Amsterdam University Press

NEDERLANDSE TAALKUNDE 26.2 (2021) 195-221
HTTPS://DOI.ORG/10.5117/NEDTAA2021.2.003.DEYG

Begrijpelijkheid en accent bij beginnende NT2-leerders

Het belang van achtergrondvariabelen

Bart Deygers
Universiteit Gent
Bart.Deygers@ugent.be

Carolien Frijns
Universiteit Gent
Carolien.Frijns@ugent.be

Abstract

In social interaction, foreign language accent and comprehensibility impact how we perceive our conversational partners. In recent years, research interest in these constructs has been on the rise, while many issues remain underexposed. These issues include the relationship between comprehension and accent on the one hand, and background variables of both learner and assessor on the other. Since most research to date has been conducted with highly educated and advanced learners of English as a second or foreign language, we do not know to what extent those results can be generalised to a wider population that includes beginning learners of Dutch as a second language from various educational backgrounds. In addition, little research has been done into the comparability of the judgements of trained and non-trained assessors. In the current study, we compared the judgments of four trained evaluators with the intuitive judgments of 272 non-trained evaluators (first-year students at various Educational Bachelor's Programmes in Primary Education). The first group of raters evaluated the speaking performance of 116 learners of Dutch as a second language using standardised criteria, the second group used more subjective criteria. The results show that the two groups of evaluators make very similar judgements and that these judgements are mainly related to two background variables: the nature of the NT2 course the learners followed

© Bart Deygers & Carolien Frijns

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

(also an indicator of prior education and cognitive ability) and the level of language proficiency of the course.

Keywords: L2 accent, comprehensibility, rater behavior, intuitive judgment, impact of background variables

1. Sociolinguïstische perspectieven op anderstalig accent en begrijpelijkheid

In deze studie bekijken we hoe getrainde en niet-getrainde beoordelaars het accent en de begrijpelijkheid van beginnende leerders van het Nederlands als tweede taal (NT2) percipiëren en op welke wijze die percepties samenhangen met zowel achtergrondvariabelen van de NT2-leerder als die van de beoordelaars. In wat volgt, bespreken we (1) de constructen *anderstalig accent* en *begrijpelijkheid* en (2) de relatie tussen uitspraak en sociale verbondenheid of affiliatie, een concept dat in recent tweedetaalverwervingsonderzoek steeds meer aan belang wint en boeiende perspectieven op uitspraakverwerving biedt.

1.1 Anderstalig accent en begrijpelijkheid

De voorbije jaren is er heel wat onderzoek verschenen naar anderstalig accent, en de relatie tussen accent, verstaanbaarheid en begrijpelijkheid (zie bijvoorbeeld Crowther 2020; Suzuki & Kormos 2020). Onderzoek van Munro & Derwing (1995a, 1995b) toont aan dat het om drie afzonderlijke maar sterk verbonden constructen gaat. Hoewel er van 'anderstalig accent' verschillende definities in omloop zijn, definiëren we dat construct in navolging van Munro (2008) als spraak die op verschillende, niet-pathologische manieren afwijkt van die van moedertaalsprekers. Het gaat met andere woorden om het gepercipieerde verschil tussen een anderstalig accent en het accent van een moedertaalspreker. Hoewel er ook onder moedertaalsprekers soms grote uitspraakverschillen bestaan, worden die accentvariaties doorgaans niet als 'anderstalig' gepercipieerd.

Wanneer een accent van een niet-moedertaalspreker sterk aansluit bij dat van een bepaalde subpopulatie uit de groep moedertaalsprekers, spreken we van *nativeness* (Derwing & Munro 2015). *Nativeness* bereik je niet alleen door verbaal in een groep te passen door een accent dat door de gemeenschap herkend wordt, maar ook door het gebruik van typische discourse markers of tussenwerpsels zoals 'zeg' of 'hoor' en veelgebruikte interactiepatronen (LeVelle & Levis 2014; Miller 2003). Doorgaans worden niet-moedertaalsprekers namelijk pas gezien als 'accentvrij' wanneer ze zich niet enkel de standaardtalige uitspraakvariant van de doeltaal eigen hebben gemaakt, maar ook

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

regionale varianten en interactiepatronen hanteren met typische discourse markers voor de betreffende taalgroep (LeVelle & Levis 2014).

Verstaanbaarheid gaat over de mate waarin de luisteraar de spreker correct begrijpt, terwijl *begrijpelijkheid* naar de ervaring van verstaanbaarheid verwijst. Het eerste concept verwijst dus naar de mate van overeenstemming tussen de bedoelde boodschap en de begrepen boodschap (Derwing & Munro 2015; Suzuki & Kormos 2020). Waar verstaanbaarheid veeleer over objectief begrip gaat, verwijst begrijpelijkheid naar de mate waarin de luisteraar de spreker moeilijk of makkelijk te begrijpen vindt. Begrijpelijkheid wordt doorgaans gezien als een van de meest fundamentele aspecten van mondelinge communicatie: als de boodschap nauwelijks overkomt, wordt communicatie bemoeilijkt (Derwing & Munro 2015).

Accent, begrip en verstaanbaarheid staan niet los van elkaar, maar zijn deels overlappende constructen.[1] Zo kan een zwaar accent begrip en/of verstaanbaarheid in de weg staan, maar dat hoeft niet zo te zijn (Crowther 2020). Begrijpelijkheid zoals beoordeeld door moedertaalsprekers is volgens heel wat studies niet noodzakelijk verbonden met anderstalig accent (Pickering & Baker 2014). Sprekers met een beperkt anderstalig accent kunnen namelijk nog steeds moeilijk te begrijpen zijn (Munro & Derwing 1995a). Andere studies laten zien dat bepaalde aspecten van anderstalig accent compenseren voor begrijpelijkheid, terwijl andere aspecten er geen effect op hebben (Derwing & Munro 2015; Levis & Cortes 2008). In deze bijdrage zetten we in op de relatie tussen accent en begrijpelijkheid, en bestuderen we de verbanden met achtergrondvariabelen van zowel leerders als beoordelaars. We hebben daarbij bijzondere aandacht voor affiliatie (d.w.z. sociale verbondenheid) in relatie tot uitspraak.

1.2 De relatie tussen affiliatie en uitspraak

Historisch werd lang de nadruk gelegd op het belang van accentloze uitspraak bij tweede- of vreemdetaalleerders. Er werden dan ook al tal van studies uitgevoerd naar variabelen die accent beïnvloeden. Onderzoek geeft aan dat leeftijd (*age of acquisition*) en moedertaal daarbij belangrijke mediërende, maar geen allesbepalende, factoren zijn (Ortega 2008; Paradis 2018; Sereno & Wang 2007). Andere factoren die accent kunnen beïnvloeden, zijn taalcontact, taalattitude, motivatie en leerstijl (Crowther et al. 2015).

Waar in onderzoek de nadruk lang heeft gelegen op achtergrondvariabelen van leerders, is er recent ook steeds meer aandacht voor de sociale context waarin leerders zich bevinden. Sinds *the social turn* in

1 Voor een uitgebreide bespreking van de constructen en de relatie ertussen verwijzen we naar Simon & Lybaert, in dit nummer.

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

taalverwervingsonderzoek (The Douglas Fir Group 2016) nemen studies niet alleen leerdersvariabelen onder de loep, maar ook contextvariabelen en de interactie tussen leerder en context. Die focus sluit aan bij onderzoek naar de relatie tussen taalverwervingsprocessen (onder meer verwerving van uitspraak) en de sociale context (Levis & Moyer 2014). In navolging van de interactiehypothese, die stelt dat naast taalinput ook authentieke interactie van belang is om taal te verwerven (Long 1996), toont recent werk het belang aan van voldoende betekenisvolle interactie in de doeltaal voor accentverwerving (Moyer 2014). Studies van LeVelle & Levis (2014) en Levis & Moyer (2014) wijzen ook op het belang van affiliatie- en identificeringsmogelijkheden met de gemeenschap waarin de tweede taal wordt gesproken. Affiliatie verwijst daarbij naar sociale verbondenheid die zowel jonge als volwassen tweedetaalleerders motiveert én een belangrijke bron van taalinput en interactie vormt (Joaquin & Schumann 2013), en is steeds centraler komen te staan in taalverwervingsonderzoek. Affiliatie houdt ook verband met de theorie van *the interactional instinct,* die stelt dat taalverwerving gestuwd wordt vanuit inherente drijfveren om zich te verbinden met significante anderen (Frijns 2019; Lee et al. 2009; Joaquin & Schumann 2013). Taalverwerving, en specifiek verwerving van uitspraak, wordt daarbij niet als een louter cognitief of biologisch vermogen gezien, maar ook als een sociaal fenomeen waarbij affiliatie een sterke rol speelt. Sociale factoren die taalverwerving, en specifiek uitspraak, beïnvloeden zijn zo sinds *the social turn* steeds meer in de belangstelling komen te staan.

