rol. Dit effect houdt in dat mensen met weinig kennis van een bepaald onderwerp zich in een zelfbeoordeling zullen overschatten: “incompetence (…) not only causes poor performance but also the inability to recognize that one's performance is poor” (Kruger & Dunning, 1999). Daartegenover staat dat experts zich in het algemeen juist onderschatten. Dit wordt toegeschreven aan het idee dat zij het succes van anderen erkennen en daardoor een bescheiden zelfbeoordeling maken. Binnen SA in tweede- en vreemdetaalverweving is dit effect bijvoorbeeld zichtbaar bij uitspraaktoetsen (Trofimovich, Isaacs, Kennedy, Saito, & Crowther, 2016). Taalleerders met een zwaar accent en lage verstaanbaarheid schatten hun niveau hoger in dan hun docenten ze inschatten, omdat ze zelf het verschil in accenten nog niet goed kunnen horen. Taalleerders met een licht accent en een hoge verstaanbaarheid schatten zichzelf juist lager in, omdat ze al taalvaardig genoeg zijn om

de verschillen goed te kunnen horen. Met andere woorden: beginnende taalleerders zullen zichzelf in het algemeen overschatten, terwijl ervaren taalleerders zichzelf onderschatten.

De taalvaardigheid die wordt gemeten lijkt ook van belang te zijn. In de introductie was al te lezen hoe sociale vaardigheden minder accuraat werden ingeschat dan bijvoorbeeld bij sport, maar ook de taalvaardigheden verschillen onderling van elkaar. Uit de meta-analyse van Ross (1998) blijkt bijvoorbeeld dat SA bij lezen en luisteren een hogere gemiddelde correlatie heeft met vaardigheidsmetingen dan bij de productieve vaardigheden. Ook was de *range* bij spreekvaardigheid groter; de minimum en maximum gevonden correlaties liggen verder uit elkaar dan bij de andere vaardigheden. Dat zou erop kunnen wijzen dat de zelfbeoordeling bij spreekvaardigheid meer onderhevig zou kunnen zijn aan externe factoren (Ross, 1998).

Tot nu toe is er in dit theoretisch kader consequent gesproken over zowel vreemde- als tweedetaalverwerving. Deze twee komen dan ook grotendeels overeen, in de zin dat de taalleerders een extra taal leren, naast hun moedertaal. Het verschil zit in de context waarin ze de taal leren. Een tweede taal wordt geleerd in het desbetreffende taalgebied, een vreemde taal daarbuiten. Wie bijvoorbeeld Frans leert op een Nederlandse middelbare school, leert een vreemde taal. Wanneer deze (Nederlandssprekende) leerling naar Frankrijk gaat en daar Frans leert, leert zij een tweede taal. Dit heeft ook een effect op de accuraatheid van een zelfbeoordeling (Blue, 1994). Wie een tweede taal leert, vergelijkt zichzelf met de moedertaalsprekers van de betreffende taal, terwijl een vreemdetaalleerder voornamelijk contact heeft met andere niet-moedertaalsprekers van deze taal. De vreemdetaalleerder zal haar taalvaardigheid waarschijnlijk hoger inschatten dan de tweedetaalleerder, omdat het taalniveau van de sprekers rondom een vreemdetaalleerder in het algemeen lager zal zijn.

Ook de manier van het introduceren van een zelfbeoordeling kan effect hebben op de validiteit ervan. Door de kandidaten te laten oefenen met SA en ze duidelijke instructies te geven, kan die ook meer valide worden (Kruger & Dunning, 1999). Dat betekent dat de ervaring van een kandidaat met zelfbeoordelingen ook een variabele is met invloed op de resultaten. Uit verschillende onderzoeken blijkt inderdaad dat de resultaten van SA accurater worden wanneer de deelnemers daar meer ervaring mee hebben (zoals Jamrus & Razali, 2018; Lappin-Fortin & Rye, 2014; Butler & Lee, 2010; Chen, 2008). In het onderzoek van Chen (2008) werd aan studenten gevraagd om hun eigen spreekvaardigheid te beoordelen. Dit onderzoek duurde twaalf weken; aan het einde van deze periode waren de studenten significant beter geworden in hun zelfbeoordelingen. Ook experimenten met een controlegroep lieten ditzelfde effect zien (Jamrus & Razali, 2018; Butler & Lee, 2010).

Kortom, we zien dat er behoorlijk wat variabelen invloed hebben op een zelfbeoordeling. Naast de hoogte van de stakes van een SA, kunnen ook karaktereigenschappen als leeftijd, geslacht, achtergrond, opleidingsniveau en taalvaardigheidsniveau een rol spelen bij de accuraatheid van een zelfbeoordeling. Ook de gemeten taalvaardigheid, de talige context waarin de kandidaten zich bevinden en hun voorgaande ervaring met SA lijken een effect te hebben op zelfbeoordelingen. In het tweede deel van dit onderzoek (de meta-analyse) zal bekeken worden of deze variabelen inderdaad een significant effect hebben op de validiteit.

Methode

Meta-analyse

Er is gekozen voor een meta-analyse om de centrale vraag te beantwoorden. Een empirische studie had ook bij kunnen dragen aan de beantwoording van deze vraag, maar was noodzakelijkerwijs kleinschalig geweest. Bovendien zijn er al behoorlijk veel experimenten gedaan om zelfbeoordeling te onderzoeken. Dit baant de weg voor een meta-analyse waarin deze experimenten samen worden gevat in een overkoepelend onderzoek. Een meta-analyse is een belangrijke onderzoeksmethode om individuele onderzoeksresultaten te generaliseren (Field & Gillett, 2010; Lipsey & Wilson, 2001). In deze paragraaf wordt uitgelegd wat een meta-analyse inhoudt en wat de voordelen ervan zijn. Ook worden er eerdere meta-analyses over zelfbeoordelingen bij tweede- en vreemdetaalleerders besproken.

Een meta-analyse is een methode waarin meerdere, empirische onderzoeken worden behandeld. In deze onderzoeken wordt een antwoord gezocht op dezelfde vraag. In het huidige onderzoek is dat bijvoorbeeld de vraag naar de concurrent validity van zelfbeoordelingen bij tweede- of vreemdetaalleerders. In principe kan een meta-analyse al over twee onderzoeken gaan, zolang ze maar een zo compleet mogelijke set opneemt van de resultaten binnen het betreffende gebied (Rosenthal & DiMatteo, 2001). Simpel gezegd combineert een meta-analyse onderzoeksresultaten tot een gewogen gemiddelde van een set correlaties (Oswald & Plonsky, 2010). Deze cijfers zijn geen individuele scores, zoals bij andere onderzoeksmethodes gebruikelijk is, maar zijn statistische gegevens uit ander onderzoek. De cijfers die als afhankelijke variabele gebruikt worden in meta-analyses zijn effectgroottes, om precies te zijn. Daarvoor kunnen verschillende gestandaardiseerde maten gebruikt worden (zoals de correlatiecoëfficiënt, de standardized mean difference, of de odds ratio), zolang die maar de grootte en de richting van een effect weergeven (Lipsey & Wilson, 2001).

Het meest voor de hand liggende voordeel van meta-analyses is dat ze een algemeen beeld geven van de resultaten binnen een bepaald onderzoeksveld. Met een meta-analyse vertrouwen we niet op de resultaten van een enkele studie om een fenomeen te begrijpen (Oswald & Plonsky, 2010). De resultaten van meerdere experimenten kunnen worden meegenomen in de statistische analyse. Daardoor kan een dergelijk onderzoek niet alleen een groot aantal participanten bevatten, maar wordt ook het aantal onderzoekers groter en is er meer variatie in meetinstrumenten. Daarmee beschermen ze tegen overhaaste conclusies op basis van ‘losse’ experimenten (Lipsey & Wilson, 2001; Rosenthal & DiMatteo, 2001).

Bovendien kunnen door meta-analyses patronen worden ontdekt die verborgen blijven in andere vormen van onderzoek (Lipsey & Wilson, 2001). Hierbij kan bijvoorbeeld gedacht worden aan onderzoeken die conflicterende resultaten rapporteren. In een meta-analyse worden deze studies met elkaar vergeleken en kan de onderzoeker bekijken of er een verschillende aanpak was. Kleinere onderzoeken kunnen dan niet alleen op een betekenisvolle

wijze bijdragen aan een groter onderzoek, ook kan worden onderzocht of er eventuele confounders zijn die de tegengestelde resultaten veroorzaken (Rosenthal & DiMatteo, 2001). In het huidige onderzoek zou het bijvoorbeeld kunnen dat de participanten uit twee conflicterende studies uit verschillende landen komen. Zoals eerder in het theoretisch kader staat beschreven, is dit een variabele die invloed kan hebben op het uiteindelijke resultaat. Een meta-analyse biedt hier dus een mogelijkheid om patronen te vinden.

Het voordeel van meta-analyses is ook dat de focus minder op significantie komt te liggen en meer op de grootte van een effect (Lipsey & Wilson, 2001). Het bezwaar bij NHST (null-hypothesis significance testing) is dat de resultaten worden gereduceerd tot een dichotomie: wel of niet significant (Oswald & Plonsky, 2010). De significantie wordt door vier cijfers bepaald: sample mean difference, sample variances of the groups, alpha level en de groepsgrootte (Oswald & Plonsky, 2001). Van deze vier is vooral de groepsgrootte aan verandering onderhevig. Dat heeft invloed op het significantieniveau. Twee studies die bijvoorbeeld precies dezelfde effectgroottes vermelden, kunnen ogenschijnlijk tegenstrijdige resultaten laten zien wanneer het aantal participanten van elkaar verschilt (Lipsey & Wilson, 2001; Rosenthal & DiMatteo, 2001). In een meta-analyse wordt de aandacht voor deze dichotomie verschoven naar de grootte en de richting van het gevonden effect.

