

Radboud Universiteit

MA Taalwetenschap: Nederlands als Tweede Taal: Docent en Expert

Masterscriptie

Het verschil in strategiegebruik tussen een DOMC- toets en een standaard meerkeuzetoets bij NT2-leerders

Anne van Asseldonk

prof. dr. A (eerste begeleider)

dr. B (tweede begeleider)

21-1-2021

Voorwoord

Voor u ligt de scriptie ‘Het verschil in strategiegebruik tussen een DOMC-toets en een standaard meerkeuzetoets bij NT2-leerders’. Het is een onderzoek naar het strategiegebruik van NT2-leerders in een luistertoets en de invloed die het toetsformat daarop heeft. Deze masterscriptie is geschreven in het kader van de afronding van mijn masteropleiding Taalwetenschap met als specialisatie Nederlands als Tweede Taal aan de Radboud Universiteit. Van september 2020 tot en met januari 2021 ben ik bezig geweest met het onderzoek en het schrijven van deze scriptie.

Het schrijven van deze scriptie gaf me veel inzicht in mijn academische ontwikkeling. Sinds (en mede dankzij) het schrijven van mijn bachelorscriptie ben ik in dat opzicht gegroeid, waardoor ik nu relatief zelfstandig dit onderzoek kon uitvoeren. Daarnaast is voor mij nog duidelijker geworden waar mijn interesses binnen de taalwetenschap liggen. Met deze scriptie hoop ik bij te dragen aan het verbeteren en vernieuwen van taaltoetsing, iets wat ik ook in de toekomst graag zou willen blijven doen.

Bij het opzetten en schrijven van mijn scriptie ben ik begeleid door A. Hij was ook de eerste begeleider van mijn bachelorscriptie en die begeleiding heb ik destijds als prettig ervaren. Dit was wederom het geval. Ik wil A daarom bij dezen graag bedanken voor zijn feedback en het meedenken over problemen en eventuele oplossingen. Daarnaast wil ik mijn tweede lezer B bedanken voor het beoordelen van deze scriptie.

Mijn dank gaat ook uit naar alle mensen die mij hebben geholpen bij het werven van proefpersonen. Het zijn er te veel om hier op te noemen, maar ik hoop dat ze zich aangesproken voelen wanneer ze dit lezen. Zij hebben de cursisten die binnen mijn doelgroep vielen benaderd om mee te doen aan mijn onderzoek. Dankzij hen heb ik een grote groep cursisten kunnen bereiken. De cursisten die besloten deel te nemen, ben ik ook zeer dankbaar. Zonder hen had ik mijn onderzoek niet kunnen uitvoeren.

Ten slotte wil ik C en D bedanken. Zij hebben mij het patent op de DOMC-vraag verleend, waardoor ik dit onderzoek kon uitvoeren.

Anne van Asseldonk,

Nijmegen, 21 januari 2021

Inhoudsopgave

Voorwoord	i
Inhoudsopgave.....	ii
Abstract.....	1
Hoofdstuk 1: Inleiding	2
1.1 Relevantie.....	2
1.2 Onderzoeksvraag.....	3
1.3 Hypothesen	4
1.4 Leeswijzer	4
Hoofdstuk 2: Theoretisch kader	4
2.1 Meerkeuzevragen.....	4
2.2 Drie soorten toetsstrategieën	6
2.2.1 Taalleerderstrategieën	7
2.2.2 Toetsmanagementstrategieën	7
2.2.3 Toetswijsheidstrategieën	8
2.3 Manieren om toetswijsheidstrategieën te voorkomen	9
2.3.1 <i>Discrete-option multiple-choice</i>	11
2.4 Strategieën in luistervaardigheid.....	12
2.5 Mondelinge rapporten en vragenlijsten over strategiegebruik	13
2.6 Vaardige versus minder vaardige kandidaten	15
2.7 Staatsexamens NT2	15
2.7.1 Luistervaardigheidsexamen	16
Hoofdstuk 3: Methode	16
3.1 Onderzoeksontwerp.....	16
3.2 Materiaal	17
3.2.1 Hardopdenkprotocol.....	18
3.2.2 Implementatie Qualtrics	19
3.3 Proefpersonen.....	19
3.4 Procedure	20
3.4.1 Hardopdenkprotocol.....	20
3.5 Data-analyse.....	21
Hoofdstuk 4: Resultaten.....	21
4.1. Analyse vragenlijst over strategiegebruik.....	21

4.1.2 Strategiegebruik (kwantitatief)	21
4.1.3 Strategiegebruik (kwalitatief).....	24
4.2 Tijdsduur.....	28
4.3 Gemiddelde score.....	29
4.4 Terughoudendheid	30
Hoofdstuk 5: Discussie en conclusie.....	31
5.1 Bespreking resultaten.....	31
5.2 Verklaringen resultaten	32
5.3 Generaliseerbaarheid naar Staatsexamen.....	33
5.4 Meer onderzoek naar DOMC-toetsen	34
5.5 Conclusie	35
Literatuurlijst.....	36
Bijlage I: Alle toetsvragen en antwoorden	39
Bijlage II: Lijst met strategieën en bijbehorende categorie	43

Abstract

In taaltoetsen kan construct-irrelevante variantie voorkomen door ongewenst strategiegebruik van de kandidaat. De kandidaat doorloopt dan niet de beoogde processen bij het beantwoorden van een vraag. De *discrete option multiple choice* (DOMC) vraag biedt hier mogelijk een oplossing voor. Dit type vraag, waarbij de antwoordopties sequentieel worden getoond, lijkt in eerdere onderzoeken ongewenst strategiegebruik te verminderen ten opzichte van de standaard meerkeuzevraag. In de huidige studie werd dit onderzocht in de context van een luistertoets voor tweedetaalleerders van het Nederlands. De vraag die centraal stond, was: 'Hoe verschillen de strategieën die kandidaten inzetten in een DOMC-toets van de strategieën die kandidaten inzetten in een standaard meerkeuzetoets?' Om deze vraag te beantwoorden werd een luistertoets met een vragenlijst over het strategiegebruik van de kandidaat uitgezet. In totaal maakten 45 tweedetaalleerders een van de twee versies van de toets. Enkel van hen namen deel aan een hardopdenkprotocol. Er bleken met name veel overeenkomsten in strategiegebruik tussen de twee condities te zijn. Taalleerderstrategieën werden in beide condities het minst ingezet. Desondanks bleek uit de kwalitatieve analyse van het strategiegebruik dat er voornamelijk construct-relevant gedrag vertoond werd. Er waren echter ook aanwijzingen voor verschillen in strategiegebruik: de proefpersonen die de standaard meerkeuzetoets maakten, gebruikten het proces van eliminatie vaker en proefpersonen in de DOMC-conditie gaven aan het moeilijker te vinden om het goede antwoord te gokken. Er is nog weinig onderzoek gedaan naar de DOMC-vraag, waardoor er nog veel vragen onbeantwoord zijn. Hoewel uit de huidige studie geen grote voordelen van het DOMC-format bleken, was dat in andere onderzoeken wel het geval. Het is daarom de moeite waard om te onderzoeken of en in welke gevallen het DOMC-format valide ingezet kan worden.