Onder invloed van die social turn zien we in de tweedetaalverwervingsliteratuur twee complementaire onderzoeksperspectieven: het ene is voornamelijk gericht op variabelen die rechtstreeks of onrechtstreeks accentverwerving beïnvloeden, het andere op hoe een spreker gepercipieerd wordt. In verband met de perceptie van de spreker: het oordeel van luisteraars over accent en begrijpelijkheid is het resultaat van talige kenmerken en affiliatieve factoren. Tweedetaalleerders en moedertaalsprekers nemen bijvoorbeeld fonologische kenmerken van een etnolect over, omdat ze zich identificeren met de sprekers van de betreffende variëteit (Marzo & Ceuleers 2011). In dat verband heeft Miller (2003: 47) het over *audibility*, een term die verwijst naar het oordeel dat moedertaalsprekers vellen over de sociale en talige integratie van niet-moedertaalsprekers, op basis van hun uitspraak. De percepties van de moedertaalspreker-luisteraar staan vanuit dat kader dus centraal als indicatie van de mate waarin tweedetaalleerders als *audible* worden gezien en de mogelijkheid hebben om aan sociale interacties te participeren (Miller 2003). Het concept *audibility* verwijst zo naar het idee van *legitimate peripheral participation*, dat sociale interactie niet louter

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

ziet als een kwestie van *willen*, maar zeker ook van *mogen* participeren in bepaalde sociale structuren (Lave & Wenger 1992; The Douglas Fir Group 2016). Een accent kan mee bepalen of iemand gezien wordt als deel van een bepaalde sociale *in-group* (McNamara 2020) en kan leiden tot bepaalde oordelen ten aanzien van de spreker (Lippi-Green 2012). Accent kan ook invloed hebben op percepties van taalvaardigheid, intelligentie, opleiding en sociale status (Derwing & Munro 2015; Lippi-Green 2012). Het geheel van die – vaak onbewuste – oordelen bepaalt mee of iemand gezien wordt als lid van een *in-group* en bijgevolg als legitieme conversatiepartner.

In interacties kennen luisteraars dus vaak impliciet en onbewust bepaalde kenmerken aan sprekers toe op basis van uitspraak en accent (*linguistic stereotype hypothesis*; Lambert, Hodgson, Gardner & Fillenbaum 1960). Zelfs wanneer luisteraars niets weten over de achtergrond van een spreker, maken ze op basis van onder andere uitspraak veronderstellingen over zijn of haar sociale en demografische achtergrond en de mate waarin de sprekers bijvoorbeeld vriendelijk of intelligent zijn (Lindemann 2002). Tweedetaalleerders lopen op basis van uitspraak overigens sterker het risico om negatiever te scoren voor gepercipieerde intelligentie dan eerstetaalleerders (Lindemann, Litzenberg & Subtirelu 2014). Ze kunnen ook als geschikte en minder geschikte representanten van hun eerste taal (T1) en tweede taal (T2) worden gezien. Zo toonden Gatbonton, Trofimovich & Magid (2005) aan dat tweedetaalleerders met een sterker anderstalig accent beschouwd worden als geschikte representanten of leiders van de T2-groep, terwijl tweedetaalleerders zonder een dergelijk anderstalig accent veeleer gezien worden als communicatoren van de T1-groep die geschikt zijn om relaties te onderhouden met 'outsiders' zoals T2-sprekers. De mate waarin een anderstalig accent gepercipieerd wordt, hangt met andere woorden sterk samen met percepties van integratie of groepsaffiliatie. Accent vervult dus een belangrijke functie inzake sociale en situationele affiliatie (Moyer 2015).

1.3 Percepties van ervaren en onervaren beoordelaars van beginnende NT2-leerders

Een *in-group* van moedertaalsprekers is uiteraard geen homogeen gegeven. Demografische en sociaal-economische verschillen tussen T1-beoordelaars maken dat de ene moedertaalspreker waarschijnlijk anders oordeelt over de uitspraak van een T2-leerder dan de andere moedertaalspreker. Onderzoeksdata hierover zijn echter beperkt: tot op heden gingen studies vooral na wat de impact is van achtergrondvariabelen van T2-leerders op het oordeel van moedertaalsprekers, en niet welke impact demografische kenmerken van moedertaalsprekers hebben op hun oordelen (zie bijvoorbeeld

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

Pinget et al. 2014). Onderzoek in gerelateerde disciplines geeft echter aan dat achtergrondvariabelen zoals migratie-achtergrond, thuistaal (één- of meertalig) en omgeving (bijvoorbeeld opgegroeid zijn in een stad met meer kansen om met diversiteit in contact te komen) een belangrijke en doorgaans onbewuste impact kunnen hebben op oordelen over diversiteit (Charlesworth & Banaji 2013; Hogan & Mallot 2005; Maly 2009). In deze studie nemen we daarom niet alleen achtergrondvariabelen van de leerders maar ook van de beoordelaars onder de loep.

1.4 Focus en onderzoeksvragen

In dit artikel hanteren we, in navolging van *the social turn* in tweedetaalverwervingsonderzoek en studies naar de relatie tussen sociale dynamiek en anderstalig accent (Levis & Moyer 2014), een sociolinguïstische oriëntatie op de concepten *begrijpelijkheid* en *anderstalig accent*. We volgen daarbij het hierboven gesitueerde belang van *audibility*, waarbij percepties van luisteraars een rol spelen in de mate waarin tweedetaalleerders beschouwd worden als volwaardige participanten van de T1-groep. Deze studie draagt op drie manieren bij aan de onderzoeksliteratuur.

Allereerst wordt er in deze studie niet alleen rekening gehouden met achtergrondvariabelen van leerders die het Nederlands als tweede taal verwerven maar ook met die van de beoordelaars (Pickering & Baker 2014), waaronder vertrouwdheid met anderstalige accenten en training in de beoordeling van spreekvaardigheid van NT2-leerders (Derwing & Munro 2015). Om de subjectiviteit van beoordelaars te reduceren, werken onderzoekers vaak met getrainde beoordelaars of met beoordelaars die vertrouwd zijn met het taalgebruik van tweedetaalleerders (Kang & Rubin 2014). Het is echter de vraag of dergelijke interventies de betrouwbaarheid van de score en de validiteit van het oordeel verhogen. Mensen die sterk gewend zijn aan T2-uitspraak oordelen immers niet noodzakelijk milder of anders dan mensen die geen ervaring hebben met niet-moedertaalsprekers (Moyer 2014). Daarnaast blijkt dat getrainde beoordelaars niet altijd in staat zijn om hun intuïtie opzij te zetten (Lumley 2002) of om dezelfde criteria op een uniforme manier te interpreteren (Deygers & Van Gorp 2015; Harsch & Martin 2012). Onderzoek heeft bovendien aangetoond dat ook niet-getrainde beoordelaars op een betrouwbare manier kunnen oordelen over uitspraak, grammatica en woordenschat (Saito et al. 2017) en over begrijpelijkheid en accent (Derwing et al. 2004). Het blijft echter onduidelijk in hoeverre de oordelen van getrainde en niet-getrainde beoordelaars overeenkomen (Suzuki & Kormos 2020). Ook vergelijkend onderzoek naar beoordelaars die meer of minder ervaring hebben met accenten blijft een zeldzaamheid

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

(Saito et al. 2017). In dit onderzoek vergelijken we het oordeel van getrainde beoordelaars die werken met gestandaardiseerde beoordelingscriteria met het oordeel van niet-getrainde beoordelaars die aan de start van hun lerarenopleiding staan en veelal werken vanuit hun intuïtie. De resultaten van ons onderzoek kunnen, ten eerste, een kritisch licht werpen op de assumptie dat beoordeling van accent en begrijpelijkheid getraind kan worden.