De nadelen van meta-analyses zijn paradoxaal genoeg zowel dat alle resultaten worden meegenomen, als dat niet alle resultaten meegenomen kunnen worden. Het eerste deel van deze schijnbare tegenstelling wordt door Rosenthal en DiMatteo (2001) ook wel omschreven als “garbage in, garbage out”. Dit houdt in dat meta-analyses zowel goede als slechte onderzoeken includeren. De kwaliteit van een meta-analyse is dus zo goed als de kwaliteit van de onderzoeken die zij opneemt. Dit zou opgelost kunnen worden door de effectgroottes te wegen aan de hand van de kwaliteit van het betreffende onderzoek, of bepaalde onderzoeken op basis van hun methode te blokkeren (Rosenthal & DiMatteo, 2001). Het tweede nadeel is dat niet alle resultaten kunnen worden meegenomen omdat ze niet zijn gepubliceerd. Dit wordt ook wel publication bias genoemd (Plonsky & Oswald, 2012; Lipsey & Wilson, 2001; Rosenthal & DiMatteo, 2001). Niet alle onderzoeken worden daadwerkelijk gepubliceerd, zeker wanneer het resultaat ervan niet significant is. Dit geeft een vertekend beeld van de werkelijkheid. In een meta-analyse kan door middel van een funnel plot worden beredeneerd of er sprake is van publication bias (Field & Gillett, 2010).

Het huidige onderzoek is de derde meta-analyse die is uitgevoerd naar de validiteit van zelfbeoordelingen binnen vreemde- en tweedetaalverwerving. Het eerste onderzoek is uitgevoerd door Blanche en Merino (1989), daarna werd een onderzoek gedaan door Steve Ross (1998) dat nog steeds veel wordt geciteerd. In de volgende alinea’s worden deze twee meta-analyses besproken.

Het onderzoek van Blanche en Merino (1989) is uitgevoerd in de tijd dat SA als onderwerp “just begun to expand as a distinct field of interest in language testing and evaluation” (Blanche & Merino, 1989). In deze meta-analyse worden zestien onderzoeken besproken, waarvan er maar zeven een correlatiecoëfficiënt rapporteerden. Er is daarom ook geen

gemiddelde effectgrootte berekend; de onderzoekers gaven enkel aan dat de meeste van deze studies een correlatiecoëfficiënt tussen de 0,5 en 0,6 rapporteerden. De andere studies zijn kwalitatief besproken. Het hoofdresultaat uit deze meta-analyse is dat er een redelijke samenhang is tussen SA en vaardigheidsmetingen. Ook vonden Blanche en Merino (1989) dat de accuraatheid van de zelfbeoordelingen afhankelijk was van de taalvaardigheid van de participanten.

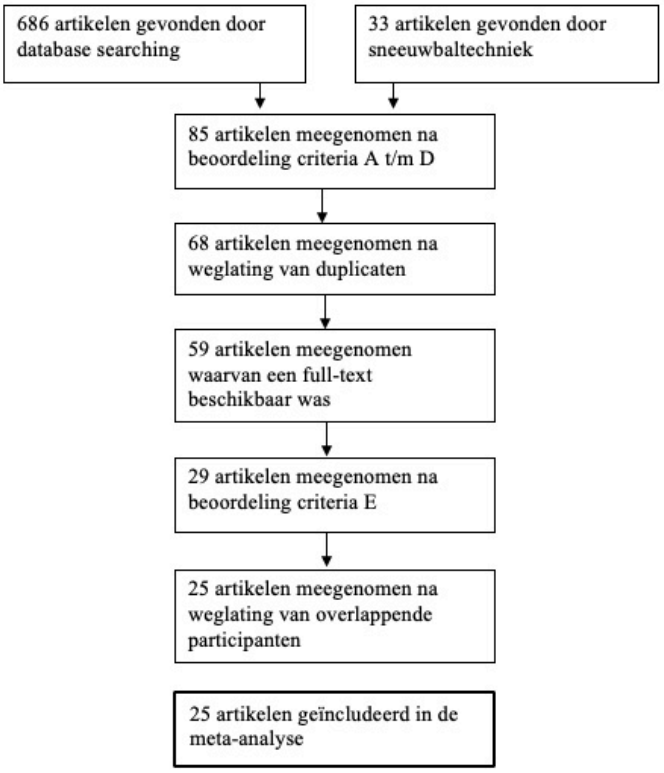
Negen jaar later is de meta-analyse van Ross (1998) gepubliceerd. Daarin zijn elf onderzoeken opgenomen, die samen zestig effectgroottes rapporteerden. De gemiddelde correlatiecoëfficiënt is 0,63. Er is veel variatie gevonden tussen de verschillende effectgroottes: de range liep van 0,09 tot 0,8. Daarvoor zijn verschillende redenen mogelijk; Ross (1998) onderzoekt in zijn meta-analyse de invloed van de gemeten vaardigheid. De studies waarin de participanten hun receptieve taalvaardigheden moesten beoordelen, bleken hogere effectgroottes te rapporteren dan de studies waarin de productieve vaardigheden werden getest. In zijn conclusie staat het volgende: “the literature base for self-assessment in second language learning is not extensive, but fortunately is growing at a steady rate” (Ross, 1998).

Het aantal onderzoeken dat is gedaan naar de validiteit van SA, is dus gegroeid. Sinds de publicatie van de meta-analyse van Steve Ross (1998) is er veel onderzoek geweest naar het onderwerp. Het was daarom interessant om een geüpdatete versie te maken, waarin onder andere nieuwere vormen van toetsing mee konden worden genomen. Ook worden in het huidige onderzoek meer potentiële confounders besproken, om eventuele patronen te kunnen herkennen.

Verzamelen van data

Om een meta-analyse naar behoren uit te voeren moet een verzameling van onderzoeken met dezelfde onderzoeksvraag worden gedefinieerd. De experimenten in die onderzoeken moeten tevens een vergelijkbare uitkomstmaat hebben. Daarvoor worden criteria opgesteld waaraan deze onderzoeken moeten voldoen. Deze criteria zijn idealiter breed genoeg om voldoende onderzoeken toe te laten in de analyse, maar zijn bovenal precies genoeg om scheve vergelijkingen te voorkomen (Oswald & Plonsky, 2010). Dit in het achterhoofd houdende, zijn de volgende criteria opgesteld: (a) het gaat om empirisch onderzoek dat (b) de validiteit van zelfbeoordeling meet door (c) T2-leerders (d) twee toetsen te laten maken (een zelfbeoordeling en een taalvaardigheidsmeting) om daar ten slotte (e) een effect size over te rapporteren (of voldoende data om een effect size af te leiden).

Het zoeken naar relevante artikelen gebeurde aan de hand van key word searches. In drie academische databases - Linguistics and Language Behavior Abstracts (LLBA), the Educational Information Resource Center (ERIC) en Google Scholar – werd gezocht op 'self assessment' dan wel 'self evaluation', in combinatie met 'second language learning'. Ook werd de sneeuwbaltechniek toegepast door in de literatuurlijsten van gevonden artikelen te zoeken naar nieuwe bronnen. Deze zoektocht eindigde toen de databases naar alle waarschijnlijkheid uitgeput waren (er werden geen nieuwe bronnen meer gevonden). De vele honderden artikelen die hieruit kwamen werden beoordeeld op criteria a-e en daarnaast werd gekeken of onderzoeken van eenzelfde onderzoeker niet deels overlapten. Daarbij was het vooral belangrijk dat dezelfde participanten niet in verschillende onderzoeken voorkwamen. Dit proces (zie figuur 1) resulteerde in een verzameling van 25 artikelen die gebruikt konden worden in het huidige onderzoek.



Figuur 1: Flowchart van het zoekproces naar relevante artikelen

Coderen van data

Na het verzamelen van de data werden de artikelen gecodeerd in een bestand waarin om verschillende kenmerken van de onderzoeken werd gevraagd. In dit bestand werd het aantal participanten en de gerapporteerde correlatiecoëfficiënt opgenomen omdat deze cijfers gebruikt worden in de meta-analyse. Daarnaast werd gevraagd om de T1 en T2 van de participanten, hun taalvaardigheidsniveau, hun opleidingsniveau, hun leeftijd en de gemeten taalvaardigheid. Deze variabelen zouden namelijk van invloed kunnen zijn op de validiteit van de SA, zoals beschreven in het theoretisch kader. Wanneer een eigenschap niet expliciet stond aangegeven in de tekst, is die door de auteur afgeleid uit verder gegeven informatie. In die gevallen is ook de oorspronkelijke aanduiding opgenomen in het bestand. De T1 stond maar in enkele gevallen expliciet aangegeven; meestal werd alleen gerapporteerd in welk land of aan welke universiteit het onderzoek plaatsvond. Aangezien het land waarin een participant woonachtig is niet direct een bepaalde moedertaal impliceert, is een extra kolom toegevoegd.

De SA's uit de verschillende experimenten zijn beoordeeld op de vier kwaliteitskenmerken zoals benoemd in het theoretisch kader:

1. De toets is geschreven in de moedertaal van de participant;
2. De zelfbeoordeling bestaat uit can do statements;
3. De antwoordmogelijkheden bevinden zich op een schaal van minstens drie;

4. De zelfbeoordeling is on-task afgenomen.

Deze vier kwaliteitskenmerken zijn per artikel beoordeeld door de auteur van het huidige onderzoek en een medestudent van de Master Taalwetenschap. Daarbij konden ze kiezen uit verschillende opties. Bij het eerste kenmerk (De toets is geschreven in de moedertaal van de participant) kon worden gekozen uit: *Ja, Nee, Waarschijnlijk wel, Waarschijnlijk niet, Niet vermeld.* De mogelijkheden bij het tweede kwaliteitskenmerk (De zelfbeoordeling bestaat uit can do statements) waren: *Ja, Nee, Niet vermeld.* Het derde kwaliteitskenmerk (De antwoordmogelijkheden bevinden zich op een schaal van minstens drie) had vier opties: *Ja, Nee, Ander soort vragen, Niet vermeld.* De vierde (De zelfbeoordeling is on-task afgenomen) had er ook vier: *Ja, Nee, Beide, Niet vermeld.*

De interbeoordelaarsbetrouwbaarheid tussen de auteur en de medestudent was 80%. De twee codeurs verschilden het meest van mening wat betreft het eerste kwaliteitskenmerk. Aangezien de moedertaal niet in elk artikel expliciet werd benoemd en ook de taal van de toets niet altijd werd aangegeven, konden niet alle onderzoeken precies worden gecodeerd. Het tweede en het derde kenmerk konden redelijk eenvoudig worden geïnterpreteerd, meestal werden die expliciet genoemd of waren ze te vinden in de bijlagen. Het vierde kwaliteitskenmerk leverde wel wat onduidelijkheid omdat de volgorde van de toetsen niet altijd werd vermeld. Er waren in totaal 20 punten waarop de twee beoordelaars het in eerste instantie niet eens waren. Daarvan ging het in 70% van deze verschillen om een beoordelaar die iets voorzichtiger was dan de ander (zoals een *Niet* en een *Waarschijnlijk niet)*. Alle 20 gevallen zijn door de beoordelaars nog eens samen doorgenomen om tot consensus te komen. Deze uiteindelijke cijfers zijn gebruikt in de analyse.