Hoofdstuk 1: Inleiding

Het is een gegeven in de wereld van de toetsing dat een toets valide moet zijn. Volgens Messick (1989) is validiteit de mate waarin het gerechtvaardigd is om op basis van de toetsscore gevolgtrekkingen over een construct te maken. Hij noemt construct-irrelevante variantie als bedreiging voor de toetsvaliditeit. Construct-irrelevante variantie houdt in dat de toetstaak teveel aspecten meet, waaronder aspecten die irrelevant zijn voor het construct. Dit is het geval wanneer externe hints, los van vaardigheid, naar het goede antwoord leiden (waardoor de toets te makkelijk is) of wanneer externe hints, los van vaardigheid, voor onnodig foute antwoorden zorgen (waardoor de toets te moeilijk is). Het is niet gemakkelijk om construct-irrelevante variantie volledig te voorkomen. Zelfs de meest gebruikte toetsformats, zoals de meerkeuzetoets, zijn onderhevig aan construct-irrelevante variabelen. De zoektocht naar de manier hoe het (nog) beter kan, blijft daarom bestaan. Door middel van de huidige studie hoop ik een steentje bij te dragen aan deze zoektocht. De standaard meerkeuzevraag wordt onder de loep genomen en vergeleken met een meerkeuzevraag die in 2009 voor het eerst geïntroduceerd werd door Foster en Miller: de *discrete option multiple choice* (DOMC) vraag. Het is een alternatieve manier van toetsen, waarbij de antwoordmogelijkheden sequentieel getoond worden in plaats van simultaan (Papenberg, Willing, & Musch, 2017). Wanneer de kandidaat een incorrecte antwoordmogelijkheid als incorrect aanwijst, verdwijnt het alternatief en krijgt hij het volgende alternatief te zien. Wordt het correcte antwoord als incorrect afgedaan, gaat de kandidaat naar de volgende vraag. Ook als de kandidaat een incorrecte antwoordmogelijkheid als correct markeert, krijgt hij de andere alternatieven niet meer te zien en gaat hij door naar de volgende vraag. De DOMC-vraag kent volgens een aantal wetenschappers enkele voordelen boven de standaard meerkeuzevraag (o.a. Foster & Miller, 2009; Papenberg, Willing, & Musch, 2017; Willing, Ostapczuk, & Musch, 2015). Zij benoemden onder andere dat kandidaten de toets sneller kunnen maken en dat fraude beperkt wordt doordat niet alle alternatieven worden getoond. Een ander belangrijk voordeel is dat er minder construct-irrelevante variantie, veroorzaakt door het gebruik van ongewenste toetsstrategieën, lijkt te zijn. In deze scriptie ga ik onderzoeken wat (on)gewenste strategieën zijn en of het strategiegebruik van kandidaten in een DOMC-toets verschilt van dat van kandidaten in een standaard meerkeuzetoets. Dit wordt onderzocht door het strategiegebruik van tweedetaalleerders van het Nederlands (NT2-leerders) in een luistertoets (ofwel DOMC, ofwel standaard meerkeuze) van een oefen-Staatsexamen NT2 te vergelijken. Hierbij wordt een onderscheid gemaakt tussen drie soorten strategieën: taalleerder-, toetsmanagement- en toetswijsheidstrategieën. In het kort zijn taalleerderstrategieën de manieren waarop kandidaten hun taalvaardigheden en de gerelateerde vaardigheden van woordenschat, grammatica en vertaling inzetten om tot het goede antwoord te komen. De toetsmanagementstrategieën zijn de strategieën die ingezet worden om betekenisvol op vragen en taken te kunnen antwoorden. De categorie toetswijsheidstrategieën wordt gedefinieerd als het in je voordeel kunnen inzetten van items die externe hints bevatten, zonder kennis te hebben van het onderdeel dat getoetst wordt. Uitbreidende definities volgen in paragraaf 2.2.

1.1 Relevantie

De Staatsexamens NT2 zijn *high stakes* toetsen. De resultaten van de examens zijn belangrijk, want ze kunnen toegang tot de Nederlandse maatschappij geven of ontzeggen. Het is daarom van belang dat de examenuitslag een eerlijke en rechtvaardige uitslag is over het kennen en kunnen van de kandidaat. Een valide toets draagt bij aan die eerlijke en rechtvaardige uitslag. Er moet zo min mogelijk construct-irrelevante variantie zijn. In die zin contribueert dit onderzoek aan het waarborgen van de kwaliteit

van het luisterexamen NT2. Het DOMC-format kan daarnaast mogelijk een oplossing bieden voor het fraudegevoelige systeem van de Staatsexamens.

Het onderzoek naar DOMC-vragen is tot op heden beperkt. Er is nog weinig bekend over de processen die kandidaten doorlopen bij het beantwoorden van een DOMC-vraag. Papenberg, Diedenhofen en Musch (2019) kaartten aan dat de kennis over de beslissingen (inclusief gokstrategieën die daarbij ingezet worden) van kandidaten in sequentiële formats vergroot moet worden. Bovendien is het onderzoek naar luistervaardigheid relatief summier. Meer inzicht in de luisterstrategieën die NT2-leerders inzetten of zouden moeten inzetten om succesvolle luisteraars te zijn, kan van hulp zijn voor de leerders zelf en voor hun NT2-docenten. Je bewust zijn van de strategieën die je in zou moeten zetten lijkt namelijk een positieve invloed te hebben op het luistergedrag van een taalleerder (Vandergrift, Goh, Mareschal, & Tafaghodtari, 2006). NT2-docenten kunnen hier in hun lessen op inspelen. Door te onderzoeken welke strategieën kandidaten gebruiken tijdens het maken van een DOMC-luistertoets en een meerkeuzeluistertoets hoop ik de kennis over zowel DOMC als luistervaardigheid te vergroten.

1.2 Onderzoeksvraag

De vraag die in de huidige studie centraal staat, luidt als volgt:

Hoe verschillen de strategieën die kandidaten inzetten in een DOMC-toets van de strategieën die kandidaten inzetten in een standaard meerkeuzetoets?

De deelvragen die onderzocht worden, zijn:

1. Wat zijn construct-irrelevante en -relevante strategieën?

2. Welke strategieën zetten kandidaten in tijdens de DOMC-toets?

- Tot welke categorieën behoren deze strategieën?

- Zijn dit construct-relevante strategieën?

3. Welke strategieën zetten kandidaten in tijdens de standaard meerkeuzetoets?

- Tot welke categorieën behoren deze strategieën?

- Zijn dit construct-relevante strategieën?

4. Wat zijn de verschillen tussen het strategiegebruik van de kandidaten in de DOMC-toets en het strategiegebruik van de kandidaten in de standaard meerkeuzetoets?

- Is het strategiegebruik in een van deze toetsen te prefereren boven het strategiegebruik in de andere toets?

De eerste deelvraag zal door middel van literatuuronderzoek beantwoord worden in het theoretisch kader. De overige drie deelvragen worden experimenteel onderzocht met behulp van een Qualtrics-vragenlijst die bestaat uit een luistertoets en vragen over het strategiegebruik van de proefpersonen in die toets. Enkele proefpersonen zullen ook deelnemen aan een hardopdenkprotocol. De doelgroep van dit onderzoek zijn NT2-leerders op B1-niveau.

In de huidige studie worden de twee toetsen ook vergeleken op twee variabelen die in eerdere onderzoeken zijn genoemd. Het betreft de gemiddelde score en de tijdsduur. Deze variabelen bleken in andere onderzoeken over het algemeen lager te liggen bij de DOMC-toets (Foster & Miller, 2009;

Papenberg, Willing, & Musch, 2017). Daarnaast zal het aantal getoonde alternatieven in de DOMC-conditie worden berekend. Ook zal er een vraag met betrekking tot het al dan niet snel willen accepteren van een antwoord worden gesteld om te onderzoeken of er een verband is met de gemiddelde score.