Ten tweede zetten we in op beginnende taalleerders met verschillende opleidingsachtergronden die Nederlands als doeltaal hebben. Het meeste vergelijkbaar onderzoek is zo goed als uitsluitend gebaseerd op gevorderde, hoger geschoolde taalleerders (Andringa & Godfroid 2020; Henrich et al. 2010; Kang & Rubin 2014; Tarone 2010), Dergelijke leerders verschillen onder andere in voorkennis, ervaring met cognitief belastende taaltaken en werkgeheugencapaciteit (Rüsseler et al., in review; Vagvoelgyi et al. 2016). Volgens de *trade off*-hypothese (Skehan 2009; Skehan & Foster 2001) – die stelt dat door de beperkingen van het werkgeheugen complexere inhoud kan leiden tot een *trade off* ten koste van bijvoorbeeld complexiteit, correctheid of vlotheid – is het dan ook mogelijk dat cognitief belastende taken een negatieve impact hebben op de begrijpelijkheid van startende tweedetaalleerders (Crowther et al. 2015). Het *trade off*-effect zou immers sterker kunnen doorwegen bij lager geschoolde taalleerders aangezien scholingsgraad en geletterdheid het werkgeheugen beïnvloeden (Kurvers 2015; Unsworth & Engle 2007; Vagvoelgyi et al. 2016). Daarnaast is er slechts weinig gerelateerd onderzoek verricht naar sprekers van niet-Indo-Europese moedertalen (Saito et al. 2017), doeltalen anders dan het Engels (Suzuki & Kormos 2020) of beginnende taalleerders. Concreet richt het huidige onderzoek zich op niveau A1 en A2 van het Europees Referentiekader (ERK; Council of Europe 2018). Die niveaus situeren zich op het niveau van een basisgebruiker (tegenover een onafhankelijke en vaardige gebruiker die zich respectievelijk op B- en C-niveau bevinden). Het A2-niveau heeft het voorbije decennium sterk aan belang gewonnen in de context van inburgering in de lidstaten van de Raad van Europa zoals Vlaanderen en Nederland. Van inburgeraars wordt doorgaans verwacht dat ze de taal op A2-niveau beheersen (Rocca et al. 2019).

Ten derde maken we gebruik van dialogische taken. In onderzoek tot nu toe worden vooral monologische taken gebruikt waarbij leerders vertellen wat ze zien op een reeks tekeningen (*picture-narrative tasks*), die moeilijk blijken te zijn voor lager opgeleide, beginnende taalleerders (zie verderop). Aangezien begrijpelijkheid gelinkt is aan taakcomplexiteit (Crowther 2020; Sasayama 2016), is het belangrijk om ook andere taaktypes te gebruiken (Suzuki & Kormos 2020). Bovendien worden vaak fragmenten

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

van 30 seconden gebruikt, terwijl samples van volledige gesprekken grotere ecologische validiteit hebben (Isaacs & Thomson 2013).

In deze studie pakken we dus enkele lacunes in het bestaande onderzoek aan. Samengevat bekijken we de rol van en interactie tussen sprekers- en beoordelaarsvariabelen in het oordeel over accent en begrijpelijkheid. We beantwoorden onderstaande onderzoeksvragen:

(1) Wat is de impact van demografische variabelen en de onderwijsachtergrond van NT2-leerders op hun uitspraakscores op een gecentraliseerde A2-toets?

(2a) Hoe verhouden oordelen over accent en begrijpelijkheid van getrainde en niet-getrainde beoordelaars zich tot elkaar rekening houdend met vertrouwdheid met een anderstalig accent?

(2b) Welke achtergrondvariabelen van NT2-leerders (zoals NT2-traject, ERK-niveau en verblijfsduur) beïnvloeden de oordelen van niet-getrainde beoordelaars over accent en begrijpelijkheid?

(2c) Welke gepercipieerde kenmerken (inburgering, intelligentie en sympathie) beïnvloeden de oordelen van niet-getrainde beoordelaars over accent en begrijpelijkheid?

(2d) Welk verband kunnen we vaststellen tussen de gepercipieerde intelligentie en gemeten intelligentie van NT2-leerders?

(3) Welk verband kunnen we vaststellen tussen de scores voor accent en begrijpelijkheid?

2. Materiaal en methode

2.1 Dataset

Voor dit onderzoek evalueren getrainde en niet-getrainde beoordelaars de gesproken prestaties van beginnende leerders van het Nederlands als tweede taal. De NT2-leerders die de prestaties leverden, legden twee taken af uit de gestandaardiseerde test die in Vlaanderen gebruikt wordt om na te gaan of inburgeraars voldoen aan de A2-taalvoorwaarde. In de eerste taak werd een gestructureerd gesprek gevoerd over de vrijetijdsbesteding en de woonsituatie van de deelnemer. In de tweede taak werd aan de deelnemers gevraagd om een korte reeks tekeningen te beschrijven.

De volledige NT2-dataset bestaat uit prestaties van 116 volwassenen die in Vlaanderen taallessen volgden in het kader van hun inburgeringstraject en vrijwillig deelnamen aan dit onderzoek. De mediaanleeftijd was 32 jaar (*Gem.* = 34, *SD* = 10). Op het moment van de spreektaken waren de deelnemers gemiddeld 4 jaar in België (*SD* = 5, *Med.* = 2). Ruwweg de helft van hen had net het A1-niveau afgewerkt, terwijl de andere helft net klaar was met het

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

A2-niveau. Tabel 1 bevat een overzicht van de demografische variabelen van de deelnemers.

Tabel 1 Demografische samenstelling dataset NT2-prestaties

Variabele		n	%
Geslacht	V	63	54%
	M	53	46%
ERK-niveau cursus	A1	57	49%
	A2	59	51%
NT2-traject	Traag	56	48%
	Standaard	31	27%
	Snel	29	25%
Moedertaal (top 3)	Arabisch	25	22%
	Berber	8	7%
	Engels	8	7%
Land van herkomst (top 3)	Afghanistan	16	14%
	Marokko	13	11%
	Irak	11	10%

De leerders werden gerekruteerd in drie verschillende types NT2-traject: traag, standaard en snel. Trage trajecten zijn gericht op leerders die functioneel ongeletterd zijn en niet verder hebben gestudeerd dan lager secundair onderwijs. Binnen de trage trajecten zijn er variaties in onderwijsduur en aanpak, afhankelijk van de onderwijsachtergrond en geletterdheid van de leerders. In deze studie worden alle trajecten van 240 uur of meer per ERK-niveau geclassificeerd als traag. Het publiek van standaardtrajecten is geletterd en heeft lager of hoger secundair afgerond. Mensen in snelle trajecten hebben hoger secundair of universitair onderwijs genoten. Per ERK-niveau voorziet het snelle traject 60 uur instructie, het standaardtraject 120 uur, het trage traject 240 uur. Nieuwkomers die geen diploma kunnen voorleggen, krijgen bij oriëntatie naar een NT2-traject vaak een COVAAR-test voorgelegd (Magez 2007; Verschueren et al. 2011). Deze gestandaardiseerde, taalonafhankelijke test van cognitieve vaardigheden en redeneer- en leervermogen kent verschillende drempelwaarden. Een score tot en met 22 impliceert dat iemands cognitieve vaardigheden beter passen bij een traag traject. Vanaf een score van 30 wordt een standaardtraject aanbevolen. Leerders met een score van 40 of hoger kunnen een versneld traject volgen.

2.2 Participanten

De gesprekken werden beoordeeld door getrainde en niet-getrainde beoordelaars. De getrainde beoordelaars waren vier vrouwelijke masterstudenten

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

aan een Vlaamse universiteit. Op het moment van dataverzameling waren ze alle vier twintig jaar oud en volgden ze de opleiding Taal- en Letterkunde met specialisatie Nederlands. Voordat ze aan de slag gingen met de scoring van de prestaties van de 116 deelnemers, volgden ze een standaardiserings-training van twee dagen en legden ze een proef af waarbij ze gekalibreerde samples beoordeelden. Wanneer hun scores binnen de foutenmarge van de gekalibreerde samples vielen, konden de beoordelaars beginnen met de beoordelingen van de prestaties in de eigenlijke studie.

Alle niet-getrainde beoordelaars zaten op het moment van hun deelname aan dit onderzoek in het eerste jaar van hun Educatieve Bacheloropleiding Lager Onderwijs aan verschillende Vlaamse hogescholen. De dataverzameling vond plaats aan het einde van het eerste semester of bij de start van het tweede semester 2019-2020 en werd uitgevoerd in het kader van een opleidingsonderdeel over taaldidactiek. Op die manier hadden alle niet-getrainde participanten één semester opleidingsonderdelen gevolgd aan de lerarenopleiding. Ze volgden geen beoordelaarstraining. Het profiel van de niet-getrainde beoordelaar was relatief homogeen: de typische niet-getrainde beoordelaar was 18 jaar (*Gem.* = 20, *SD* = 5, *Med.* = 18) en vrouwelijk (82%), groeide op in een landelijke of randstedelijke context (75%) en spreekt thuis Nederlands (89%). 56% volgde vooraf Algemeen Secundair Onderwijs (ASO), 37% Technisch Secundair Onderwijs (TSO) en 7% Beroeps Secundair Onderwijs (BSO).[2]