Analyse

In de meeste onderzoeken werd één effect size gerapporteerd om aan te geven in hoeverre de zelfbeoordeling samenhing met de vaardigheidsmeting. Deze score kon dan vrij simpel worden toegevoegd aan de huidige meta-analyse. Er was echter een aantal artikelen waarin meerdere effect sizes werden gerapporteerd. Dit gebeurde vaak wanneer er in het betreffende onderzoek meerdere vaardigheden waren gemeten. In een ander onderzoek (Laufer & Yano, 2001) waren twee self-assessments gedaan, een off-task en een on-task. In dit soort gevallen bestaan er twee oplossingen (Lipsey & Wilson, 2001). De eerste is om een willekeurige effect size uit het artikel te halen als representatief voor het hele artikel en bij de tweede aanpak wordt het gemiddelde van de effect sizes berekend. Aangezien er bij de eerste aanpak op een systematische manier waardevolle data verloren gaat, is in het huidige onderzoek gekozen om in dit soort gevallen het gemiddelde te berekenen. Dat gemiddelde wordt dan in de meta-analyse behandeld als een zelfstandige effect size. Wanneer naar deelvaardigheden werd gekeken, werd wel de specifieke correlatiecoëfficiënt gebruikt.

In sommige gevallen is er niet gekozen voor het berekenen van een gemiddelde. In één artikel is een willekeurige effect size genomen, omdat het niet duidelijk was in hoeverre de participanten overlapten in de metingen van de verschillende vaardigheden (Janssen-van

Dieten, 1989). Vandaar dat hier een willekeurige vaardigheid (spreken) is gekozen. Er was ook een speciaal type artikel, waarbij bewust is gekozen voor een van de gerapporteerde effect sizes. In een aantal artikelen werd namelijk het effect van een bepaald traject gemeten op de overeenkomstigheid van zelfbeoordelingen en vaardigheidsmetingen. De participanten kregen daarin bijvoorbeeld training in het beoordelen van hun eigen kwaliteiten. In dit type artikel was er dan een nulmeting en een posttest. In deze gevallen is altijd voor de nulmeting gekozen, omdat de nog ongetrainde participanten meer overeen kwamen met de participanten uit de andere onderzoeken.

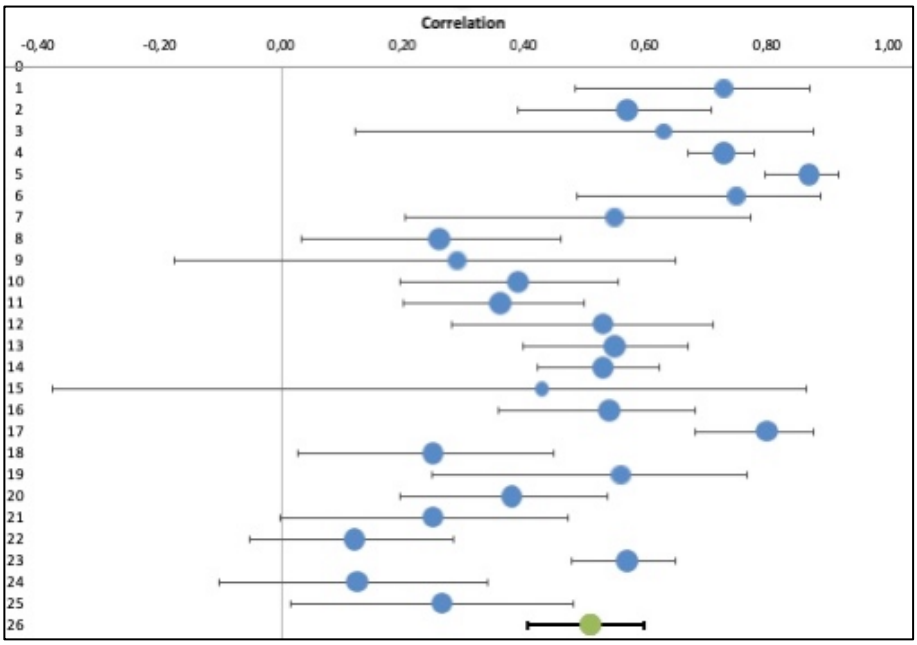
De effect sizes die tijdens het coderen van de data zijn gevonden, zijn overwegend gerapporteerd als Pearsons *r* en Spearmans rho. Wanneer deze niet waren gegeven, is de Pearsons *r* berekend aan de hand van de gegeven resultaten. Deze cijfers zijn geanalyseerd met behulp van het correlational data workbook uit de *Meta-Essentials tools* (Suurmond, Van Ree & Hak, 2017). In dit werkboek worden de effect sizes allereerst getransformeerd naar Fishers *z.* Deze correlatiecoëfficiënt neigt namelijk – meer dan Pearons *r* en Spearmans rho – naar een normale verdeling en de r-to-z transformatie is variantiestabiliserend (Van Ree, Suurmond & Hak, 2015). De meta-analyse wordt gedaan met deze Fishers *z,* daarna worden de cijfers weer getransformeerd naar de 'normale' effect size *r* voor een duidelijke presentatie. Ook de analyses voor de subvragen (de one-way ANOVA en de onafhankelijke t-test) zijn gedaan met behulp van Fishers *z.* In de analyses is uitgegaan van het aantal participanten, zodat een experiment met 289 participanten (Barrow, Nakanishi & Ishino, 1999) niet evenveel invloed zou hebben op de resultaten als een experiment met 10 participanten (Liu & Brantmeier, 2019).

Resultaten

In totaal zijn de gegevens van 2138 participanten meegenomen, die zijn verkregen uit 25 onderzoeken. Deze studies en enkele van hun gegevens staan weergegeven in bijlage 1. De studiegroottes en effect sizes zijn gebruikt om tot de volgende analyse te komen.

Homogeniteitsanalyse en publication bias

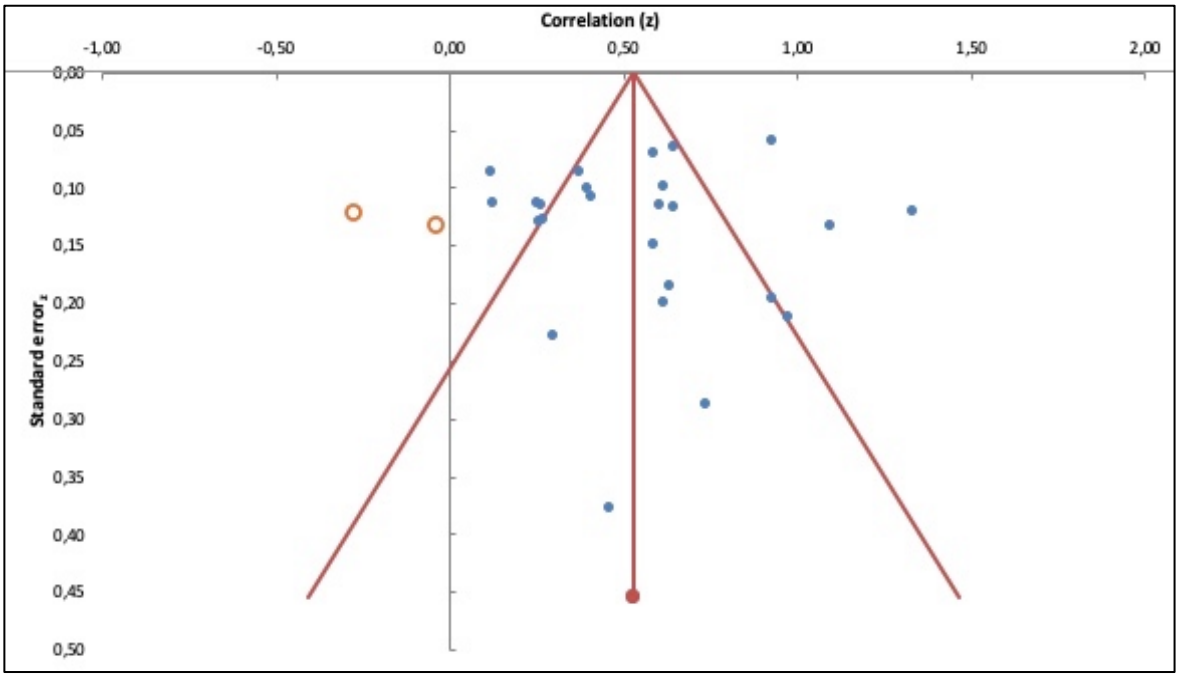
Een homogeniteitsanalyse is uitgevoerd om te bepalen of de in de meta-analyse opgenomen experimenten vergelijkbare effect sizes rapporteerden. Hier kwam een significant resultaat uit, $Q(24) = 246$, $p < 0.01$, wat betekent dat homogeniteit niet kon worden aangenomen. De verzameling opgenomen onderzoeken is dus heterogeen. Dit is ook te zien in de forest plot in figuur 2. De 25 gerapporteerde effect sizes zijn hier met een 95% betrouwbaarheidsinterval weergegeven. Het zesentwintigste punt in deze figuur geeft de gemiddelde effect size ($r = 0.52$) weer. De meeste individuele effect sizes liggen hier niet bij in de buurt en vallen ook niet binnen het 95% betrouwbaarheidsinterval van de gecombineerde effect size.