1.3 Hypothesen

Ik verwacht dat zowel in de DOMC-toets als in de standaard meerkeuzetoets vooral gewenste strategieën uitgelokt worden. Taalleerderstrategieën zullen het meest gebruikt worden, gevolgd door toetsmanagementstrategieën en ten slotte toetswijsheidstrategieën. De proefpersonen in de huidige studie zijn immers gevorderde (en vaak ook hoogopgeleide) tweedetaalleerders. Deze groep blijkt minder snel terug te vallen op compensatiestrategieën en kan beter gebruik maken van taalleerderstrategieën. Deze hypothese sluit tevens aan bij de bevinding van Cohen en Upton (2006) dat de gevorderde leerders uit hun onderzoek meer toetsmanagement- dan toetswijsheidstrategieën gebruikten.

Het verschil tussen de DOMC-toets en de standaard meerkeuzetoets zit naar verwachting niet in de volgorde van de categorieën die ingezet worden, maar in de omvang van het strategiegebruik. Eerdere onderzoeken lieten zien dat het gebruik van toetswijsheidstrategieën vermindert in een DOMC-toets. De antwoordmogelijkheden zijn namelijk niet zo gemakkelijk met elkaar te vergelijken als bij een meerkeuzetoets. In de standaard meerkeuzetoets voorspel ik daarom een grotere rol voor toetswijsheidstrategieën dan in de DOMC-toets. Wanneer dit het geval is, zorgt het DOMC-format ervoor dat de construct-irrelevante variantie veroorzaakt door toetswijsheidstrategieën kleiner wordt.

1.4 Leeswijzer

In Hoofdstuk 1 werd de centrale onderzoeksvraag ingeleid en geïntroduceerd. Hoofdstuk 2 vormt de wetenschappelijke basis voor de huidige studie. Ook zal deelvraag 1 hier worden beantwoord. In Hoofdstuk 3 wordt de methodologie van de experimentele deelvragen besproken. De resultaten daarvan zijn in Hoofdstuk 4 te vinden. In Hoofdstuk 5 worden de resultaten bediscussieerd en zal er een conclusie worden getrokken.

Hoofdstuk 2: Theoretisch kader

2.1 Meerkeuzevragen

De meerkeuzevraag is veelvoorkomend in het Nederlandse onderwijssysteem. Hij wordt in allerlei soorten toetsen, opgaven en vakgebieden gebruikt: van *high stakes* tot *low stakes* toetsen, van productieve opgaven tot receptieve opgaven en van de wiskunde tot de psychologie. De meerkeuzevraag is zo frequent, omdat hij goed onderzocht is en allerlei voordelen kent. De meest genoemde voordelen van een meerkeuzevraag zijn de snelle en betrouwbare scoring ervan. Met name in grote groepen is de efficiënte beoordeling, die vaak ook geautomatiseerd gebeurt, een groot voordeel. Tevens kan er weinig onduidelijkheid omtrent de antwoorden van de kandidaten ontstaan. Daarnaast worden meerkeuzevragen over het algemeen sneller beantwoord dan open vragen, waardoor er meer vragen gesteld kunnen worden in dezelfde tijd. In het geval van een luistervaardigheidstoets zou je onder andere vragen willen stellen om verschillende deeltaalvaardigheden

te toetsen (bijvoorbeeld globaal en selectief luisteren), om begrip van diverse tekstsoorten te testen (zoals informatief, instructief en persuasief) en om begrip van teksten met verschillende thema's te testen (bijvoorbeeld werk en opleiding). De kans op onderrepresentatie van het te meten construct wordt kleiner als er veel vragen gesteld worden, wat de constructvaliditeit van de toets vergroot (Kingston, Tiemann, Miller Jr. & Foster, 2012). Een groter aantal vragen vergroot ook de betrouwbaarheid van een toets, want de toets is consistent (Hughes, 2002). Een ander voordeel van de meerkeuzevraag is het feit dat construct-irrelevante factoren zoals het handschrift van de kandidaat vermeden worden. De docent of examiner kan zich hierdoor niet laten beïnvloeden. Door het meerkeuzeformat spelen echter wel weer andere construct-irrelevante variabelen een rol. Een voorbeeld hiervan is de aanwezige gokkans en het inzetten van trucjes om die gokkans te maximaliseren (toetswijsheid).

Een nadeel van een meerkeuzetoets is dus de aanwezige gokkans. Wanneer er vier antwoordmogelijkheden zijn, heeft een kandidaat zonder enige kennis 25% kans om het goede antwoord te kiezen. Hier kan bij het bepalen van de cesuur echter betrekkelijk goed rekening mee gehouden worden door het toepassen van een gokkanscorrectie. Een ander nadeel is dat het construeren van een meerkeuzevraag een moeilijke taak is. De kwaliteit van een vraag is erg afhankelijk van de toetsontwikkelaar. Het construeren van een meerkeuzevraag is een complex en langdurig proces (Hughes, 2002). Hoewel er sprake is van tijdwinst voor de beoordelaar door de snelle beoordeling van een meerkeuzevraag, is de toetsontwikkelaar juist veel tijd kwijt aan het construeren van goede meerkeuzevragen. Om deze reden worden meerkeuzevragen vaak hergebruikt. Dit maakt de meerkeuzetoets fraudegevoelig, waardoor de antwoorden van de toets beschermd moeten worden. Docenten kiezen er daarom regelmatig voor om de toets niet in een klaslokaal te bespreken, maar laten de kandidaten naar hun kantoor komen. In werkelijkheid blijken weinig kandidaten van die mogelijkheid gebruik te maken. De kandidaten weten dan niet of de antwoorden die zij gaven goed of fout waren. Dit kan ervoor zorgen dat er incorrecte kennis gecreëerd wordt bij de kandidaten: zij blijven denken dat een door hen gegeven antwoord juist is, hoewel het in werkelijkheid een onjuist antwoord was. Roediger III (2005) merkte op dat de afleiders van een meerkeuzetoets ertoe kunnen leiden dat een kandidaat niet zoveel leert van een toets als mogelijk is. In het ideale geval zou de kandidaat namelijk leren van zijn fouten en deze in het vervolg niet meer maken. Dit leerproces wordt verstoord, omdat de kandidaten bij een meerkeuzevraag over het algemeen blootgesteld worden aan meer incorrecte dan correcte antwoorden. De proefpersonen in Roedigers onderzoek gaven in een toets incorrecte antwoorden op basis van afleiders uit een eerdere toets. Het verhogen van het aantal afleiders bleek de kans op het creëren van incorrecte kennis te vergroten. Roediger noemt dit het negatievesuggestie-effect. Als een meerkeuzevraag niet goed geconstrueerd is, veroorzaakt dat nog meer problemen. Er kan bijvoorbeeld meer dan één correct antwoord zijn. Ook komt het voor dat geen enkele optie volledig correct is. Ten slotte kunnen er hints naar het goede antwoord voorkomen, waardoor de gokkans boven het veronderstelde kansniveau uitkomt. Dit geeft een oneerlijk voordeel aan kandidaten met een grote toetswijsheid; dat zijn kandidaten die goed gebruik kunnen maken van deze hints. Dit was onder andere het geval in de onderzoeken van Allan (1992), Katz, Lautenschlager, Blackburn & Harris (1990) en Daneman en Hannon (2001). De toets meet in deze gevallen niet wat er gemeten moet worden, waardoor de validiteit in het geding komt. Het zoeken naar externe hints is een toetsstrategie die een kandidaat kan inzetten om tot het goede antwoord te komen. Over het algemeen worden er drie soorten strategieën onderscheiden, sommige relevanter voor het te meten construct dan andere (Cohen, 2006).