2.3 Materiaal en procedure

Om de interbeoordelaarsbetrouwbaarheid van de getrainde beoordelaars te kunnen schatten, werd één op de vier prestaties beoordeeld door alle beoordelaars, wat resulteerde in een hoge mate van exacte overeenkomst en een non-significant verschil inzake strengheid (83%, $\chi^2(2) = 2.8$, $p = .24$). Prestaties werden gescoord op inhoudelijke en vormelijke aspecten. Inhoudelijke aspecten werden binair gescoord: wanneer een leerder een bepaald inhoudelijk element op een begrijpelijke manier aan bod brengt, werd er een score van 1 toegekend voor dat item. Vormelijke aspecten werden gescoord op een vijfpuntenschaal die gebaseerd is op het ERK. Een score van 0 staat voor een prestatie die voor een bepaald criterium niet voldoet aan het pre-A1-niveau, terwijl een score van 4 aangeeft dat een prestatie het B1-niveau haalt of overstijgt. Het beoordelingsmodel bevatte 28 inhoudelijke items (18 voor taak 1; 10 voor taak 2) en 11 vormelijke items (taak 1: grammatica, woordenschat, uitspraak, vlotheid, algemeen linguïstisch bereik; idem voor

2 Ruwweg equivalent aan respectievelijk VWO, HAVO, VMBO.

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

Tabel 2 Overzicht geselecteerde prestaties voor niet-getrainde beoordelaars

	Land van Herkomst	Leef-tijd	Onderwijs	T1	Maanden in België	Totaal-score	Uit-spraak-score /8
A1_ST_1	Marokko	26	Lager secundair	Arabisch	24	12	2
A2_TR_1	Nigeria	38	Universitair	Bini	36	25	2
A1_TR_1	Nigeria	21	Lager	Engels	36	24	2
A1_SN_1	Thailand	31	Hoger secundair	Thai	2	26	4
A1_TR_2	Somalië	33	Lager	Soma-lisch	35	25	4
A1_SN_2	Frankrijk	28	Universitair	Frans	6	25	4
A2_TR_2	Marokko	41	Lager secundair	Arabisch	120	31	4
A1_TR_3	Ghana	37	Lager secundair	Twi	96	26	6
A2_TR_4	Bulgarije	43	Lager secundair	Bulgaars	14	32	6
A2_SN_1	VK	31	Universitair	Engels	36	32	6
A2_SN_2	Syrië	20	Hoger secundair	Arabisch	18	35	7
A2_SN_3	Albanië	20	Hoger secundair	Albanees	7	38	8
A2_ST_1	Syrië	42	Hoger secundair	Arabisch	12	33	8
A2_TR_3	Marokko	41	Lager secundair	Arabisch	252	37	8

taak 2, plus cohesie). Tabel 3 geeft een omschrijving van de verschillende beoordelingscriteria zoals ze geoperationaliseerd werden. Aangezien beide taken behoren tot een *high stakes* gestandaardiseerde test die momenteel in gebruik is en onderhevig is aan strikte copyrightrestricties, kunnen we de exacte vragen en het volledige beoordelingsmodel niet delen.

De niet-getrainde beoordelaars scoorden veertien geselecteerde prestaties uit de gestandaardiseerde toets. Die prestaties werden intuïtief beoordeeld aan de hand van negen criteria gebaseerd op de constructen *anderstalig accent* en *begrijpelijkheid* (zoals besproken in de literatuurstudie) en op de niveau-omschrijvingen van uitspraak en accent in het ERK. De prestaties werden geselecteerd op basis van totaalscore, uitspraakscore en vaardigheidsschatting volgens het *Item Response Theory*-model dat gebruikt werd voor onderzoeksvraag 1 (zie verderop). De geselecteerde

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

Tabel 3 Beoordelingscriteria niet-getrainde beoordelaars

Criterium	Vraag	Niveaus likertschaal
Begrijpelijkheid 1 (begrip taak 1)	Hoe goed begrijp je wat deze persoon in het weekend doet?	4
Begrijpelijkheid 2 (begrip taak 2)	Hoe goed begrijp je de woonsituatie van deze persoon?	4
Begrijpelijkheid 3 (boodschap)	Hoe goed kan deze persoon een boodschap overbrengen?	5
Accent 1 (correctheid)	Welke omschrijving past het best bij de uitspraak?	4
Accent 2 (storend)	Welke omschrijving past het best bij de uitspraak?	4
Accent 3 (anderstalig)	Welke omschrijving past het best bij het accent?	4
Sympathie	Hoe sympathiek vind je deze persoon?	5
Intelligentie	Hoe intelligent vind je deze persoon?	5
Ingeburgerd	Hoe ingeburgerd is deze persoon volgens jou?	5

prestaties vertegenwoordigen een breed spectrum van uitspraakscores, zoals gescoord door de getrainde beoordelaars. Daarnaast verschillen de geselecteerde prestaties op het vlak van de achtergrondvariabelen land van herkomst, leeftijd, onderwijsniveau, moedertaal en aantal maanden in België. Tabel 2 biedt een overzicht van de geselecteerde prestaties. De code van elke prestatie (kolom 1) is een combinatie van ERK-niveau (A1 of A2), NT2-traject (traag, standaard, snel) en een rangnummer.

De niet-getrainde beoordelaars scoorden prestaties aan de hand van negen criteria gerelateerd aan begrijpelijkheid en anderstalig accent en een Likertschaal (Tabel 3). Ze kregen geen training. Ze beluisterden elke prestatie twee keer en gaven een score in met hun smartphone, via een online platform. Omdat het niet mogelijk was om elke niet-getrainde beoordelaar alle veertien prestaties te laten beoordelen, construeerden we een *missing data design*. Elke niet-getrainde beoordelaar beoordeelde tussen de zes en de negen prestaties en twee prestaties werden collectief beoordeeld (A2_SN_1 en A2_ST_1). Daardoor was er voldoende overlap om het beoordelingspatroon van elke beoordelaar te schatten en voldoende data om tot een robuuste schatting van het oordeel over elke prestatie te komen. Het laagste aantal beoordelingen voor een prestatie was 124, het hoogste 272. Het *Many Facet Rasch*-model dat we opstelden (zie sectie 2.4), is ontworpen om met dergelijke designs om te gaan (Linacre 2012).

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

De niet-getrainde beoordelaars gaven na afloop van de beoordeling van de dataset op een vierpuntenschaal (helemaal niet vertrouwd/belangrijk tot zeer vertrouwd/belangrijk) antwoord op drie bijkomende vragen:

(1) Hoe vertrouwd ben je met de uitspraak van niet-moedertaalsprekers van het Nederlands? (*M* (*sd*) = 1.88 (0.78), *Med* = 2)
(2) Hoe belangrijk vind je het dat niet-moedertaalsprekers klinken zoals Vlamingen? (*M* (*sd*) = 2.85 (0.39), *Med* = 3)
(3) Hoe belangrijk vind je het dat migranten snel Nederlands leren? (*M* (*sd*) = 1.64 (0.78), *Med* = 1)

2.4 Analyse

Many Facet Rasch-analyse (MFRA) – een vorm van *item response theory* (IRT) – veronderstelt dat de score op een door menselijke beoordelaars gescoord item bepaald wordt door minstens drie parameters: de vaardigheid van de leerder, de moeilijkheid van een item of taak, en de strengheid van de beoordelaar. De inschattingen van vaardigheid van de leerder worden dus gezien als een functie van taakprestaties en beoordelaarsparameters. MFRA plot die gegevens op eenzelfde intervalschaal en gebruikt *logits* om vaardigheden, strengheid en moeilijkheid aan te duiden. In de MFR-output zijn we vooral geïnteresseerd in de zogeheten *measures*, die een beeld geven van de vaardigheid van een leerder, de gemiddelde vaardigheid van een populatie of de moeilijkheid van een item of taak. Hogere *measures* duiden op een verhoogde relatieve moeilijkheid (van items of taken) of vaardigheid (van leerders). Een andere belangrijke maat is *model fit,* een score die aangeeft of een bepaald element in de meting past in het geconstrueerde statistische model. *Infit*-waarden rond 1 geven aan dat een item of leerder in het statistische model past, terwijl die onder .50 kunnen duiden op *overfit* (redundantie). Waarden boven 1.5 signaleren een *misfit* en kunnen het statistische model verstoren (McNamara et al. 2019), waardoor ze het best verwijderd worden uit de analyse. In dit onderzoek werden de scores van 15 leerders uit de initiële analyse verwijderd.

Een bijkomende kracht van MFRA-analyses bestaat in de mogelijkheid om *differential item functioning* (DIF) na te gaan in achtergrondvariabelen. DIF treedt op wanneer bepaalde items of taken consequent en significant een andere moeilijkheidsgraad opleveren bij verschillende populaties, na controle voor het vaardigheidsniveau van de testpersonen (Aryadoust et al. 2011). DIF is op zichzelf niet noodzakelijk problematisch, maar *bias* is dat wel (Deygers 2019; McNamara et al. 2019). Er kan sprake zijn van *bias* wanneer grote aantallen items systematisch en aantoonbaar leiden tot hogere of lagere scores bij specifieke populaties, bijvoorbeeld omwille

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

van mogelijk construct-irrelevante factoren, zoals geslacht, educatieve achtergrond of thuistaal (Shaw & Imam 2013; Xi 2010). Daar waar DIF dus een louter statistische aangelegenheid is, vraagt het oordeel over *bias* een weloverwogen interpretatie (McNamara & Roever 2006).