Figuur 2: Forest plot van de gerapporteerde effect sizes (r)

Een mogelijke verklaring voor deze heterogeniteit is publication bias. Publication bias is een risico bij meta-analyses. Deze bias komt voort uit het feit dat significante resultaten vaker geplubliceerd worden dan niet-significante resultaten. Het gevolg hiervan kan zijn dat een meta-analyse een grotere effect size vindt dan de werkelijke samenhang tussen een zelfbeoordeling en een vaardigheidsmeting. De funnel plot in figuur 3 geeft een algemene indruk van de verspreiding van de gerapporteerde effect sizes. De stippen zijn de 25 geanalyseerde onderzoeken. Op de horizontale as is de effect size (in Fisher's z) aangegeven. Op de verticale as staat de standaardafwijking. Wanneer er geen sprake is van publication bias, zijn de resultaten symmetrisch rond de gemiddelde effect size verdeeld. Een funnel plot is overzichtelijk wanneer de studies met de meeste power in de top van de grafiek staan

weergegeven, daarom wordt de verticale as omgekeerd (Sterne & Egger, 2001). Deze studies hebben namelijk een kleine standaardafwijking. Bij studies met minder power, waar meer afwijkingen van het gewogen gemiddelde verwacht kunnen worden (Field, 2010), zie je dat ook de grootte van de effect size schommelt. Die kan zowel lager als hoger liggen ten opzichte van het gemiddelde. Hoe groter de power, hoe dichter de effect size bij het gemiddelde zou moeten liggen. Een homogene groep van studies zonder publication bias zou daarom in een trechtervorm passen. Wanneer er wel sprake zou zijn van publication bias, zouden bijvoorbeeld de lagere effect sizes niet gerapporteerd zijn en dus niet weergegeven in de grafiek (Sterne & Egger, 2001).



Figuur 3: Funnel plot van effect sizes (z) en standard error

In figuur 3 is te zien dat vrijwel evenveel gerapporteerde effect sizes links en rechts van het gewogen gemiddelde (z = 0.53) liggen. Slechts twee punten wijken af van dit beeld (in deze figuur zijn die ook gespiegeld weergegeven in het oranje). Deze symmetrie zou als bewijs kunnen worden gezien dat er geen sprake is van publication bias. Wel valt op dat veel datapunten buiten de trechtervorm liggen. Dat komt doordat de verzameling onderzoeken niet homogeen is. Hier zouden de variabelen zoals ze zijn genoemd in het theoretisch kader (de hoogte van de stakes van een toets; de kwaliteit van de verschillende zelfbeoordelingen; de leeftijd, het geslacht de achtergrond, het opleidingsniveau, en het taalvaardigheidsniveau van de kandidaat; de gemeten taalvaardigheid, de talige context waarin de kandidaten zich bevinden en hun voorgaande ervaring met SA) een rol kunnen spelen. In de opgenomen onderzoeken is voldoende data gevonden om het effect van de kwaliteit van de zelfbeoordelingen, de gemeten taalvaardigheid, het taalvaardigheidsniveau en de talige context waarin de kandidaten zich bevinden te kunnen meten.

Hoofdvraag

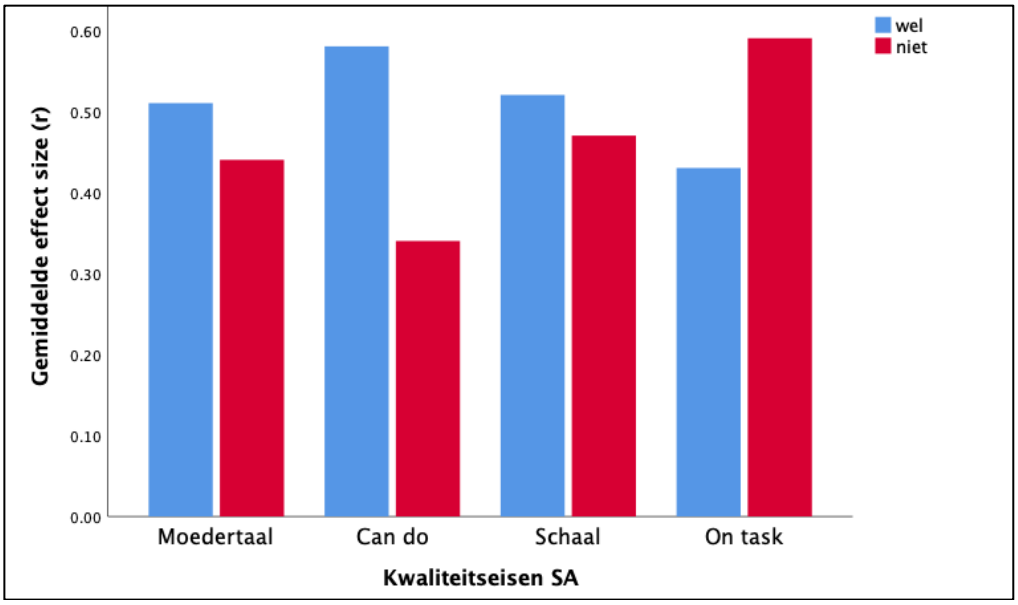
De hoofdvraag van dit onderzoek was 'In hoeverre zijn zelfbeoordelingen binnen vreemde- en tweedetaalverwerving een valide methode om taalvaardigheden te meten?' De effect sizes liepen van 0.12 tot 0.8 (zie bijlage 1) en de gemiddelde effect size (r = 0.52, zie tabel 1) laat de correlatie zien tussen zelfbeoordelingen en vaardigheidsmetingen. De determinatiecoëfficiënt r^2 toont aan dat die overlap 27% van de individuele verschillen beslaat.

Tabel 1: Gemiddelde effect size van 2138 participanten uit 25 studies

				95% confidence interval	
	r	r^2	SE	Lower	Upper
SA vs. vaardigheidsmeting	0.52	0.27	0.02	0.49	0.55

Kwaliteit van de SA

De eerste subvraag in het huidige onderzoek was 'Wat is het effect van de kwaliteit van de zelfbeoordeling op de validiteit?' Daarbij zijn vier kwaliteitskenmerken geformuleerd: de zelfbeoordeling moet in de moedertaal worden afgenomen, de items zijn can do statements, de antwoordmogelijkheden zijn op een schaal aangeboden, en de toets is on-task afgenomen. In de analyse zijn de effect sizes van 25 onderzoeken met elkaar vergeleken (n = 2138) aan de hand van het wel of niet bezitten van de vastgestelde kenmerken. De grafiek in figuur 4 geeft deze resultaten weer.



Figuur 4: Vergelijking van de gemiddelde effect size (r) tussen de kwaliteitskenmerken

Zoals te zien gedragen de gemiddelde effect sizes zich bij drie van de vier kwaliteitskenmerken zoals verwacht. De onderzoeken die deze kenmerken bezitten, rapporteren een hogere effect size dan de onderzoeken die dat niet doen. Bij de 'on task' is

een tegengesteld resultaat te zien. Er werd gemiddeld hoger gescoord op zelfbeoordelingen die off-task werden afgenomen dan zelfbeoordelingen die on-task werden afgenomen. Om te zien of deze verschillen ook significant waren, is een onafhankelijke t-toets gedaan. Hierbij is uitgegaan van het aantal participanten in plaats van het aantal studies, zodat een gewogen analyse kon worden gedaan.

Gemiddeld werd er op de zelfbeoordelingen die waren geschreven in de moedertaal van de participanten (M = 0.48, SD = 0.23, N = 1302) hoger gescoord dan op de zelfbeoordelingen die dat niet waren (M = 0.45, SD = 0.13, N = 661). Hierbij was Levenes test significant, dus zijn de vrijheidsgraden aangepast. Het verschil, 0.03, BCa 95% CI [0.01, 0.05], was significant $t(1925) = 3.66$, $p < 0.01$, hoewel het effect zeer klein was, $d = 0.16$.

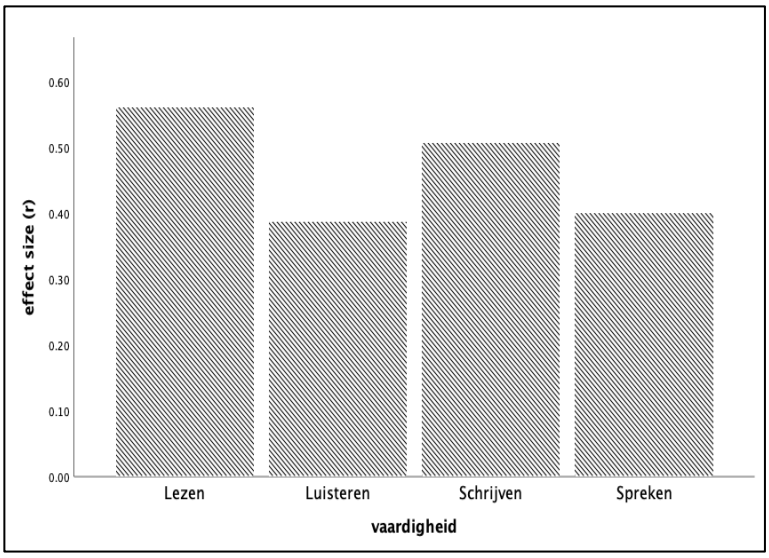
Gemiddeld werd er op de zelfbeoordelingen die als can do statemens (M = 0.56, SD = 0.18, N = 1407) hoger gescoord dan op de zelfbeoordelingen die dat niet waren (M = 0.34, SD = 0.18, N = 699). Hierbij was Levenes test significant, dus zijn de vrijheidsgraden aangepast. Het verschil, 0.22, BCa 95% CI [0.20, 0.23], was significant $t(1372) = 26.3$, $p < 0.01$ en het effect was groot $d = 1.22$.

Gemiddeld werd er op de zelfbeoordelingen met antwoordmogelijkheden op een schaal van minstens drie (M = 0.5, SD = 0.19, N = 1591) hoger gescoord dan op de zelfbeoordelingen die dat niet hadden (M = 0.46, SD = 0.24, N = 537). Hierbij was Levenes test significant, dus zijn de vrijheidsgraden aangepast. Het verschil, 0.04, BCa 95% CI [0.02, 0.6], was significant $t(781) = 3.36$, $p < 0.01$, hoewel het effect zeer klein was, $d = 0.18$.