2.2 Drie soorten toetsstrategieën

De drie soorten toetsstrategieën die onderscheiden worden zijn taalleerderstrategieën, toetsmanagementstrategieën en toetswijsheidstrategieën (Cohen, 2014). De categorie **taalleerderstrategieën** behelst de manieren waarop kandidaten hun taalvaardigheden en de gerelateerde vaardigheden van woordenschat, grammatica en vertaling inzetten om tot het goede antwoord te komen. Een voorbeeld hiervan is het zoeken naar betekenismarkeringen (zoals definities of voorbeelden) in een passage. Het inzetten van dit soort strategieën is niet kwalijk, omdat de kandidaat de beoogde vaardigheden inzet om tot het goede antwoord te komen. Taalleerderstrategieën kunnen weliswaar ingezet worden in een toets, maar het zijn eigenlijk geen toetsstrategieën. Deze strategieën kunnen namelijk ook buiten de toets ingezet worden om een gesprek of tekst te begrijpen.

De toetsmanagement- en toetswijsheidstrategieën vallen samen onder de noemer toetsafnamestrategieën (Cohen, 2006) en worden ook wel compensatiestrategieën genoemd (Cohen & Upton, 2006). De **toetsmanagementstrategieën** zijn volgens Cohen (2014) de strategieën die ingezet worden om betekenisvol op vragen en taken te kunnen antwoorden. Hieronder valt ook het keuzes maken op basis van de beschikbare tijd, zoals het toekennen van meer tijd aan vragen die de meeste punten kunnen opleveren. De laatste categorie, **toetswijsheidstrategieën**, wordt gedefinieerd als het in je voordeel kunnen inzetten van items die externe hints bevatten, zonder kennis te hebben van het onderdeel dat getoetst wordt (Tavakoli & Samian, 2014). Het verschil tussen deze twee toetsafnamestrategieën is volgens Cohen dat de eerste bijdraagt aan construct-relevante variantie, terwijl de tweede slechts bijdraagt aan construct-irrelevante variantie.

Wat deze verschillende categorieën exact inhouden, en hoe ze te classificeren, blijft echter werk in uitvoering. Er zijn in essentie twee contrasterende ideeën over wat een toetsstrategie is (Cohen, 2006). Aan de ene kant is er de visie dat een toetsstrategie specifiek, klein en deel van een grotere set strategieën is. Aan de andere kant wordt gedacht dat een toetsstrategie meer globaal, flexibel en algemeen is. In de huidige studie wordt uitgegaan van de eerste visie. De strategieën die in de huidige studie onderzocht worden, vallen namelijk onder een van de drie grotere sets strategieën die zojuist besproken zijn. Het is belangrijk om op te merken dat algemeen wordt aangenomen dat kandidaten toetsstrategieën bewust inzetten. Als dit een voornamelijk onbewust proces was, zouden kandidaten namelijk niets zinnigs kunnen zeggen over hun strategiegebruik. Het onderscheid tussen de toetsmanagement- en toetswijsheidstrategieën is ook niet altijd helder. Een strategie kan in een bepaald onderzoek onder de noemer toetsmanagement vallen, hoewel de strategie binnen de definitie van toetswijsheid valt. Dit is bijvoorbeeld het geval bij de strategie 'Wanneer ik een vraag beantwoord, gebruik ik hints uit andere vragen om het antwoord te gokken' die door Wu en Stone (2016) geclassificeerd wordt als toetsmanagementstrategie, hoewel deze strategie volgens de hierboven genoemde definitie onder de noemer toetswijsheidstrategieën zou moeten vallen. Winke en Lim (2014) maken in hun onderzoek zelfs geen onderscheid meer tussen de twee typen toetsafnamestrategieën. Cohen (2014) heeft geprobeerd de drie soorten strategieën van elkaar te onderscheiden, maar mijns inziens houdt het onderscheid dat Cohen maakt tussen toetsmanagement- en toetswijsheidstrategieën geen stand. Het efficiënt kunnen toekennen van de beschikbare tijd zegt niets over de te meten taalvaardigheid en is daarom naar mijn mening niet construct-relevant. Logistieke strategieën zoals het letten op de tijd worden in de huidige studie daarom niet meegenomen. De eerste deelvraag 'Wat zijn construct-irrelevante en -relevante strategieën?' kan dan als volgt in het kort beantwoord worden: taalleerderstrategieën zijn altijd construct-relevant en toetswijsheidstrategieën zijn altijd construct-irrelevant. Toetsmanagementstrategieën zijn alleen construct-relevant wanneer het geen logistieke zaken betreft.

In de komende drie paragrafen worden de definities van de drie categorieën verder uitgewerkt. De definities zijn voornamelijk gebaseerd op die van Cohen (2014), maar aangepast naar mijn eigen interpretatie.

2.2.1 Taalleerderstrategieën

Cohen (2014, op basis van zijn eigen werk in 2011) definieert taalleerderstrategieën als volgt: *'Thoughts and actions, consciously chosen and operationalized by language learners, to assist them in carrying out a multiplicity of tasks from the very onset of learning to the most advanced levels of target-language performance.'*

Taalleerderstrategieën worden ingezet om de betekenis van een tekst te begrijpen of om een vraag te kunnen beantwoorden. In het geval van luistervaardigheid zijn dit de luisterstrategieën. Bij een luisterbegripvraag kan het bijvoorbeeld nodig zijn om inferentiestrategieën in te zetten. Cohen (2014) illustreert dit in zijn artikel met een tekst waarin het woord *updates* wordt gebruikt. De kandidaten moeten hieruit afleiden dat het om een computer gaat. Dit soort subvaardigheden zijn construct-relevant, omdat de toetsontwikkelaars willen dat die strategieën ingezet en die processen doorlopen worden. In het dagelijks leven moet dat namelijk ook gebeuren. In de context van een toets zijn de subvaardigheden nodig om de vraag te kunnen beantwoorden. In de context van het dagelijks leven zijn ze nodig om een boodschap te begrijpen en eventueel een gepaste reactie te formuleren. Taalleerderstrategieën worden over het algemeen in drie categorieën onderverdeeld (Taguchi, 2001). Het betreft cognitieve strategieën (weer onderverdeeld in *top-down* en *bottom-up*), metacognitieve strategieën (kennis over en het reguleren van cognitie), en emotionele strategieën (zoals het aanmoedigen van jezelf).

In een valide toets zouden kandidaten die deze strategieën veelvuldig inzetten een hogere score moeten behalen dan kandidaten die voornamelijk gebruikmaken van toetswijsheidstrategieën. In het onderzoek van Wu en Stone (2016) werd op basis hiervan de pilot leestoets van het *Canadian English Language Proficiency Index Program-General (CELP-IP-General)* gevalideerd. Het gebruik van taalleerderstrategieën was een positieve voorspeller voor de prestaties op de toets (.24). Het gebruik van toetsmanagementstrategieën (-.23) en toetswijsheidstrategieën (-.50) bleken juist negatieve voorspellers te zijn. De taalleerderstrategieën die in dit onderzoek onderscheiden werden, waren het begrijpen van het hoofdidee van een passage, het begrijpen van specifieke zinnen in een passage en het herlezen van stukken tekst. Het probleem met deze indeling is dat hij beperkt is en daardoor nuance mist. Hierdoor worden de verschillen tussen de categorieën (in dit geval taalleerder- en toetsmanagementstrategieën) moeilijker te duiden.