Op basis van de logitscores van het IRT-model berekenden we *fair scores* voor uitspraak. Die scores zijn logitscores die getransformeerd zijn naar de oorspronkelijke beoordelingsschaal en compenseren voor eventuele afwijkingen van beoordelaars. Die scores kunnen gebruikt worden als afhankelijke variabelen in MANCOVA's en lineaire regressiemodellen. MFRA gebeurde met FACETS (Linacre 2015), MANCOVA met SPSS26, en alle andere analyses met *R* (R Core Team 2018). Binnen R gebruikten we de pakketten *car* (Fox & Weisberg 2011) voor regressieanalyses, *mlogit* (Croissant 2018) voor logistische regressie, *pwr* (Champely 2018) voor poweranalyses, en *psych* (Champely 2018) voor descriptieve statistiek, principale-componentenanalyse, correlaties, effectgroottes en *intraclass correlation* (ICC). We construeerden een MFRA-model bestaande uit tien facetten. De term *facet* staat voor een factor die een invloed kan hebben op de meting en de schatting van de *true score*. De facetten binnen het model waren leerder, beoordelaar, NT2-traject (Traag, Standaard, Snel), ERK-niveau (A1, A2), regio van herkomst (OECD-lidstaat of niet), leeftijd (15-24, 24-39, 40+), geslacht (M/V), verblijfsduur (1-2 jaar / meer dan 2 jaar), moedertaal (Indo-Europees of niet).

3. Resultaten

3.1 Impact van achtergrondvariabelen van leerders op uitspraakscores

Het gehanteerde MFR-model verklaarde 83.26% van de geobserveerde scorevariatie in de gestandaardiseerde spreektoets, zoals beoordeeld door getrainde beoordelaars. Een DIF-analyse van de achtergrondvariabelen verklaarde een aanvullende 4.99%. Zoals aangegeven, werd dit model geconstrueerd op basis van alle items op de spreektoets, dus niet louter de uitspraakgerelateerde items, maar ook de items inhoud, algemeen linguïstisch bereik, woordenschat, grammatica, coherentie en vlotheid. De resultaten in de volgende alinea schetsen dus een beeld van de kwaliteiten van de gehele toets. De DIF-resultaten gaan echter louter over de uitspraakscores.

Tabel 4 toont de vaardigheidsschattingen (*measures*) van de verschillende subgroepen, opgedeeld volgens achtergrondvariabelen. Het IRT-model werd

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

Tabel 4 MFR-resultaten, scoreverschillen tussen demografische groepen

		Meas.	*(se)*	*Infit*	χ^2	df	*p*	Range	Betrouwbaarheid
ERK-niveau	A1	34.58	(.55)	.94	358.2	1	.000	14%	.99
	A2	48.90	(.51)	.91					
	Snel	50.49	(.69)	1.02					
NT2-traject	Standaard	37.43	(1.04)	.92	254.7	2	.000	13%	.98
	Traag	37.30	(.5)	.87					
Regio	OECD	46.13	(.87)	.86	82.1	1	.000	9%	.99
	Non-OECD	37.35	(.42)	.94					
Leeftijd	18-24	52.80	(.93)	.9	258.6	2	.000	18%	.99
	24-39	37.62	(.5)	.93					
	40+	34.80	(.73)	.92					
Verblijf	<2j	44.70	(.53)	.98	61.5	1	.000	6%	.97
	>2j	38.78	(.54)	.87					
Taal	Indo-Eur	44.97	(.54)	.86	73.1	1	.000	6%	.97
	Non-Indo-Eur	38.51	(.52)	.98					

zo opgesteld dat alle variabelen interageren met alle andere variabelen, waardoor geschatte vaardigheidsscores de impact van alle andere variabelen in rekening brengen. Tabel 4 bevat enkel de variabelen waarbinnen significante verschillen werden gevonden en biedt een overzicht van de vaardigheidsschattingen voor elk facet. De tabel geeft ook aan hoe groot het verschil is in *measures* binnen een facet (*range*) en hoe betrouwbaar de schatting is. Betrouwbaarheid is hier het Rasch-equivalent van KR-20, de proportie ware variantie ten opzichte van de geobserveerde variantie voor de elementen in een facet.

De data tonen aan dat de grootste prestatieverschillen te vinden zijn binnen de variabelen leeftijd, ERK-niveau en NT2-traject. Het verschil in vaardigheid tussen leerders in het trage traject en het standaardtraject is verwaarloosbaar, maar leerders in het snelle traject doen het significant en substantieel beter dan leerders in de andere trajecten. Hetzelfde patroon zien we terugkeren bij leeftijd, waar de jongste groep significant hoger scoort dan de twee andere groepen.

Na controle voor beoordelaarsafwijking en de genoemde achtergrondvariabelen zien we dat het moeilijkste item (het item waarbij de kans het kleinst is dat een kandidaat met een gemiddeld vaardigheidsniveau de hoogste score behaalt) uit de A2-spreektoets een *measure* heeft van 66.38 en het makkelijkste een *measure* van 0.38 ($\chi^2(38) = 1134.2$, $p = .000$). De scores voor

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

uitspraak verschillen significant per taak: een hoge uitspraakscore halen, is significant moeilijker voor de *picture-narrative task* (Taak 2 measure = 53.89 (1.16), Infit = 1.06) dan voor de dialoog (Taak 1 measure = 44.08 (1.18), Infit = 1). De DIF-analyses werden uitgevoerd per taak, omdat we niet wilden veronderstellen dat de impact van bijvoorbeeld cognitieve belasting gelijk was over taken heen (zie de *trade-off hypothesis* hierboven), en ook niet dat de dialogische taak en de *picture-narrative task* eenzelfde construct zouden meten. De DIF-analyse voor uitspraakscores op Taak 1 resulteerde enkel in significante verschillen voor NT2-traject (snel = 64.6 (*se* 2.7), traag = 54.9 (*se* 2); $t(97) = -2.78$, $p = .006$) en ERK-niveau (A2 = 61.7 (*se* 2.1), A1 = 54 (*se* 2.4); $t(114) = -2.41$, $p = .017$). Voor Taak 2 vonden we significante verschillen voor wat betreft NT2-traject (snel = 51.9 (*se* 2.7), traag = 44.3 (*se* 2); $t(97) = -2.19$, $p = .03$). Aangezien hogere uitspraakscores geassocieerd zijn met hogere ERK-niveaus en met NT2-programma's waarin het niveau doorgaans substantieel hoger ligt (Deygers et al., in review), is er geen directe reden om aan te nemen dat er in deze uitspraakscores sprake is van bias.

Aangezien uit de MFR-analyses blijkt dat de uitspraakscores op de twee taken significant en substantieel verschillen, construeerden we een MANCOVA-model met de uitspraakscores op de twee taken als afhankelijke variabelen. Poweranalyse toonde aan dat de steekproefgrootte afdoende was om kleine tot medium effecten ($f^2 = .06$) te vinden bij $\alpha = .05$ en $\beta = .8$. Het model toont aan dat ERK-niveau ($F(2) = 24.613$, $p < .001$, partial $\eta^2 = .317$) en NT2-traject ($F(2) = 31.640$, $p < .001$, partial $\eta^2 = .374$) allebei een significante impact hebben op de uitspraakscores, en dat de leeftijd en T1 van de respondent op hun beurt een significant effect hebben op ERK-niveau en NT2-traject. Tabel 5 toont de resultaten, uitgesplitst per taak. Het MANCOVA-model verklaart 60% van de variantie voor uitspraak in Taak 1 (= .602) en in Taak 2 (= .6). Wanneer gecontroleerd wordt voor leeftijd en T1 blijkt de impact van ERK-niveau en NT2-traject op uitspraakscores significant te zijn, met een medium effectgrootte. Een mogelijk interactie-effect tussen NT2-traject en ERK-niveau blijkt niet significant te zijn.