Op zelfbeoordelingen die on-task waren afgenomen (M = 0.39, SD = 0.2, N = 946) werd gemiddeld juist lager gescoord dan op zelfbeoordelingen die off-task waren afgenomen (M = 0.56, SD = 0.17, N = 1192). Dit verschil, -0.17, BCa 95% CI [-0.19, -0.16, was significant $t(1801) = -21.66$, $p < 0.01$, met een groot effect, $d = 0.92$.

Confounders

De tweede subvraag was naar de invloed van de in het theoretisch kader genoemde variabelen op de accuraatheid van de zelfbeoordelingen. Daarvan worden de verschillende taalvaardigheden geanalyseerd, de eventuele verschillen tussen vreemde- en tweedetaalleerders, en de onderlinge verschillen tussen de taalvaardigheidsniveaus. Onderstaande grafiek geeft het antwoord op de vraag of het vermogen van taalleerders om hun capaciteiten in te schatten verschilt per taalvaardigheid.



Figuur 5: Gemiddelde effect size (r) per vaardigheid

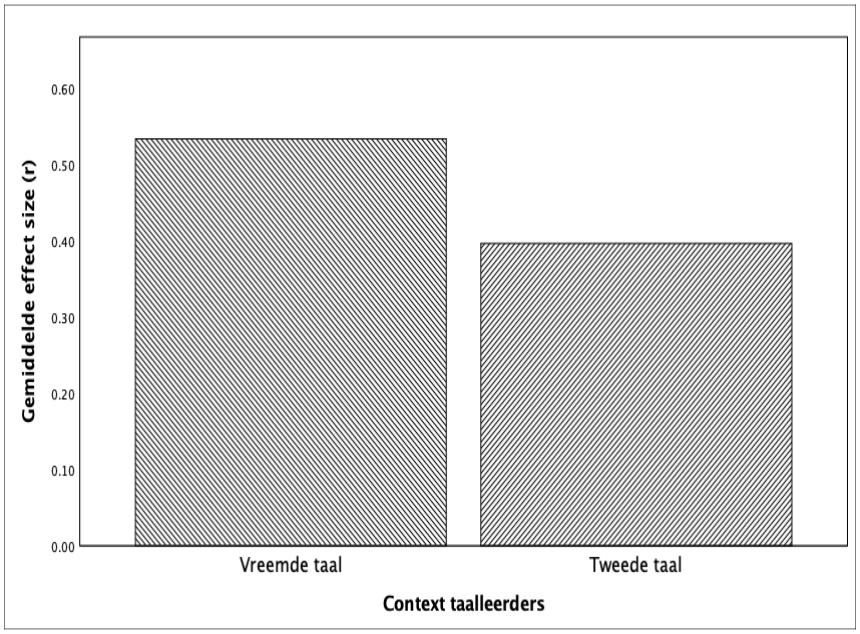
Om erachter te komen of de verschillen tussen de gemiddelde effect sizes van de zelfbeoordelingen onder de vier taalvaardigheden ook significant zijn, is een one-way ANOVA uitgevoerd (zie tabel 2). Daar kwam uit dat de gevonden effect sizes per taalvaardigheid significant van elkaar verschilden ($F (3, 3411) = 253.52$, $p < 0.05$).

Tabel 2: one way ANOVA over het effect van taalvaardigheid op de SA

Analysis of variance	Sum of squares	Df	MS	F
Model	19.5	3	6.5	253.52
Residual	87.47	3411	0.026	
Total	106.97	3314		

Op het gebied van leesvaardigheid wisten de participanten zichzelf het best in te schatten ($M = 0.56$, $SD = 0.19$, $N = 1076$), terwijl luistervaardigheid juist het minst goed ging ($M = 0.37$, $SD = 0.04$, $N = 593$). Schrijfvaardigheid ($M = 0.51$, $SD = 0.06$, $N = 493$) en spreekvaardigheid ($M = 0.4$, $SD = 0.19$, $N = 1253$) stonden respectievelijk op de tweede en derde plek. De Tukey post hoc test gaf aan dat er geen significant verschil was tussen spreekvaardigheid en luistervaardigheid. De overige verschillen tussen de vaardigheden waren wel significant ($p < 0.01$).

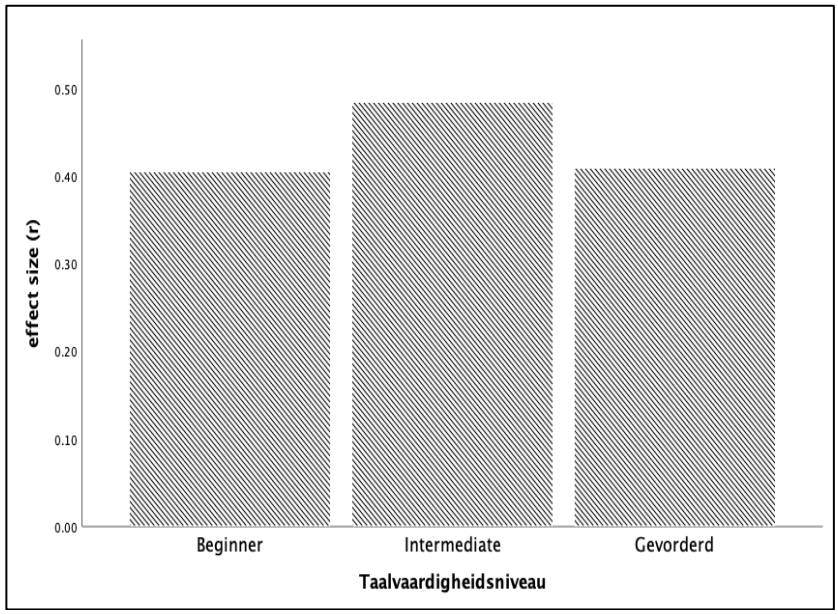
Een tweede variabele in het huidige onderzoek was of het verschil tussen het leren van een vreemde taal en een tweede taal invloed had op de accuraatheid van een zelfbeoordeling. In figuur 6 is te zien hoeveel deze twee van elkaar verschillen. De participanten die een zelfbeoordeling invulden met betrekking tot een vreemde taal waren daarin accurater ($M = 0.53$, $SD = 0.18$, $N = 1255$) dan de participanten die hun tweede taal beoordeelden ($M = 0.4$, $SD = 0.17$, $N = 684$).



Figuur 6: Gemiddelde effect size (r) tussen vreemdetaalleerders en tweedetaalleerders

Een onafhankelijke t-test is uitgevoerd om te beoordelen of het verschil tussen deze twee significant is. Het verschil tussen de participanten die hun vreemde taal beoordeelden ten opzichte van een tweede taal, 0.13, BCa 95% CI [0.12, 0.15], was significant $t(1937) = 16.45$, $p < 0.01$, met een middelgroot effect, $d = 0.74$.

De derde variabele die in de huidige meta-analyse wordt onderzocht, is het effect dat het taalvaardigheidsniveau van de participant heeft op de accuraatheid van de SA. Figuur 7 geeft de gemiddelde effectgroottes weer.



Figuur 7: Gemiddelde effect size (r) per taalvaardigheidsniveau

Er is een one way ANOVA uitgevoerd om de verschillen tussen deze drie te beoordelen (zie tabel 3). Daarbij is een significant verschil gevonden tussen de drie taalvaardigheidsniveaus ($F(2, 1170) = 15.3$, $p < 0.05$).

Tabel 3: one way ANOVA over het effect van het taalvaardigheidsniveau op de SA

Analysis of variance	Sum of squares	Df	MS	F
Model	1.401	2	0.701	15.3
Residual	53.563	1170	0.046	
Total	54.965	1172		

De Tukey post hoc test gaf aan dat er een significant verschil was tussen intermediate (M = 0.48, SD = 0.13, N = 355) ten opzichte van beginner (M = 0.4, SD = 0.21, N = 99) en tussen intermediate ten opzichte van gevorderd (M = 0.41, SD = 0.25, N = 719), maar niet tussen beginner en gevorderd (p = 0.4).

Discussie

Met deze meta-analyse is geprobeerd een antwoord te vinden op de vraag: 'In hoeverre zijn zelfbeoordelingen binnen vreemde- en tweedetaalverwerving een valide methode om taalvaardigheden te meten?' De 25 experimenten die in deze analyse waren opgenomen vergeleken een zelfbeoordeling over de taalvaardigheid van kandidaten met een taalvaardigheidsmeting. De gemiddelde correlatie die daaruit kwam was 0.52, wat betekent dat de SA's en vaardigheidsmetingen in deze onderzoeken gemiddeld 27% met elkaar overlappen. Deze correlatiecoëfficiënt is bijna twee keer zo hoog als de effectgrootte die gevonden was in een onderzoek naar zelfbeoordelingen in het algemeen (r = 0.29) (Mabe & West, 1982). Het lijkt er dus op dat mensen relatief gezien goed zijn in het beoordelen van hun taalvaardigheid.

Hoe hoog de correlatiecoëfficiënt moet zijn, ligt aan het doel van de betreffende toets. De stakes kunnen laag zijn, zoals bij plaatsingstoetsen, of hoog, wanneer er een certificaat aan de uitslag verbonden is. Wanneer de stakes hoog zijn, wordt er ook een hogere validiteit van de toets verlangd. Er zijn weinig specifieke standaarden wat betreft de minimale waarde van de correlatiecoëfficiënt (Crocker & Angina, 2008). Over het algemeen bekeken, lijkt een correlatiecoëfficiënt van 0.8 aangehouden te worden voor high stakes toetsing, en 0.6 voor low stakes toetsing. Dit zou betekenen dat een gemiddelde effectgrootte van 0.52 voor zelfbeoordelingen te laag is om in afzonderlijk ingezet te worden als vaardigheidsmeting.

Uit de homogeniteitsanalyse bleek dat de resultaten uit de opgenomen experimenten heterogeen waren. Dit kan worden verklaard door de verschillende variabelen die in het theoretisch kader zijn genoemd. Deze variabelen waren als volgt: de hoogte van de stakes van een SA, de kwaliteit van de verschillende zelfbeoordelingen, kandidaateigenschappen als leeftijd, geslacht, achtergrond, opleidingsniveau en taalvaardigheidsniveau, de gemeten taalvaardigheid, de talige context waarin de kandidaten zich bevinden en hun voorgaande ervaring met SA. In de huidige analyse zijn de kwaliteitskenmerken, de taalvaardigheden, het verschil tussen vreemde- en tweedetaalverwerving, en het taalvaardigheidsniveau van de participanten meegenomen.