2.2.2 Toetsmanagementstrategieën

Toetsmanagementstrategieën zijn processen die bewust worden geselecteerd om verantwoord een antwoord te kiezen (Cohen, 2014). Een voorbeeld van een toetsmanagementstrategie is het kiezen van een antwoordoptie, omdat de kandidaat weet dat de andere fout zijn. Dit draagt bij aan construct-relevante variantie, omdat er kennis gebruikt is om de vraag te beantwoorden. Hoewel de kandidaat misschien niet overtuigd was van het goede antwoord, heeft hij wel opties kunnen elimineren op basis van wat hij (al dan niet) gehoord of gelezen heeft. In het licht hiervan is het vreemd dat het inzetten van toetsmanagementstrategieën een negatieve voorspeller voor de toetsresultaten was in de studie van Wu en Stone (2016). Een mogelijke verklaring hiervoor is dat zij, zoals al eerder vermeld werd, de toetswijsheidstrategie 'Wanneer ik een vraag beantwoord, gebruik ik hints uit andere vragen om het

antwoord te gokken' classificeerden als toetsmanagementstrategie. Het verschil met taalleerderstrategieën zit in het feit dat taalleerderstrategieën ook buiten toetsen worden ingezet, hoewel toetsmanagementstrategieën alleen tijdens toetsen worden ingezet.

Ook wanneer kandidaten toetsmanagementstrategieën zouden moeten gebruiken om een vraag te beantwoorden, kan het zo zijn dat de kandidaten niet de cognitieve processen doorlopen die volgens de toetsontwikkelaars nodig zijn om tot het correcte antwoord te komen. Dit was het geval in een onderzoek van Cohen en Upton (2007). Zij onderzochten de strategieën die kandidaten inzetten bij twee typen meerkeuzevragen: de meer traditionele meerkeuzevraag met één correct antwoord en een innovatieve meerkeuzevraag waarbij meerdere alternatieven gekozen moeten worden met behulp van slepen en neerzetten. De meer innovatieve meerkeuzevraag was ontworpen om de academische taak van het vormen van een begrijpelijke en coherente representatie van een gehele tekst te simuleren. Het bleek dat het beantwoorden van deze vragen in grote mate gefaciliteerd werd door het feit dat ze altijd aan het einde van de toets kwamen. De kandidaten hadden de tekst dan al meerdere keren gelezen om de voorafgaande vragen te beantwoorden. Terwijl het doel van de toetsontwikkelaars was om academisch veeleisende items te construeren, was de realiteit dat de kandidaten de beoogde strategieën niet hoefden in te zetten om de vraag te beantwoorden. Dit resultaat benadrukt het belang van onderzoek naar strategiegebruik: het is namelijk niet altijd duidelijk welke processen doorlopen worden bij het beantwoorden van een vraag.

Cohen en Upton (2006) beschreven de strategieën die kandidaten gebruikten bij de verschillende leesvragen die ontworpen waren om kandidaten bekend te maken met de vernieuwde TOEFL. Zij concludeerden dat academische leesvaardigheid nodig was om de passages goed te begrijpen en de vragen te beantwoorden. De kandidaten gebruikten echter ook een verzameling van toetsafnamestrategieën, waarvan de toetsmanagementstrategieën het grootste gedeelte besloegen. De zes meest voorkomende toetsmanagementstrategieën waren het herlezen van de vraag, het parafraseren van de vraag, het lezen van een passage om hints naar het antwoord te vinden, het overwegen (of het uitstellen daarvan) van de alternatieven en het selecteren of uitsluiten van een alternatief op basis van woordenschat, structuur of betekenis.

2.2.3 Toetswijsheidstrategieën

Het toepassen van toetswijsheidstrategieën houdt in dat kennis van toetsformats en overige irrelevante informatie wordt ingezet om een antwoord te kunnen geven, zonder de benodigde taalkennis en -vaardigheden in te zetten (Cohen, 2014). Ze zijn onwenselijk, omdat ze een onevenredig voordeel geven aan kandidaten van verschillende vaardigheidsniveaus. Het is daarom een bron van invaliditeit: de constructvaliditeit van de toets komt in het geding. Er wordt dan namelijk niet alleen gemeten wat gemeten moet worden (hoe goed de luistervaardigheid is bijvoorbeeld), maar ook hoeveel kennis iemand heeft van de manier van toetsen. Studenten met een grote toetswijsheid kunnen dus bepaalde strategieën inzetten om op die manier tot het goede antwoord te komen. Radboud in'to Languages, het talencentrum van de Radboud Universiteit met een eigen toetsafdeling, bespreekt op hun website vijf toetswijsheidstrategieën 'geheimen' om elke meerkeuzetoets te halen (en hoe dit te voorkomen als docent). Een eerste 'geheim', of strategie, is het inzetten van algemene kennis om de vraag te beantwoorden. Studenten moeten op basis van puur en alleen voorkennis de vraag niet goed kunnen beantwoorden of foute antwoorden kunnen elimineren. Een docent moet daarom geen feitenvragen stellen, maar verwijzen naar de tekst. Ook moet de docent erop letten dat het antwoord op een vraag niet in een andere vraag gegeven wordt. Ten derde zijn docenten geneigd om het goede antwoord in het midden te plaatsen, dus bij twijfel kiezen studenten met een grote toetswijsheid voor antwoord B of C, afhankelijk van het aantal alternatieven. Om dit te voorkomen

wordt aangeraden om de antwoorden op alfabetische of chronologische volgorde te plaatsen. Andere strategieën zijn het kiezen van het langste antwoord en het kiezen van een genuanceerd antwoord. Douglas en Hegelheimer (2005) noemen nog het proces van eliminatie (het selecteren van een optie, ondanks dat het niet begrepen wordt, uit een vaag vermoeden dat de andere opties niet correct zijn) en het selecteren van een optie, omdat het een woord of zinsdeel uit een passage lijkt te hebben – mogelijk een sleutelwoord. Indien zulke strategieën toepasbaar zijn op een meerkeuzetoets, is er sprake van construct-irrelevante variantie. De verschillen in prestaties op een item worden veroorzaakt door kenmerken van de opgave die niet relevant zijn voor wat de opgave beoogt te meten (Bügel & Sanders, 1998). In dit geval zijn dat de vragen of antwoordmogelijkheden, die sturen richting het goede antwoord, waardoor de kans dat het goede antwoord wordt gekozen door te gokken groter is dan kansniveau.

Uit verschillende onderzoeken is gebleken dat het inzetten van dergelijke toetswijsheidstrategieën de kandidaat kan helpen bij het kiezen van het correcte antwoord. Allan (1992) ontwierp een toets van kennisvragen waarin bewust hints werden toegevoegd. De enige informatie die de kandidaten tot hun beschikking hadden, was de informatie in de vraag en in de antwoordopties. De proefpersonen konden met behulp van vier verschillende strategieën de juiste antwoorden geven. De items bevatten namelijk een van de volgende hints: a) een grammaticale aanwijzing, b) een associatie tussen een woord in de vraag en een van de alternatieven, c) soortgelijke alternatieven waardoor het overgebleven antwoord juist moest zijn of d) de benodigde informatie werd ergens anders in de toets gegeven. De gemiddelde resultaten op de toets (18.3) waren boven kansniveau (circa 8), wat suggereert dat de proefpersonen niet willekeurig hebben gegokt, maar gebruik hebben gemaakt van de hints. De hints in categorie a) en c) werden het meest opgepakt. Andere toetswijsheidstrategieën die Allan bespreekt, en hier nog niet eerder genoemd zijn, zijn e) de absurde optie: het elimineren van opties die niet correct kunnen zijn of die geen logische associatie met het sleutelwoord in de vraag hebben (zoals: 'bergen' in de vraag, 'Sahara' in een antwoord), f) optie-inclusie: het kiezen van een antwoord, omdat dat andere alternatieven omvat, waarvan de kandidaat weet dat ze juist zijn, g) de precieze optie: het kiezen van het alternatief dat meer precies wordt omschreven dan de andere, omdat dat het juiste antwoord kan zijn, h) kies geen van óf beide opties die de correctheid van elkaar impliceren en ten slotte i) kies geen van óf een van beide opties, waarvan een, indien correct, de incorrectheid van de andere optie impliceert.