3.2 De rol van achtergrondvariabelen

3.2.1 De samenhang van getrainde en niet-getrainde beoordelaars

Om na te kunnen gaan of de oordelen van getrainde en niet-getrainde beoordelaars samenhangen, is het noodzakelijk om de samenhang tussen de twee groepen te bekijken. De interbeoordelaarsbetrouwbaarheid van de getrainde beoordelaars is hoog (zie boven). Hetzelfde geldt voor de niet-getrainde beoordelaars. De Chi-kwadraatstest die de interbeoordelaarsbetrouwbaarheid

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

Tabel 5 MANCOVA, Fairscores uitspraak ~ achtergrondvariabelen (between-subject effects)

	df	*F*	*p*	Partial η^2	*B (SE)*
Uitspraak, taak 1					
Model	5	34.865	.000	0.62	
Intercept	1	797.342	.000	0.882	-2.365 (0.123)
leeftijd[a]	1	5.077	.026	0.045	-0.245 (0.109)
T1[b]	1	4.229	.042	0.038	-0.171 (0.083)
ERK-niveau	1	49.691	.000	0.317	-0.652 (0.094)
NT2-traject	1	61.425	.000	0.365	-0.742 (0.135)
Uitspraak, taak 2					
Model	5	34.561	.000	0.618	
Intercept	1	714.133	.000	0.87	-2.188 (0.122)
leeftijd[a]	1	5.556	.02	0.049	-0.255 (0.108)
T1[b]	1	4.455	.037	0.04	-0.174 (0.082)
ERK-niveau	1	49.235	.000	0.315	-0.660 (0.093)
NT2-traject	1	58.496	.000	0.353	-0.735 (0.134)

[a] dummy variabele: 0 = ≤24, 1 = >24
[b] dummy variabele: 0 = Indo-Europees, 1 = niet-Indo-Europees

voor de niet-getrainde beoordelaars nagaat, is niet significant (χ^2 (100) = 97.2, p = .56). Ter controle berekenden we de ICC voor accentcriteria (correctheid, storend of hinderlijk accent, anderstalig accent) op basis van de twee prestaties die door alle niet-getrainde beoordelaars beoordeeld werden. De ICC is hoog, zowel globaal (.96) als uitgesplitst per criterium (.97 voor Accent 1, .95 voor Accent 2 en .96 voor Accent 3).

In een volgende stap vergeleken we de uitspraakoordelen van de getrainde en de niet-getrainde beoordelaars door de fairscores van de prestaties te correleren met de fairscores van diezelfde prestaties na beoordeling door getrainde beoordelaars. Wanneer de overeenkomst niet exact was, verschilde de rangorde maximaal één plaats, wat resulteerde in een sterke rangcorrelatie tussen de beide groepen beoordelaars (τ = .85, p < .001).

3.2.2 *De impact van achtergrondvariabelen van leerders en beoordelaars*

In een eerste stap creëerden we een MFR-model om de *measures* van de criteria gekoppeld aan accent, begrijpelijkheid, intelligentie, inburgering en sympathie in te schalen volgens moeilijkheid (zie Tabel 6). In de analyse

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

Tabel 6 MFR-itemschatting

	Meas. (se)	Infit	Discrim.
Accent 1 (correctheid)	58.27 (.96)	.60	1.43
Accent 3 (anderstalig)	52.96 (.97)	1.41	.52
Accent 2 (storend)	52.30 (.97)	.63	1.36
Intelligent	51.09 (1.02)	.86	1.15
Ingeburgerd	49.90 (1.02)	1.15	.83
Begrijpelijkheid 2 (begrip taak 2)	44.67 (.99)	.91	1.19
Begrijpelijkheid 1 (begrip taak 1)	41.08 (1.01)	1.16	.84
Sympathiek	36.47 (1.04)	1.28	.69
Begrijpelijkheid 3 (boodschap)	34.23 (1.09)	.96	1.09
($\chi^2(8) = 507.6$, $p = .000$)			

werd gecontroleerd voor strengheid van beoordelaar. Een stippellijn geeft aan dat het verschil in *measure* met aanpalende criteria significant is. De criteria die rechtsreeks peilden naar de beoordeling van uitspraak en accent bleken het strengst beoordeeld te zijn. Criteria die peilden naar de mate waarin NT2-leerders de boodschap konden overbrengen en naar hoe sympathiek ze leken, werden significant hoger gescoord dan de andere criteria. De verschillen tussen de drie uitspraakgerelateerde criteria waren significant ($p < .001$), met medium effectgrootte r (.34 – .47). Ook tussen alle andere criteria bleken de verschillen significant te zijn ($p < .001$).

In een volgende stap gingen we na of bepaalde beoordelaarskenmerken een impact hadden op de scores voor accent. Vervolgens voegden we naast beoordelaarskenmerken ook leerderskenmerken toe aan het IRT-model. Het enige significante verschil dat overblijft, is dat beoordelaars met een stedelijke achtergrond significant milder in uitspraakoordelen zijn dan mensen die in een landelijke of randstedelijke context opgegroeid zijn ($\chi^2(1) = 5.9$, $p = .02$), maar de verschillen zijn klein. Gepercipieerde vertrouwdheid met accenten van niet-moedertaalsprekers van het Nederlands bleek geen significante impact te hebben op de beoordeling. Daarnaast zien we dat de grootste verschillen in oordelen toe te schrijven zijn aan het ERK-niveau van de leerder (Separation 13.90, Reliability.99; $\chi^2(1) = 388.5$, $p = .000$) en het NT2-traject, al is het effect kleiner dan bij het ERK-niveau (Separation 6.75, .98; $\chi^2(1) = 93.1$, $p = .000$). Ook voor moedertaal vonden we een significant effect (Separation 6.48, Reliability .98; $\chi^2(1) = 86$, $p = .000$). Prestaties van NT2-leerders met een hoger ERK-niveau,

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

Tabel 7 Lineaire regressie, Begrijpelijkheid

	B	SE	β	*t*	*p*	Betrouwbaarheids-interval	
						2.5 %	97.5%
(Constant)	.205	.035		5.873	0	.136	.273
Accent 1 (correctheid)	.321	.016	.306	19.536	.000	.289	.354
Accent 2 (storend)	.266	.017	.247	15.670	.000	.233	.299
Accent 3 (anderstalig)	.057	.011	.065	5.059	.000	.035	.079
Sympathiek	.032	.013	.028	2.438	.015	.006	.059
Intelligent	.035	.016	.029	2.159	.031	.003	.066
Ingeburgerd	.288	.014	.303	21.319	.000	.262	.315
							R^2_{adj} = .768

een Indo-Europese thuistaal en een sneller NT2-traject werden dus significant hoger beoordeeld dan prestaties van NT2-leerders met andere achtergrondvariabelen.

3.2.3 Het verband tussen gepercipieerde en gemeten intelligentie

Via logistische regressie bepaalden we in hoeverre de oordelen over intelligentie de COVAAR-score konden voorspellen. Een COVAAR-score van 30 geeft aan dat iemand zich kan inschrijven voor een standaardtraject. Leerders die 30 of hoger scoorden, codeerden we met 1, leerders die lager scoorden met 0. De oordelen van de niet-getrainde beoordelaars over gepercipieerde intelligentie vormden een niet-significante en zwakke voorspeller van de COVAAR-score (Est.(se) = 0.24(1.15), z = .21, p = .8, 95% CI: .12 < 1.27 > 14.68 / Nagelkerke pseudo R^2 = .004). Met andere woorden, gepercipieerde intelligentie is een slechte voorspeller voor werkelijke cognitieve vaardigheden, zoals vastgesteld via de COVAAR-test of via een attest hoger onderwijs. Intelligentie-oordelen worden vooral bepaald door de mate waarin een uitspraak als storend gepercipieerd wordt (B(se) = .36(.02), t(3) = 17.97, p < .000) en door begrijpelijkheid (B(se) = .22(.01), t(3) = 15.88, p < .000). Ook uitspraakfouten (B(se) = .06(.02), β = .07, t(3) = 3.25, p < .000) droegen significant, maar in mindere mate, bij tot het model (.52).

3.3 De relatie tussen accent en begrijpelijkheid

Om de relatie tussen accent en begrijpelijkheid verder uit te diepen, modelleerden we twee stepwise lineaire regressiemodellen. In een eerste

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

Tabel 8 Lineaire regressie, accent

	B	SE	*β*	*t*	*p*	Betrouwbaarheids-interval	
						2.5 %	97.5%
(Constant)	-0,188	0,031		-6,035	0	-0,249	-0,127
Begrijpelijkheid 3 (Boodschap)	0,163	0,014	0,203	11,955	0	0,136	0,189
Begrijpelijkheid 2 (begrip taak 2)	0,231	0,013	0,267	17,489	0	0,205	0,257
Ingeburgerd	0,199	0,014	0,225	14,288	0	0,171	0,226
Begrijpelijkheid 1 (begrip)	0,141	0,012	0,161	12,103	0	0,119	0,164
Intelligent	0,133	0,014	0,122	9,652	0	0,106	0,16
							$R^2_{adj} = .74$

model (Tabel 7) werd de gemiddelde score van de begripscriteria (Begrip 1 en Begrip 2, Boodschap) gebruikt als afhankelijke variabele. Het meest robuuste model (Tolerance = 0.34; VIF = 3.11; Durbin-Watson = 1.7) was significant meer voorspellend dan eerdere modellen in de stepwise regressiesequentie ($p < .000$). Dat model verklaarde zo'n 77% van de scorevariatie. De steekproefgrootte ($n = 272$) is adequaat om kleine effecten ($f^2 = .05$) vast te stellen bij $\alpha = .05$ $\beta = .9$. Uit de regressie blijkt dat naast Accent 1 en Accent 2 ook de gepercipieerde mate van inburgering een sterke voorspeller van begrijpelijkheid is.