Er zijn in het theoretisch kader vier kwaliteitskenmerken opgesteld die een positieve invloed zouden kunnen hebben op de validiteit van een SA. Deze waren de volgende:

1. De toets is geschreven in de moedertaal van de participant;
2. De zelfbeoordeling bestaat uit can do statements;
3. De antwoordmogelijkheden bevinden zich op een schaal van minstens drie;
4. De zelfbeoordeling is on-task afgenomen.

Het eerste kenmerk (de moedertaal) had een positief, significant effect op de accuraatheid van de zelfbeoordeling, maar het effect was erg klein. De reden voor het includeren van dit kwaliteitskenmerk was dat de participanten een SA beter zouden begrijpen wanneer het in de moedertaal was geschreven en ze in de moedertaal beter over hun eigen taalvaardigheden na kunnen denken (Hasselgreen, 2003). Een verklaring voor het kleine effect zou kunnen zijn dat veruit de meeste participanten uit de opgenomen experimenten een intermediate of gevorderd taalvaardigheidsniveau bezaten. Dat zou kunnen betekenen dat de participanten de korte can do statements ook in de tweede of vreemde taal goed konden begrijpen.

Uit de meta-analyse is gebleken dat het tweede kenmerk uit deze lijst (de can do statements) de grootste positieve invloed uitoefende op de validiteit. Het verschil tussen zelfbeoordelingen met en zonder can do statements was significant en het effect was groot. Can do statements gaan namelijk over specifieke taalsituaties en daardoor kunnen participanten preciezer inschatten wat ze wel en niet kunnen (Oskarsson, 1980; Bradshaw, 2001; Little, 2002).

Het derde kwaliteitskenmerk (de grootte van de schaal van antwoordmogelijkheden) had ook een significant en positief effect op de validiteit, maar ook hier was het effect klein. Dit zou kunnen liggen aan het feit dat het verschil tussen twee en drie antwoordmogelijkheden niet heel groot is. Volgens Oskarsson (1980) moest er minimaal een neutrale optie bestaan om een preciezere indruk te krijgen van een zelfbeoordeling. Zelf raadde hij echter vijf antwoordmogelijkheden aan. Wellicht had de huidige analyse een groter effect laten zien wanneer de cut-off hoger had gelegen.

Het vierde kwaliteitskenmerk (de zelfbeoordeling is on-task afgenomen) had juist een negatieve invloed. Dit is een opvallend resultaat dat rechtstreeks tegen de hypothese in gaat. De verwachting was namelijk dat on-task zelfbeoordelingen de kandidaten meer informatie verschaffen om een oordeel te vellen, waardoor ze accurater worden gemaakt. Todd (2002) beargumenteert daarnaast dat kandidaten beter in staat zouden zijn om hun vaardigheden accuraat in te schatten wanneer ze die kort geleden nog hadden uitgevoerd. In het onderzoek van Butler & Lee (2006) werd bijvoorbeeld wel een positief, significant effect gevonden onder kinderen van 8-11 jaar oud die een off-task en on-task zelfbeoordeling invulden. In het onderzoek van Laufer en Yano (2001) - waarin de participanten zowel een off-task als een on-task SA maakten - werd geen significant verschil gevonden, hoewel de participanten iets accurater scoorden op de on-task SA. Een mogelijke verklaring voor de resultaten uit het huidige onderzoek zou in het coderen kunnen liggen. In de meeste onderzoeken werd

namelijk niet expliciet vermeld of de SA off-task of on-task was afgenomen. Daarom is dit afgeleid aan de volgorde waarin de toetsen werden afgenomen: wanneer de vaardigheidsmeting voor de zelfbeoordeling kwam, is dit vaak opgevat als on-task toetsing. Het is mogelijk dat er daardoor verkeerde interpretaties zijn gedaan die tot deze verwarrende resultaten leiden. Het zou ook kunnen dat de participanten na een vaardigheidsmeting al moe zijn voordat ze de zelfbeoordeling kunnen maken. Een dergelijk effect zou de negatieve uitkomst kunnen verklaren.

Na de kwaliteitskenmerken zijn enkele mogelijke confounders onderzocht. Ten eerste zijn de opgenomen onderzoeken opgesplitst per onderzochte taalvaardigheid (lezen, luisteren, schrijven, en spreken). Uit de resultaten volgt dat de participanten hun leesvaardigheid het meest accuraat konden beoordelen. Dat lijkt erop te wijzen dat de receptieve vaardigheden accurater kunnen worden ingeschat (zoals ook bleek uit de meta-analyse van Ross, 1998), ware het niet dat de participanten op luistervaardigheid juist het minst goed scoorden. Op spreekvaardigheid werd ook onder gemiddeld beoordeeld. Participanten zijn dus beter in het beoordelen van hun taalvaardigheden wanneer die op schrift staan (lezen en schrijven) en minder goed wanneer het gaat om het voeren van gesprekken (luisteren en spreken). Een verklaring hiervoor kan zijn dat men doorgaans meer tijd heeft om iets te lezen of te schrijven, en daarbij ook zinnen opnieuw kan lezen. Deze mogelijkheid is er minder bij luisteren en spreken. Daardoor is er bij de schriftelijke vaardigheden meer ruimte om de vaardigheden te beoordelen.

De experimenten zijn tevens opgesplitst in de leerders van een vreemde taal en die van een tweede taal. In het theoretisch kader was de talige context (komt de taalinput vanuit moedertaalsprekers of niet-moedertaalsprekers) als mogelijke confounder genoemd, omdat een vreemdetaalleerder haar taalvaardigheid waarschijnlijk hoger zal inschatten dan een tweedetaalleerder. Uit de analyse bleek dat de vreemdetaalleerders hun eigen taalvaardigheden accurater beoordeelden dan de tweedetaalleerders. Toevallig bleek ook dat vrijwel alle onderzoeken waarin een vreemde taal werd onderzocht waren uitgevoerd in Azië (zie bijlage 1). De onderzoeken waarin een tweede taal werd onderzocht waren voornamelijk in de Verenigde Staten en een aantal Europese landen uitgevoerd. Aangezien er niet voldoende vreemdetaal-onderzoeken uit niet-Aziatische landen zijn gevonden, is het in deze meta-analyse niet mogelijk om de talige context en de achtergrond van de participanten van elkaar los te koppelen. Het gevonden effect kan door één van deze twee mogelijke confounders zijn veroorzaakt of door de combinatie ervan.

Deze meta-analyse heeft ook gezocht naar het mogelijke effect van het taalvaardigheidsniveau op de zelfbeoordeling. De onderzoeken zijn hierbij in drie niveaus opgedeeld: beginner, intermediate, en gevorderd. Van deze drie scoorden de participanten met een intermediate taalvaardigheidsniveau significant accurater dan de beginners en de gevorderden. Het is mogelijk dat hierin het Dunning-Kruger effect (Kruger & Dunning, 1999) te zien is. Dit houdt in dat de beginnende taalleerders hun eigen niveau overschatten, en daardoor niet accuraat zijn, en gevorderde taalleerders zichzelf juist onderschatten, en daardoor niet accuraat zijn.

Dat betekent dat de participanten met een intermediate taalvaardigheidsniveau het meest accuraat zouden scoren, en deze theorie wordt bevestigd door de resultaten.

In het theoretisch kader zijn nog meer mogelijke confounders genoemd, maar de opgenomen experimenten lieten niet alle analyses toe. In niet een van de onderzoeken is gebruik gemaakt van een high stakes toets, dus het verschil tussen low en high stakes kon niet worden geanalyseerd. Ook de leeftijd en het opleidingsniveau zijn om een dergelijke reden buiten beschouwing gelaten; de participanten uit de verschillende experimenten waren overwegend universitaire studenten van rond de 20 jaar oud. Het geslacht van de participanten is meestal niet meegenomen in de onderzoeken (in het onderzoek waar het verschil wel is gemeten, is geen significant verschil gevonden; Laufer & Yano, 2001). De culturele achtergrond van de participanten kon niet worden geanalyseerd omdat daar te weinig data over beschikbaar was; alleen het land van onderzoek werd gerapporteerd. Tenslotte is ook over de eerdere ervaring van de participanten met SA weinig bekend.

De gevonden effect size ($r = 0.52$) komt overeen met de schatting van Blanche en Merino (1989), die op basis van zeven correlatiecoëfficiënten aangaven dat de meeste zich tussen 0.5 en 0.6 bevonden. Het huidige onderzoek bevestigt daarmee hun vermoeden. Blanche en Merino (1989) rapporteerden daarnaast dat veel van hun opgenomen onderzoeken een effect zagen van het taalvaardigheidsniveau van de participanten op de accuraatheid van zelfbeoordelingen. De sterke taalleerders zouden zichzelf vaak onderschatten, terwijl zwakkere taalleerders zichzelf juist overschatten. Ook hier heeft het huidige onderzoek aanvullend bewijs voor gevonden. De intermediate participanten scoorden namelijk het meest accuraat in vergelijking met de beginnende en gevorderde taalleerders.