Onderzoekers hebben daarnaast aangetoond dat sommige meerkeuzevragen in leestoetsen beantwoord kunnen worden zonder überhaupt de bijbehorende passage te hebben gelezen. Het aantal goede antwoorden dat gegeven werd zonder de passages van de *Scholastic Assessment Test* (SAT) te hebben gelezen, was boven kansniveau (20%), namelijk: 37,6% in het onderzoek van Katz, Lautenschlager, Blackburn en Harris (1990) en 35% in het onderzoek van Daneman en Hannon (2001). Uit dat laatste onderzoek bleek echter ook, niet geheel onverwacht, dat het lezen van de passages wel degelijk loont. De deelnemers van de toets scoorden het best (71%) wanneer zij de gehele passage vooraf lazen en delen herlazen om de vragen te beantwoorden en gemiddeld (63%) wanneer zij de vragen gebruikten om de antwoorden te zoeken, zonder de passage globaal te lezen.

2.3 Manieren om toetswijsheidstrategieën te voorkomen

Een manier om te corrigeren voor de gokkans is het toepassen van een gokcorrectie. De cesuur van een toets kan bepaald worden aan de hand van een berekening waarin de gokkans wordt meegenomen. Als er 40 meerkeuzevragen met vier antwoordmogelijkheden zijn, is er bij gokken zonder enige kennis 25% kans om een vraag goed te beantwoorden. Statistisch gezien kunnen dan dus 25% van de $40 = 10$ vragen sowieso goed worden beantwoord. Daarnaast dienen de studenten van de

overige 30 vragen er 16,5 goed te hebben voor een voldoende (bij een kennispercentage van 55%). In totaal moeten dus $10 + (55\% \text{ van } 30) = 26,5$ vragen goed worden beantwoord. De cesuur ligt in dit voorbeeld bij 27 goed beantwoorde vragen. Een andere manier om voor gokkans te corrigeren is het aftrekken van punten voor het geven van een fout antwoord. De kandidaten krijgen dan ook de antwoordoptie 'Weet ik niet', waarvoor geen punten worden afgetrokken. Er kan op deze manier echter nog steeds gebruik worden gemaakt van eventueel aanwezige hints, waardoor een beredeneerde gok alsnog aantrekkelijk is. Het is daarom beter om te voorkomen dat externe hints gebruikt kunnen worden.

Om toetswijsheid in meerkeuzetoetsen te voorkomen, zijn er globaal twee mogelijkheden. Ten eerste, kan de toetsontwikkelaar ervoor zorgen dat er geen externe hints in zijn toets voorkomen. Ten tweede, kan voor een andere toetsvorm gekozen worden, waar het risico op het voorkomen of kunnen gebruikmaken van externe hints kleiner is. De eerste oplossing kan bewerkstelligd worden door toetsontwikkelaars bewust te maken van de externe hints die kunnen voorkomen in meerkeuzetoetsen. Zij kunnen hier bij het ontwerpen van een toets alert op zijn. Richtlijnen voor het schrijven van meerkeuzevragen kunnen daarbij helpen. Haladyna, Downing en Rodriguez (2002) valideerden een lijst van 31 richtlijnen voor meerkeuzevragen door onderzoeken en boeken over toetsing in een educatieve setting te lezen. Er was een consensus over de meeste richtlijnen, zoals het onafhankelijk houden van vragen en het gelijk houden van de lengte van de alternatieven. Het gebruiken van richtlijnen en het lezen van relevante literatuur kan toetsontwikkelaars helpen om construct-irrelevante variantie te voorkomen. Daarnaast is het sterk aan te raden om bij *high stakes* meerkeuzetoetsen tijdens de pre-test te controleren op mogelijke externe hints. De tweede oplossing, het veranderen van het toetsformat, zou ervoor kunnen zorgen dat externe hints minder gemakkelijk in de toets voorkomen of gebruikt kunnen worden door de kandidaten. De keuze voor een bepaald toetsformat heeft namelijk invloed op het construct dat gemeten wordt en op de resultaten op de toets. Onderzoekers die naar het effect van toetsformat keken, ontdekten dat meerkeuzetoetsen hoger gescoord werden (Wolf, 1993) en minder goed discrimineerden (Tsagari, 1994) dan *constructed-response* toetsen. Uit een onderzoek van Currie en Chiramane (2010) bleek dat *constructed-response* vragen en meerkeuzevragen, ook al was de vraag hetzelfde, slechts in 26% van de gevallen hetzelfde beantwoord werden. De onderzoekers wijten dit aan het format van de meerkeuzetoets, die het meten van taalgerelateerde vaardigheden verstoort. Hoewel bij de *constructed-response* toets 91% van de antwoorden gebaseerd was op kennis (of het gebrek daaraan), werd 64% van de antwoorden in de meerkeuzetoets gebaseerd op *cued recall* (het herinneren van iets op basis van een woord of frase in het item), een toetsafnamestrategie of simpelweg gokken. Het gebruiken van *constructed-response* vragen vermindert het gebruik van toetswijsheidstrategieën dus wellicht, al heeft dit type vraag wel als nadeel dat bepaalde voordelen van meerkeuzevragen (het snel en betrouwbaar scoren) wegvallen. Ook de modus van de toets kan het strategiegebruik van kandidaten veranderen. Tayakoli en Samian (2014) vergeleken de attitude van de kandidaten ten aanzien van het gebruik van strategieën in een papieren toets met een internettoets. De onderzoekers merkten op dat de proefpersonen meer toetswijs waren in de pen-en-papiertoets. Ook viel op dat de proefpersonen in de internettoets niet zoveel voordeel haalden uit post-toetsstrategieën als in de papieren toets. Tayakoli en Samian wijten dit aan de aard van internettoetsen, waarbij de reeds beantwoorde vragen niet meer zichtbaar zijn. Enkele proefpersonen gaven aan dat zij in de papieren toets hulp kregen van andere items; deze hulp is in een internettoets lastiger te verkrijgen.