Verder modelleerden we accent als afhankelijke variabele (Tabel 8). Dit model (Tolerance = 0.33; VIF = 3.14 ; Durbin-Watson = 1.68) stond toe om kleine effecten ($f^2 = .049$) te meten bij $\alpha = .05$ en $\beta = .9$. Het verklaarde zo'n 74% van de scorevariatie voor het criterium en behield, naast Begrijpelijkheid 3 (boodschap) en Begrijpelijkheid 2 (begrip taak 2), de gepercipieerde mate van inburgering als sterkste voorspeller.

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

4. Discussie

De resultaten van dit onderzoek tonen aan dat twee variabelen de uitspraakscores op een gecentraliseerde spreektoets, gescoord door getrainde beoordelaars, het sterkst bepalen: NT2-traject (hoe sneller het traject, hoe hoger de score) en ERK-niveau (A2-leerders scoren hoger dan A1-leerders). Die variabelen worden gemedieerd door leeftijd en moedertaal: oudere leerders met een niet-Indo-Europese moedertaal scoren lager op uitspraak in beide taken, ook na controle voor ERK-niveau en NT2-traject. We zien geen effecten van verblijfsduur of herkomstregio. Deze resultaten bevestigen wat we vinden in de literatuur, waar ook leeftijd en T1 vaak genoemd worden als beïnvloedende variabelen (Ortega 2008; Paradis 2018; Sereno & Wang 2007), net zoals de beginsituatie of de voorkennis van leerders.

Een belangrijke toevoeging die het huidige onderzoek biedt, is de variabele NT2-traject. Onderzoek heeft zich tot nu toe zo goed als uitsluitend toegelegd op universitair geschoolde leerders (Andringa & Godfroid 2019; Kang & Rubin 2014). Onze studie toont echter aan dat het belangrijk is om fijnmaziger te werk te gaan en het specifieke traject in het model mee te nemen: het traject verklaart meer scorevariantie dan het ERK-niveau van de cursus die de taalleerders volgen. Natuurlijk is NT2-traject als voorspeller niet helemaal eenduidig: mensen worden aan een bepaalde stroom toegewezen omwille van cognitieve vaardigheden en vooropleiding. Toekomstig onderzoek kan nagaan wat het precieze aandeel van die twee elementen in de scoreverdeling is. In elk geval toont deze studie aan dat educatieve of cognitieve variabelen een plaats hebben in dit soort onderzoek.

Dit onderzoek bevestigt ook dat taaktype een impact heeft op uitspraakscores (Crowther et al. 2015). De uitspraakscores op de cognitief relatief eenvoudige dialogische taak lagen substantieel en significant hoger dan de scores op de cognitief meer belastende monologische *picture-narrative* taakDat bevestigt het belang van variatie in taaktypes in uitspraakonderzoek (Suzuki & Kormos 2020) én ondersteunt de *trade off*-hypothese (Skehan & Foster 2001). Die hypothese voorspelt dat een grotere cognitieve belasting resulteert in mindere resultaten voor bijvoorbeeld vlotheid. Ook Robinsons *cognition hypothesis* voorspelt een negatieve impact van cognitieve belasting op vlotheid (Robinson 2001).

De resultaten met betrekking tot onderzoeksvraag 2a tonen duidelijk aan dat de rangorde van kandidaten die de getrainde en de niet-getrainde

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

beoordelaars scoorden, zeer gelijkaardig was (zie ook Isaacs & Thomson 2013). Ook binnen de groep van de niet-getrainde beoordelaars was de interbeoordelaarsovereenkomst groot. In die zin bevestigt dit onderzoek eerder werk dat de betrouwbaarheid van niet-getrainde beoordelaars naging (Isaacs & Thomson 2013). De vaststelling dat de oordelen van de twee groepen beoordelaars zo gelijklopend zijn, kan betekenen dat uitspraakbeoordeling een intuïtieve kwestie is die weinig training behoeft, of waar training weinig effect op heeft. Interviews of luidopdenkprotocollen kunnen helpen om meer inzicht te krijgen in de redeneringen achter de oordelen.

Verder zagen we dat beoordelaars met een stedelijke achtergrond doorgaans positiever oordeelden over uitspraak dan mensen die in een meer rurale context opgroeiden. Een verklaring hiervoor kan liggen in het feit dat mensen met een stedelijke achtergrond meer contact hebben met niet-moedertaalsprekers dan mensen met een meer landelijke achtergrond. Eerder onderzoek (Kang & Rubin 2014) stelt ook dat de mate van blootstelling aan de uitspraak van niet-moedertaalsprekers beoordelaars milder zou kunnen maken ten opzichte van non-standaard uitspraak en accenten, maar dat effect zagen we niet terug in de data. Gepercipieerde blootstelling aan NT2-accenten bleek de beoordeling niet significant te beïnvloeden. Verder gaven zowel de getrainde als de niet-getrainde beoordelaars significant hogere scores aan NT2-leerders met een hoger ERK-niveau, een Indo-Europese moedertaal en een sneller NT2-traject.

Ten slotte tonen de data dat accent en begrijpelijkheid onderling sterk verklarende factoren zijn, maar dat ze eveneens sterk verklaard worden door inschattingen van hoe ingeburgerd een NT2-leerder is. De mate van veronderstelde inburgering was een robuuste verklarende variabele in de regressiemodellen van accent en begrijpelijkheid. Dat veronderstellingen over de persoon achter de gesproken prestatie zeker niet juist hoeven te zijn, bleek uit het feit dat de inschattingen van hoe intelligent een spreker was, geen betrouwbare voorspeller waren voor de scores op een test van cognitieve vaardigheden.

5. Conclusie en implicaties

In deze studie onderzochten we de perceptie van begrijpelijkheid en accent bij beginnende NT2-leerders door getrainde en niet-getrainde beoordelaars. De impact van leeftijd, moedertaal, en taaktypes op oordelen over uitspraak werd in dit onderzoek bevestigd. Het idee dat beoordelaars uit een stedelijke context die gewend zijn aan diversiteit, milder staan tegenover vreemde

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

accenten werd in deze studie geconsolideerd. Belangrijke nieuwe inzichten die uit dit artikel naar voren komen, zijn het belang van NT2-traject als voorspellende variabele, of de mate waarin getrainde en niet-getrainde beoordelaars gelijklopen in hun oordeel. De verblijfsduur van de leerders bleek geen effect te hebben op uitspraakscores. Mogelijks is verblijfsduur geen betrouwbare proxy voor de mate van interactie: NT2-leerders kunnen immers al enige tijd in België verblijven zonder veel in betekenisvolle interactie te kunnen participeren met Nederlandssprekenden – terwijl betekenisvolle interactie cruciaal blijkt te zijn om uitspraak te verwerven. Het onderwijs zet daarom het best in op de kracht van betekenisvolle interactie, authentiek taalcontact en de ontwikkeling van een positieve houding ten opzichte van diversiteit in accenten (Derwing & Munro 2015). Toekomstig onderzoek over de mate waarin NT2-sprekers toegang krijgen tot betekenisvolle interactie binnen en buiten de NT2-klas kan verder uitwijzen hoe de combinatie van verblijfsduur en kwaliteitsvolle interactie bijdraagt aan uitspraakverwerving en participatiekansen aan de maatschappij.

Referenties

Andringa, S. & A. Godfroid (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics* 40, 134-142.

Aryadoust, V., C. Goh & L. Kim (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly* 8, 361-385.

Champely, S. (2018). *Pwr: Basic Functions for Power Analysis*. R package version 1.2-2.

Charlesworth, T. & M. Banaji (2019). Patterns of implicit and explicit attitudes: I. Long- term change and stability from 2007 to 2016. *Psychological Science* 30, 174-192.

Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors.* Brussels: Council of Europe Language Policy Division.

Croissant, Y. (2018). *Mlogit: Multinomial Logit Models*. R package version 0.3-0.

Crowther, D. (2020). Rating L2 speaker comprehensibility on monologic vs. interactive tasks: What is the effect of speaking task type? *Journal of Second Language Pronunciation* 6, 96-121.

Crowther, D., P. Trofimovich, T. Isaacs & K. Saito (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal* 99, 80-95.

Derwing, T. & M. Munro (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research.* Amsterdam/Philadelphia: Benjamins.