De meta-analyse van Ross (1998) rapporteerde een correlatiecoëfficiënt van 0.63. Het verschil met het huidige onderzoek is redelijk groot. Dit zou verklaard kunnen worden doordat het gemiddelde door Ross (1998) niet gewogen is, waardoor alle opgenomen onderzoeken even zwaar meewegen. In het huidige onderzoek zijn de effect sizes gewogen aan de hand van het aantal participanten. De range van de gevonden effect sizes komt wel redelijk overeen: bij Ross (1998) liep die van 0.09 tot 0.8, en in het huidige onderzoek van 0.12 tot 0.8. Ook is bij beide onderzoeken een homogeniteitsanalyse gedaan die wees op een heterogene verzameling van correlatiecoëfficiënten. Ross (1998) wijt deze heterogeniteit aan de verschillen tussen de taalvaardigheden. Zijn bevindingen waren dat participanten accurater scoren op de receptieve vaardigheden dan op de productieve vaardigheden, terwijl in het huidige onderzoek een verschil is gevonden tussen schriftelijke en gesproken vaardigheden. Dit wordt veroorzaakt door het grote verschil in de luistervaardigheid: Ross (1998) rapporteert een effect size van 0.65, het huidige onderzoek rapporteert 0.37. De mogelijke verklaring hiervoor is methodisch van aard. In het onderzoek van Ross (1998) is geen selectie gedaan wat betreft effect sizes, waardoor participanten vaak dubbel zijn meegerekend. In het huidige onderzoek zijn maar vier experimenten gevonden waarin de luistervaardigheid werd beoordeeld, waardoor ook dit gemiddelde niet ontegenzeglijk waar is. Naast de verschillende taalvaardigheden zijn in de meta-analyse van Ross (1998) geen andere mogelijke veroorzakers van de heterogeniteit onderzocht.

Conclusie

Alhoewel mensen redelijk accuraat zijn in het inschatten van hun taalvaardigheden, zeker ten opzichte van zelfbeoordelingen binnen andere gebieden, lijkt het niet mogelijk om SA als een op zichzelf staande toets in te zetten. De reeds gevalideerde taalvaardigheidsmetingen en docentoordelen overlappen maar 27% met de zelfbeoordelingen. Dat betekent dat een SA waarschijnlijk niet accuraat genoeg is om in te zetten ter vervanging van een vaardigheidsmeting, zelfs niet als een low stakes toets.

Op basis van de informatie die we nu hebben over confounders kunnen we ons zelfs afvragen of het wel ethisch verantwoord is om SA als vaardigheidsmeting in te zetten. Er zijn zoveel kandidaateigenschappen die een positieve dan wel negatieve invloed kunnen uitoefenen op de uiteindelijke score (waaronder niveau en culturele achtergrond), dat het niet eerlijk is om een zelfbeoordeling summatief in te zetten. De beoordeling van taalvaardigheid zou niet in een dergelijke mate afhankelijk moeten zijn van het karakter van een kandidaat.

De kwaliteit van zelfbeoordelingen kan wel op verschillende manieren worden verbeterd. In deze meta-analyse zijn vier kwaliteitskenmerken geformuleerd en onderzocht. Van deze vier leek het kwaliteitskenmerk ‘de toetsitems zijn verwoord als can do statements’ het meest positieve effect te hebben op de accuraatheid van een SA. De zelfbeoordelingen die on-task waren afgenomen scoorden juist, tegen de verwachtingen in, lager dan de zelfbeoordelingen die off-task waren afgenomen.

Vervolgonderzoek zou gedaan kunnen worden naar hoe we SA zo accuraat mogelijk kunnen maken. In het huidige onderzoek zijn een aantal kwaliteitskenmerken geformuleerd naar aanleiding van een artikel uit de jaren ’80 (Oskarsson, 1980), maar daar zouden specifiek experimenten voor opgezet kunnen worden. Vooral over het verschil tussen on-task en off-task toetsen zijn in deze meta-analyse vragen ontstaan. Een voorstel zou zijn om hier een experiment voor op te zetten waarin een experimentele groep een non-linguïstische taak doet alvorens een SA te maken, en een controlegroep enkel de SA invult. Met een dergelijk experiment kan een eventueel vermoeidheidseffect op het maken van een SA worden onderzocht.

Ook is het interessant om de vastgestelde confounders verder te analyseren. Vooral de talige context en de culturele achtergrond van een kandidaat zijn weinig onderzocht met betrekking tot SA bij vreemde- en tweedetaalleerders. Ook in de huidige meta-analyse kon niet worden vastgesteld wat het effect van deze twee variabelen is op de accuraatheid van een zelfbeoordeling. Een antwoord op de vraag naar deze eventuele effecten zou ons veel kunnen leren over de verschillende taalleerders. De doelgroep bij vreemde- en tweedetaalverwerving is tenslotte divers en deze informatie zou meer begrip kunnen brengen in taallessen waarbij cursisten uit meerdere culturen komen.

Literatuurlijst

5 grandioze geheimen om elke meerkeuzetoets te halen! Radboud Universiteit: Radboud in'to Languages, 2020. Verkregen uit: https://www.ru.nl/radboudintolanguages/over-ons/onze-blogs/5-grandioze-geheimen-elke-meerkeuzetoets-halen/

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. In: *Language Testing, 22*(3), p. 301-320.

Andringa, S. (2015). Toetsing en evaluatie. In B. Bossers, F. Kuiken, & A. Vermeer (Eds.), *Handboek Nederlands als tweede taal in het volwassenenonderwijs* (pp. 359-393). Bussum: Coutinho.

Babaii, E., Taghaddomi, S., & Pashmforoosh, R. (2016). Speaking Self-Assessment: Mismatches between Learners' and Teachers' Criteria. In: *Language Testing 33* (3), p. 411-437.

Bachman, L., & Palmer, A. (1981). The construct validity of the FSI Oral Proficiency Interview. In: *Language Learning 31*, p. 67–86.

Baleghizadeh, S., & Hajizadeh, T. (2014). Self- and Teacher-Assessment in an EFL Writing Class. In: *GIST Education and Learning Research Journal*, p. 99-117.

Barnett, J. E., & Hixon, J. E. (1997). Effects of Grade Level and Subject on Student Test Score Predictions. In: *The Journal of Educational Research 90*(3), p. 170-174.

Barrow, J., Nakanishi, Y., & Ishino, H. (1999). Assessing Japanese College Students' Vocabulary Knowledge with a Self-Checking Familiarity Survey. In: *System 27*(2), p. 223-247.

Blanche, P., & Merino, B. J. (1989). Self-Assessment of Foreign-Language Skills: Implications of Teachers and Researchers. In: *Language Learning 39*(3), p. 313-340.

Blatchford, P. (1997). Students' self assessment of academic attainment: Accuracy and stability from 7 to 16 years and influence of domain and social comparison group. In: *Educational Psychology 17*, p. 345–359.

Blue, G. M. (1994). Self-Assessment of Foreign Language Skills: Does It Work? In: *CLE Working Papers 3*, p. 18-35.

Bradshaw, B. K. (2001). Do students effectively monitor their comprehension? In: *Reading Horizons 41*, p. 143–154.

Brantmeier, C. (2006). Advanced L2 Learners and Reading Placement: Self-Assessment, CBT, and Subsequent Performance. In: *An International Journal of Educational Technology and Applied Linguistics 34* (1), p. 15-35.

Bullock, D. (2011). Learner Self-Assessment: An Investigation into Teachers' Beliefs. In: *ELT Journal 65*(2), p. 114-125.

Butler, R. (1990). The effects of mastery and competitive conditions on self-assessment at different ages. In: *Child Development 61*, p. 201–210.

Butler, Y. G. (2018). The Role of Context in Young Learners' Processes for Responding to Self-Assessment Items. In: *The Modern Language Journal 102* (1), p. 242-261.

Butler, Y. G., & Lee, J. (2006). On-Task Versus Off-Task Self-Assessments Among Korean Elementary School Students Studying English. In: *The Modern Language Journal 90*(4), p. 506-518.

Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. In: *Language Testing 27* (1), p. 5-31.

Carter, T.J., & Dunning, D. (2008). Faulty self-assessment: Why evaluating one's own competence is an intrinsically difficult task. In: *Social and Personality Psychology Compass 2*(1), p. 346–360.

Chen, P., & Zimmerman, B. (2007). A Cross-National Comparison Study on the Accuracy of Self-Efficacy Beliefs of Middle-School Mathematics Students. In: *The Journal of Experimental Education 75*(3), p. 221-244.

Chen, Y. M. (2008). Learning to Self-Assess Oral Performance in English: A Longitudinal Case Study. In: *Language Teaching Research 12*(2), p. 235-262.

Claes, M., & Salame, R. (1975). La motivation à l'accomplissement et l'auto-évaluation des performances en relation avec le rendement scolaire [Motivation toward accomplishment and the self-evaluation of performances in relation to school achievement]. In: *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement, 7*(4), p. 397–410.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, U.K: Press Syndicate of the University of Cambridge.

Denies, K., & Janssen, R. (2016). Country and Gender Differences in the Functioning of CEFR-Based Can-Do Statements as a Tool for SelfAssessing English Proficiency. In: *Language Assessment Quarterly 13*(3), p. 251-276.

Djiwandono, P. I. (2017). Character Education in Content Courses: Self-Scoring as a Means for Developing Honesty in Students. In: *Teflin Journal: A Publication on the Teaching and Learning of English 27* (2), p. 153-165.

Dolosic, H. (2018). An examination of self-assessment and interconnected facets of second language reading. In: *Reading in a Foreign Language 30*(2), p. 189-208.

Dolosic, H., Brantmeier, C., Strube, M., & Hogrebe, M. C. (2016). Living Language: Self-Assessment, Oral Production, and Domestic Immersion. In: *Foreign Language Annals 49*(2), p. 302-316.

Dunning, D., Heath, C., & Suls, J. (2004). Flawed Self-Assessment Implications for Health, Education, and the Workplace. In: *Psychological Science in the Public Interest 5*(3), p. 69-106.

Ferguson, N. (1978). Self-assessment of listening comprehension. In: *International Review of Applied Linguistics 16*, p. 149–156.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. In: *British Journal of Mathematical & Statistical Psychology 63*(3), p. 665-694.

Grahn-Saarinen, A. (2003). Chapter 7: Self-assessment, reflection and 'Can Do' statements. In A. Hasselgreen, *Bergen 'Can Do' project* (pp. 57-61). Verkregen op 19 juli 2020 van Council of Europe Publishing. Website: http://archive.ecml.at/documents/pub221E2003_Hasselgreen.pdf

Hasselgreen, A. (2003). *Bergen 'Can Do' project.* Verkregen op 19 juli 2020 van Council of Europe Publishing. Website: http://archive.ecml.at/documents/pub221E2003_Hasselgreen.pdf

Hewitt, M. P. (2005). Self-evaluation accuracy among high school and middle school instrumentalists. In: *Journal of Research in Music Education 53*, p. 148–161.