2.3.1 *Discrete-option multiple-choice*

Een ander type vraag dat toetswijsheid kan voorkomen is de *discrete-option multiple-choice* (DOMC) vraag. Dit format is veelbelovend. Het biedt een aantal grote voordelen. Allereerst kan beter gecontroleerd worden voor toetswijsheid: het is moeilijker om de antwoordmogelijkheden te vergelijken met elkaar. In het onderzoek van Papenberg, Willing en Musch (2017) werden hints in uiteenlopende kennisvragen gestopt. Er was een meerkeuzeconditie en een DOMC-conditie. De proefpersonen in de meerkeuzeconditie beantwoordden meer vragen juist (10.90) dan de proefpersonen in de DOMC-conditie (7.27), waarbij een score van 18 uit 24 als perfecte uitkomst met het gebruik van hints zou zijn. Niettemin konden ook in de DOMC-conditie hints gebruikt worden: het elimineert toetswijsheid dus niet volledig. Het in mindere mate kunnen toepassen van toetswijsheidstrategieën kan als bijkomend voordeel hebben dat organisaties in de toetsvoorbereidingsindustrie zich meer gaan richten op het onderwijzen van de daadwerkelijke vaardigheden in plaats van het aanleren van trucjes (Foster, 2016). Ook in het DOMC-format blijft een gokkans aanwezig, wanneer de kandidaat niet overtuigd is van het laatste alternatief, maar dat antwoord desondanks als correct aanwijst, omdat het de laatste mogelijkheid is. Om dit te voorkomen zou bijvoorbeeld gevarieerd kunnen worden met het aantal alternatieven, al zou dat mogelijk extra stress opleveren voor de kandidaat. Een andere optie zou zijn om de kandidaat alsnog een eerder antwoord te laten kiezen, maar daarvoor minder punten te geven indien het antwoord correct is. De kandidaten blijken echter niet vaak bij het laatste alternatief te geraken. In het onderzoek van Papenberg en collega's werden gemiddeld slechts 1.94 van de vier opties getoond. Bij Foster en Miller (2009) waren dit gemiddeld 2.64 en 2.79 van de vijf beschikbare alternatieven. Dit zorgt ervoor dat de toets sneller gemaakt kan worden én dat de veiligheid van de toets vergroot wordt. Foster en Miller zagen een afname van circa 10 procent van de tijdsduur. Bij Willing, Ostapczuk & Musch (2015), die DOMC-vragen in een medische toets onderzochten, was dit zelfs 23 tot 30 procent. Het is aannemelijk dat dit effect nog groter is wanneer de kandidaten meer gewend zijn aan het format. Ze hoeven dan geen uitgebreide instructies te lezen of oefenvragen te maken om bekend te worden met het format. Doordat de kandidaten niet alle alternatieven zien, is ook de kans op fraude en het creëren van incorrecte kennis kleiner. Het terugdringen van fraude zorgt ervoor dat vragen gemakkelijker hergebruikt kunnen worden. Dit verlaagt de kosten van toetsontwikkeling gerelateerd aan het vervangen van items (Foster, 2016). Ten slotte vermindert de simpelere aard en het verminderde aantal antwoordmogelijkheden de impact van leesvaardigheid op een item. Dit is met name handig bij het toetsen van groepen die hun leesvaardigheid nog niet volledig ontwikkeld hebben. De doelgroep van de huidige studie, tweedetaalleerders, valt daar ook onder. Foster (2016) heeft alle voordelen van DOMC-vragen in een overzicht gezet, verdeeld in verschillende categorieën en naar mate van impact. Voor uitgebreidere informatie over de voordelen van dit format verwijs ik daarom naar zijn werk.

Ook de nadelen van dit format worden beschreven in Foster (2016). Het eist op technisch gebied bijvoorbeeld meer dan een standaard meerkeuzevraag en kan daardoor minder makkelijk geïmplementeerd worden. Daarnaast zijn DOMC-vragen waarschijnlijk niet zo populair bij de kandidaten. Uit het onderzoek van Foster en Miller (2009) bleek dat circa 73% van de proefpersonen de voorkeur gaf aan het standaard meerkeuzeformat, 19% had geen voorkeur en slechts 9% gaf de voorkeur aan het DOMC-format. Het feit dat de scores in een DOMC-toets lager liggen dan in een standaard meerkeuzetoets kan hierbij meespelen (Kingston, Tiemann, Miller Jr., & Foster, 2012). Het is logisch dat de scores lager zijn, omdat de kandidaten hun manier van strategiegebruik hebben moeten aanpassen. Ze konden altijd terugvallen op compensatiestrategieën, maar nu worden ze verplicht om zich een nieuwe, mogelijk meer eisende manier van studeren aan te meten. Naarmate kandidaten meer wennen aan dit format, zal hun onzekerheid in dit format minder worden. Een andere uitdaging is de introductie van een nieuwe construct-irrelevante variabele, namelijk:

terughoudendheid wat betreft het accepteren van een antwoord (*acceptance reluctance*). Dit is een persoonlijke voorkeur om een latere antwoordmogelijkheid boven een vroege antwoordmogelijkheid te kiezen. In de dissertatie van Papenberg (2018) staat beschreven dat een hogere terughoudendheid tot hogere DOMC-toetscores leidde. Aangezien er nog weinig onderzoek naar DOMC-vragen is verricht, is het onduidelijk wat de omvang van deze variantie is. Papenberg (2018) stipt namelijk aan dat de omvang van individuele verschillen zou kunnen verminderen wanneer kandidaten gewend zijn aan het nieuwe toetsformat. De kandidaten in zijn onderzoek bleken namelijk beter te presteren in hun tweede DOMC-toets. Door het gebrek aan onderzoek is het tevens mogelijk dat meer nieuwe construct-irrelevante variabelen meespelen in een DOMC-toets. Een laatste kanttekening wordt geplaatst bij het feit dat kandidaten in het huidige DOMC-format naar de volgende vraag gaan wanneer zij een correct antwoord als incorrect aanwijzen. In dit specifieke geval kunnen de kandidaten afleiden dat dit alternatief het juiste was. Wanneer zij een incorrect antwoord als incorrect aanwijzen, krijgen zij immers wel het volgende alternatief te zien. Dit verlaagt ten eerste de toetsveiligheid, aangezien de kandidaten nu weten wat het juiste antwoord was en deze informatie zouden kunnen doorgeven. Daarnaast kan de kandidaat zich zorgen gaan maken wanneer hij weet dat hij een vraag fout beantwoord heeft. Foster en Miller besloten om deze reden een extra antwoordoptie toe te voegen. Deze optie werd getoond in de helft van de gevallen nadat de kandidaat een item eigenlijk voltooid had. Deze extra optie heeft echter geen toegevoegde waarde wanneer een kandidaat een vraag als correct aanmerkte, omdat de feedback van het doorgaan naar de volgende vraag in dat geval nihil is. De kandidaat weet namelijk niet of dat antwoord daadwerkelijk correct was. De extra antwoordoptie heeft wel een functie wanneer de kandidaat het correcte antwoord afwees. Het is echter gemakkelijker te implementeren om de kandidaten alternatieven te laten zien totdat zij een antwoord als correct aanduiden. Ook hoeven er dan geen extra afleiders verzonnen te worden. In de huidige studie is er daarom voor gekozen om kandidaten het volgende alternatief te laten zien wanneer zij het correcte antwoord afwijzen. Dit zorgt er waarschijnlijk voor dat de kandidaten meer alternatieven te zien krijgen, wat ertoe leidt dat de toetsafname iets langer zal duren. De verwachting is dat deze toename beperkt is, aangezien kandidaten een antwoord relatief snel als correct aanduiden.

Het onderzoek naar DOMC-vragen is tot op heden beperkt. De onderzoeken die er zijn, beslaan ook andere vakgebieden dan taal, zoals geneeskunde en wiskunde. Het is daardoor onduidelijk of de resultaten uit die onderzoeken gegeneraliseerd kunnen worden naar taaltoetsen. Verder onderzoek is daarom van belang. Er is tevens nog weinig bekend over de processen die kandidaten doorlopen bij het beantwoorden van een DOMC-vraag. Papenberg, Diederhoben en Musch (2019) kaartten aan dat de kennis over de beslissingen (inclusief gokstrategieën die daarbij ingezet worden) van kandidaten in sequentiële formats vergroot moet worden.