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

Derwing, T., M. Rossiter, M. Munro & R. Thomson (2004). Second language fluency: Judgments on different tasks. *Language Learning* 54, 655-679.

Deygers, B. (2019). Fairness and social justice in English language assessment. In: X. Gao, F. Davidson & C. Leung (red.), *Springer second international handbook of English language teaching*. New York: Springer.

Deygers, B. & K. Van Gorp (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing* 32, 521-541.

Deygers, B., M. Vanbuel & U. Knoch (in review). Can L2 course duration compensate for the impact of demographic and educational background variables on second language writing development? *System*.

Fox, J. & S. Weisberg (2011). *An {R} Companion to Applied Regression*. Thousand Oaks: Sage.

Frijns, C. (2019). *De vliegtuigklas. Naar sterk taalonderwijs op de basisschool*. Kalmthout: Pelckmans Pro.

Gatbonton, E., P. Trofimovich & M. Magid (2005). Learners' ethnic group affiliation and L2 pronunciation accuracy: A sociolinguistic investigation. *TESOL Quarterly* 39, 489-511.

Harsch, C. & G. Martin (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing* 17, 228-250.

Henrich, J., S. Heine & A. Norenzayan (2010). The weirdest people in the world? *Behavioral and Brain Sciences* 33, 1-75.

Hogan, D. & M. Mallot (2005). Changing racial prejudice through diversity education. *Journal of College Student Development* 46, 115-125.

Isaacs, T. & R. Thomson (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly* 10, 135-159.

Joaquin, A. & J. Schumann (2013). *Exploring the interactional instinct*. Oxford: Oxford University Press.

Kang, O. & D. Rubin (2014). Listener expectations, reverse linguistic stereotyping, and individual background factors in social judgments and oral performance assessment. In: J. Levis & A. Moyer (red.), *Social dynamics in second language accent*. Berlin: De Gruyter Mouton, 239-255.

Kurvers, J. (2015). Emerging literacy in adult second-language learners: A synthesis of research findings in the Netherlands. *Writing Systems Research* 7, 58-78.

Lambert, W., R. Hodgson, R. Gardner & S. Fillenbaum (1960). Evaluational reactions to spoken language. *Journal of Abnormal and Social Psychology* 60, 44-51.

Lave, J. & E. Wenger (1992). *Situated learning: Legitimate peripheral participation.* Cambridge: Cambridge University Press.

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

Lee, N., L. Mikesell, A. Joaquin, A. Mates & J. Schumann (2009). *The interactional instinct. The evolution and acquisition of language.* Oxford: Oxford University Press.

LeVelle, K. & J. Levis (2014). Understanding the impact of social factors on L2 pronunciation: Insights from learners. In: J. Levis & A. Moyer (red.), 79-118.

Levis, J. & V. Cortes (2008). Minimal pairs in spoken corpora: Implications for pronunciation assessment and teaching. In: C. Chapelle, Y.-R. Chung & J. Xu (red.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment*. Ames: Iowa State University, 197-208.

Levis, J. & A. Moyer (2014). *Social dynamics in second language accent*. Berlin: De Gruyter Mouton.

Linacre, J. (2012). *A User's Guide to FACETS Rasch-Model Computer Programs*. <winsteps.com>

Linacre, M. (2015). *Facets* (3.71.4) [Computer software]. <winsteps.com>

Lindemann, S. (2002). Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the United States. *Language in Society* 31, 419-441.

Lindemann, S., J. Litzenberg & N. Subtirelu (2014). Problematizing the dependence on L1 norms in pronunciation teaching: Attitudes toward second-language accents. In: J. Levis & A. Moyer (red.), 171-194.

Lippi-Green, R. (2012). *English with an accent: Language, ideology and discrimination in the United States*. New York: Routledge.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In: W. Ritchie & T. Bhatia (Eds.), *Handbook of language acquisition: Second language acquisition*. New York: Academic Press, 413-468.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing* 19, 246-276.

Magez, W. (2007). *Cognitieve vaardigheidstest* (*Covaar*) *voor anderstalige nieuwkomers e.a. Handleiding. 4. Testinstructies*. Centrum Schoolpsychologie KU Leuven Lessius Hogeschool Antwerpen – dept. Toegepaste Psychologie.

Maly, I. (2009). *De beschavingsmachine. Wij en de islam.* Berchem: EPO.

Marzo, S. & E. Ceuleers (2011). The use of Citétaal among adolescents in Limburg: The role of space appropriation in language variation and change. *Journal of Multilingual and Multicultural Development* 32, 451-464.

Miller, J. (2003). *Audible difference: ESL and social identity.* Clevedon: Multilingual Matters.

McNamara, T. (2020). The anti-shibboleth: The traumatic character of the shibboleth as silence. *Applied Linguistics* 41, 334-351.

McNamara, T. & C. Roever (2006). *Language testing: The social dimension*. Oxford: Blackwell Publishing.

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

McNamara, T., U. Knoch & J. Fan (2019). *Fairness, justice, and language assessment: The role of measurement.* Oxford: Oxford University Press.

Moyer, A. (2014). The social nature of L2 pronunciation. In: A. Moyer & J. Levis (red.), 11-31.

Moyer, A. (2015). *Foreign accent. The phenomenon of non-native speech.* Cambridge: Cambridge University Press.

Munro, M. (2008). Foreign accent and speech intelligibility. In: J. Hansen Edwards & M. Zampini (red.), *Phonology and second language acquisition*. Amsterdam/Philadelphia: Benjamins, 193-218.

Munro, M. & T. Derwing (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45, 73-97.

Munro, M. & T. Derwing (1995b). Processing time, accent, and comprehensibility in the perception of foreign-accented speech. *Language and Speech* 38, 289-306.

Ortega, L. (2008). *Understanding second language acquisition*. New York: Routledge.

Paradis, J. (2018, September 7). The L2 acquisition of children from migrant backgrounds: How age, L1, input and cognitive factors shape this process. Münster: EuroSLA.

Pickering, L. & A. Baker (2014). Suprasegmental measures of accentedness. In: A. Moyer & J. Levis (red.), 74-94.

Pinget, A-F, H. Bosker, H. Quené & N. de Jong (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing* 31, 349-365.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <www.R-project.org>

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics* 22, 27-57.

Rocca, L., C. Carlsen & B. Deygers (2019). *Linguistic Integration of adult migrants: Requirements and learning opportunities. Report on the 2018 Council of Europe and ALTE survey on language and knowledge of society policies for migrants.* Brussels: Council of Europe.

Rüsseler, J., D. Arendt, T. Münte, B. Mohammadi & M. Boltzmann (in review). Literacy affects brain structure – what can we learn for language assessment in low literates? *Language Assessment Quarterly.*

Saito, K., P. Trofimovich & T. Isaacs (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics* 38, 439-462.

Sasayama, S. (2016). Is a 'complex' task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal* 100, 231-254.

Sereno, J. & Y. Wang (2007). Behavioral and cortical effects of learning a second language. In: O.-S. Bohn & M. Munro (Eds.), *Language experience in second*

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41

language speech learning: In honor of James Emil Flege. Amsterdam/Philadelphia: Benjamins, 239-258.

Shaw, S. & H. Imam (2013). Assessment of international students through the medium of English: Ensuring validity and fairness in content-based examinations. *Language Assessment Quarterly* 10, 452-475.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics* 30, 510-532.

Skehan, P. & P. Foster (2001). Cognition and tasks. In: P. Robinson (Ed.), *Cognition and second language instruction.* Cambridge: Cambridge University Press, 183-205.

Suzuki, S. & J. Kormos (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition* 42, 143-167.

Tarone, E. (2010). Second language acquisition by low-literate learners: An understudied population. *Language Teaching* 43, 75-83.

The Douglas Fir Group (2016). A transdisciplinary framework for SLA in a multilingual world. *The Modern Language Journal* 100, 19-47.

Unsworth, N. & R. Engle (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin*, 133, 1038-1066.

Vagvoelgyi, R., A. Coldea, T. Dresler, J. Schrader & H.-C. Nuerk (2016). A review about functional illiteracy: Definition, cognitive, linguistic, and numerical aspects. *Frontiers in Psychology*, 7, 1617.

Verschueren, K., E. Buyse, V. Germeijs, R, Janssen, W. Magez, D. Van Nijlen, G. Buysse, S. Vangoetsenhoven, T. Arkens & S. Doumen (2011). *Evaluatie en aanpassing van de Covaar-II.* Faculteit Psychologie en Pedagogische Wetenschappen, KU Leuven. Onderzoeksgroep Schoolpsychologie en Ontwikkelingspsychologie van Kind en Adolescent.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing* 27, 147-170.

Universiteitsbibliotheek Gent (leesgentbel)
IP: 157.193.239.41