Huang, C. (2013). Gender differences in academic self-efficacy: A meta-analysis. In: *European Journal of Psychology of Education 28*(1), p. 1–35.

Hughes, A. (2002). *Testing for Language Teachers* (2 ed.). Cambridge: Cambridge University Press.

Jamrus, M. H. M., & Razali, A.B. (2019). Using self-assessment as a tool for English Language Learning. In: *English Language Teaching 12* (11), p. 64-73.

Janssen-van Dieten, A. M. (1989). The development of a test of Dutch as a second language: the validity of self-assessment by inexperienced subjects. In: *Language Testing 6*, p. 1–13.

Klenowski, V. (1995). Student self-evaluation processes in student-centred teaching and learning contexts of Australia and England. In: *Assessment in Education 2*(2), p. 145-163.

Kohonen, V. (2000). Student reflection in portfolio assessment: making language learning more visible. In: *Babylonia 1,* p. 13-16.

Kruger, J., & Dunning, D. (1999). Unskilled or unaware of it: Difficulties in recognizing one's own incompetence lead to inflated self-assessments. In: *Journal of Personality and Social Psychology 77*, p. 1121–1134.

Lappin-Fortin, K., & Rye, B. (2014). The Use of Pre-/Posttest and Self-Assessment Tools in a French Pronunciation Course. In: *Foreign Language Annal 47*(2), p. 300-320.

Laufer, B., & Yano, Y. (2001). Understanding Unfamiliar Words in a Text: Do L2 Learners Understand How Much They Don't Understand? In: *Reading in a Foreign Language 13*(2), p. 549-566.

Laveault, D., & Miles, C. (2002). The study of individual differences in the utility and validity of rubrics in the learning of writing ability. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Leblanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. In: *TESOL Quarterly 19*(4), p. 673-687

Lipsey, M. W., & Wilson, D. B. (2001). *Applied social research methods series; Vol. 49. Practical meta-analysis.* Sage Publications, Inc.

Little, D. (2002). The European Language Portfolio: structure, origins, implementation and challenges. In: *Language Teaching 35*(3), p.182–189.

Little, D. (2005). The Common European Framework and the European Language Portfolio: involving learners and their judgements in the assessment process. In: *Language Testing 22*(3), p. 321-336.

Liu, H., & Brantmeier, C. (2019). "I know English": Self-assessment of foreign language reading and writing abilities among young Chinese learners of English. In: *System 80*, p. 60-72.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. In: *Journal of Applied Psychology 67*, p. 280–296

Mistar, J. (2011). A Study of the Validity and Reliability of Self-Assessment. In: *TEFLIN Journal: A publication on the teaching and learning of English 22*(1), p. 45-58.

Nejad, A.M. & Mahfoodh, O. M. A. (2019). Assessment of Oral Presentations: Effectiveness of Self-, Peer-, and Teacher Assessments. In: I*nternational Journal of Instruction.12*(3), p. 615-632.

Oskarsson, M. (1980). *Approaches to self-assessment in foreign language learning.* Oxford: Pergamon, for the Council of Europe.

Oswald, F., & Plonsky, L. (2010). Meta-analysis in Second Language Research: Choices and Challenges. In: *Annual Review of Applied Linguistics 30*, p. 85-110.

Paris S. G., & Paris A. H. (2001). Classroom applications of research on self-regulated learning. In: *Educational Psychology 36(*2), p. 89–101.

Van Rhee, H.J., Suurmond, R., & Hak, T. (2015). *User manual for Meta-Essentials: Workbooks for meta-analysis (Version 1.4)* Rotterdam, The Netherlands: Erasmus Research Institute of Management. Verkregen op 10-06-20 via www.erim.eur.nl/research-support/meta-essentials

Rosenthal, R., & DiMatteo, M. R. (2001). Meta analysis: Recent developments in quantitative methods for literature reviews. In: *Annual Review of Psychology 52,* p. 59–82.

Ross, J. A. (2006). The Reliability, Validity, and Utility of Self-Assessment. In: *Practical Assessment, Research & Evaluation 11*(10), p. 1-13.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. In: *Language Testing 15*(1), p. 1-20.

Runnels, J. (2016). Self-assessment accuracy: Correlations between Japanese English Learners' Self-Assessment on the CEFR – Japan's Can Do Statements and Scores on the TOEIC. In: *Taiwan Journal of TESOL 13*(1), p. 105-137.

Salehi, M., & Masoule, Z. S. (2017). An investigation of the reliability and validity of peer-, self-, and teacher assessment. In: *Southern African Linguistics and Applied Language Studies 35*(1), p.1-15.

Sirikanjanawong, N., & Wasanasomsithi, P. (2018). Relationship between the ICAO Language Proficiency Requirements (LPRs) and Test of English for International Communication (TOEIC) Scores of Flight Attendants in Thailand. In: *LEARN Journal: Language Education and Acquisition Research Network 11(1),* p. 64-86.

Suurmond R, van Rhee, H, Hak T. (2017). Introduction, comparison and validation of Meta-Essentials: A free and simple tool for meta-analysis. In: *Research Synthesis Methods* 8 (4), p. 537-553.

Suzuki, Y. (2015). Self-Assessment of Japanese as a Second Language: The Role of Experiences in the Naturalistic Acquisition. In: *Language Testing 32* (1), p. 63-81.

Todd, R. W. (2002). Using Self-Assessment for Evaluation. In: *Forum 40* (1), p. 16-19.

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Invesitating self- and other-perception of second language speech. In: *Bilingualism 19*(1), p. 122-140.

Unaldi, I. (2016). Self and Teacher Assessment as Predictors of Proficiency Levels of Turkish EFL Learners. In: *Assessment and Evaluation in Higher Education 41*(1), p. 67-80.

Wenyue, M. A., & Winke, P. (2019). Self-assessment: how reliable is it in assessing oral proficiency over time? In: *Foreign Language Annals 52* (1), p. 66-86.

Whang, P. A., & Hancock, G. R. (1994). Motivation and mathematics achievement: Comparison between Asian American and nonAsian students. In: *Contemporary Educational Psychology 19*, p. 302–322.

What is the ELP? Council of Europe: European Language Portfolio, 2020. Verkregen op 10-07-20 van: https://www.coe.int/en/web/portfolio/introduction

Yamini, M., & Tahmasebi, S. (2012). Self-Assessment and Peer-Assessment in an EFL Context. In: *Advances in Language and Literary Studies 3*(1), p. 49-58.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekarts, P. Pintrich, & M. Zeidner (Eds.), *Self-regulation: Theory, research, and applications* (pp. 13–39). Orlando, FL: Academic.

Bijlagen

Bijlage 1: opgenomen studies in de meta-analyse

#	*Studie*	*N*	*r*	*T2*	*Vaardigheid*	*Taalniveau*	*Vreemde of tweede taal*	*Land waarin het onderzoek is uitgevoerd*
1	Babaii, Taghaddomi & Pashmforoosh (2016)	29	0.73	Engels	Spreken	Intermediate	Vreemde	Iran
2	Bachman & Palmer (1981)	75	0.57	Engels	Spreken en lezen	Intermediate	Tweede	Verenigde Staten
3	Baleghizadeh & Hajizadeh (2014)	15	0.63	Engels	Schrijven	Intermediate	Vreemde	Iran
4	Barrow, Nakanishi, & Ishino (1999)	279	0.73	Engels	Lezen	Niet genoemd	Vreemde	Japan
5	Brantmeier (2006)	71	0.87	Spaans	Lezen	Gevorderd	Niet genoemd	Verenigde Staten
6	Butler & Lee (2006)	25	0.75	Engels	Schrijven	Beginner	Vreemde	Iran
7	Chen (2008)	28	0.55	Engels	Spreken	Gevorderd	Vreemde	Taiwan
8	Dolosic (2018)	77	0.26	Engels	Lezen	Gevorderd	Vreemde	China
9	Dolosic, Brantmeier, Strube & Hogrebe (2016)	22	0.29	Frans	Spreken	Intermediate	Vreemde	Verenigde Staten
10	Ferguson (1978)	89	0.39	Engels	Spreken	Niet genoemd	Vreemde	Zwitserland
11	Janssen-van Dieten (1989)	134	0.36	Nederlands	Lezen, luisteren, spreken, schrijven	Intermediate	Tweede	Nederland
12	Lappin-Fortin & Rye (2014)	48	0.53	Frans	Spreken	Intermediate	Niet genoemd	Frankrijk
13	Laufer & Yano (2001)	106	0.55	Engels	Lezen	Gevorderd	Vreemde	Israel, Japan, China
14	Leblanc & Painchaud (1985)	200	0.53	Frans en Engels	Lezen, luisteren, spreken, schrijven	Niet genoemd	Tweede	Verenigde Staten
15	Liu & Brantmeier (2019)	10	0.43	Engels	Lezen en schrijven	Beginner	Vreemde	China
16	Mistar (2011)	78	0.54	Engels	Spreken en schrijven	Niet genoemd	Tweede	Verenigde Staten
17	Nejad & Mahfoodh (2019)	60	0.8	Engels	Spreken	Gevorderd	Vreemde	Iran

18	Runnels (2016)	80	0.25	Engels	Luisteren en lezen	Gevorderd	Vreemde	Japan
19	Salehi & Masoule (2017)	32	0.56	Engels	Schrijven	Intermediate - gevorderd	Vreemde	Iran
20	Sirikanjanawong & Wasanasomsithi (2018)	100	0.38	Engels	Spreken en schrijven	Gevorderd	Vreemde	Thailand
21	Suzuki (2015)	63	0.251	Japans	Spreken	Gevorderd	tweede	Japan
22	Trofimovich, Isaacs, Kennedy, Saito & Crowther (2016)	134	0.12	Engels	Spreken	Gevorderd	tweede	Canada
23	Unaldi (2016)	239	0.57	Engels	Lezen en luisteren	Niet genoemd	Vreemde	Turkije
24	Wenyue & Winke (2019)	80	0.125	Chinees	Spreken	Niet genoemd	Niet genoemd	Verenigde Staten
25	Yamini & Tahmasebi (2012)	64	0.264	Engels	Spreken	Beginner	Vreemde	Iran