2.4 Strategieën in luistervaardigheid

Het belang van onderzoek naar toetsvaliditeit en luistervaardigheid wordt benadrukt in verschillende theoretische publicaties (Rossa, 2016). Desondanks blijven sterke empirische inspanningen uit. De beoordeling van luistervaardigheid is een van de minst begrepen en minst ontwikkelde onderdelen van taaltoetsing- en beoordeling (Buck, 2001). Empirisch onderzoek naar de beoordeling van luistervaardigheid is relatief schaars. Dat lijkt voort te komen uit het feit dat luistervaardigheid complex en moeilijk observeerbaar is (Rossa, 2016). Een uniforme definitie voor het construct luistervaardigheid in de tweede taal ontbreekt, omdat er zoveel verschillende cognitieve processen en individuele variabelen bij komen kijken (Winke & Lim, 2014). Onderzoek naar toetsvaliditeit ontmoedigt op zijn beurt onderzoekers, omdat het een breed, complex en onhandelbaar onderwerp is. Doordat toetsvaliditeit uit zoveel verschillende componenten bestaat, worden bepaalde aspecten

over het hoofd gezien. Onderzoek naar taaltoetsing kenmerkte zich door een voornamelijk productgeoriënteerde aanpak met een focus op de toetsresultaten (Bachman, 1990). Onderzoeken naar strategiegebruik zijn procesgeoriënteerd in plaats van productgeoriënteerd. Die procesgeoriënteerde aanpak is belangrijk, omdat een mismatch tussen de intenties van toetsontwikkelaars en de denkprocessen van kandidaten de validiteit van een toets in twijfel trekt (Cohen, 1984). Het antwoordformat in elke luistertoets is potentieel een bron van construct-irrelevante variantie, omdat luisterbegrip een intern cognitief proces is (Buck, 2001).

Het is daarom belangrijk dat er meer onderzoek naar toetsvaliditeit en luistertoetsen wordt verricht. Door strategiegebruik van luisteraars te onderzoeken, wordt meer bekend over de processen die succesvolle en minder succesvolle luisteraars doorlopen. Vandergrift heeft hier veel onderzoek naar gedaan. In zijn artikel met Tafaghodtari (2010) wordt een experiment beschreven waarin een experimentele groep (die bij luisteroefeningen extra aandacht moest besteden aan metacognitieve processen onderliggend aan het luisterproces) wordt vergeleken met een controlegroep (waar dat niet gebeurde). Opdrachten die metacognitieve processen trainden, waren bijvoorbeeld het voorspellen van informatie en woorden en het vergelijken van het begrip van de tekst met andere leerders. De experimentele groep bleek beter te presteren op een luistertoets dan de controlegroep. Dit verschil was significant wanneer de minder vaardige leerders van de twee groepen met elkaar vergeleken werden. De extra aandacht voor de metacognitieve instructie had het grootste effect op minder vaardige leerders. Zij hebben dit vaak niet leren doen in hun moedertaal, terwijl hoogopgeleide leerders deze kennis van hun eerste naar hun tweede taal kunnen overbrengen (Beare, 2000; Berman, 1994).

Teng (2013) en Rossa (2016) onderzochten beiden het strategiegebruik van tweedetaalleerders in een luistertoets. Uit Tengs studie bleek dat toetswijsheid de meest voorkomende categorie was. De meest frequent gebruikte strategie was het gebruiken van het proces van eliminatie. Rossa bestudeerde de cognitieve processen die uitgelokt werden bij Duitse leerlingen in een Engelse luistertoets. Hij wilde op die manier de toets valideren. Volgens Rossa was de toets valide wanneer 1) variaties in het toepassen van strategieën leidden tot variaties in de toetsresultaten en 2) de processen die doorlopen werden bij het correct beantwoorden van een vraag overeenkwamen met de processen die volgens het construct gebruikt moesten worden. De leerlingen uit zijn onderzoek dachten hardop na bij het beantwoorden van de meerkeuzevragen. Na afloop moesten de leerlingen een mondelinge samenvatting van de tekst geven en werden ze geïnterviewd om de gegevens uit het hardopdenkprotocol te verduidelijken en bevestigen. Drie items eliciteerden processen die afweken van de voorspelde processen. Sommige kandidaten herinnerde zich de benodigde informatie, maar kozen toch het verkeerde antwoord. Andere kandidaten herinnerde zich informatie die het correcte antwoord tegensprak, waarna ze toch het correcte antwoord kozen. De toets bleek in zijn algemeenheid valide te zijn, omdat twee kernfacetten van het luisterconstruct (informatie herinneren en inferenties maken) een belangrijke rol speelden in de resultaten, terwijl gokken dat niet deed. Deze manier van valideren, kijken naar (on)gewenst strategiegebruik, wordt vaker toegepast. Wu en Stone (2015) deden dit ook voor een leestoets, zoals reeds benoemd is in paragraaf 2.2.1. Ook Salehi (2011) valideerde op die manier een Engelse leestoets. Papenberg, Diedenhofen en Musch (2019) pasten dit toe op een Duitse leestoets in het DOMC-format, die even valide bleek te zijn als zijn equivalent in het standaardmeerkeuzeformat.

2.5 Mondelinge rapporten en vragenlijsten over strategiegebruik

De construct-irrelevante variantie veroorzaakt door een bepaald strategiegebruik van kandidaten kan alleen achterhaald worden door de denkprocessen van kandidaten bloot te leggen. De (cognitieve)

processen die kandidaten doorlopen tijdens het maken van een toets zijn namelijk niet direct zichtbaar. Om deze processen zichtbaar te maken, wordt getracht de gedachtegang van de kandidaten te achterhalen door middel van mondelinge rapporten en vragenlijsten.

Mondelinge rapporten kunnen op drie manieren benaderd worden (Cohen, 2006). De eerste benadering is *self-report*. In dat geval geeft de proefpersoon een beschrijving van wat hij doet, gekarakteriseerd door algemene beschrijvingen ('Als ik een meerkeuzevraag maak, scan ik eerst de tekst'). Een tweede benadering is *self-observation*. Dat houdt in dat de proefpersoon zijn gedrag inspecteert ('Wat ik net heb gedaan is het scannen van tekst'). Deze uiting kan zowel introspectief (indien binnen 20 seconden na de handeling) als retrospectief (indien langer dan 20 seconden na de handeling) zijn. De laatste vorm van een mondeling rapport is introspectief: *self-revelation*. Het gaat dan om een hardopdenkprotocol, waarbij de proefpersoon zijn gedachten hardop uitspreekt terwijl hij zijn aandacht op bepaalde informatie richt ('Ik denk dat ik de tekst eerst moet scannen'). Het soort data dat verkregen wordt met mondelinge rapporten kent een aantal nadelen. De kandidaten kunnen geneigd zijn om sociaal geaccepteerde antwoorden te geven (Cohen, 1994). Ook worden de resultaten gekleurd door de mate waarin de proefpersonen in staat zijn hun gedachten te verbaliseren. Daarnaast is het verbaliseren van de gedachtegang tijdens het lezen of luisteren van een tekst een cognitief overbelastende taak. Om die reden kunnen mondelinge rapporten het best na de fase van taakverwerking verzameld worden. De informatie die dan gerapporteerd kan worden is beperkt (Ericsson & Simon, 1993). Onze gedachten kunnen namelijk slechts voor een bepaalde tijd in het werkgeheugen gehouden worden. Bovendien zijn niet alle gedachten beschikbaar voor verwoording. Het verzamelen van goede mondelinge rapporten is dus geen gemakkelijke taak. Er is daarom behoefte aan een oplossing voor de problemen die daarbij komen kijken. Ericsson (2003) probeert met zijn voorstel richting een oplossing voor die problemen te werken. De beste manier om de problemen te beperken is volgens hem het werken met hardopdenkprotocollen (*self-revelation*). De taak wordt dan namelijk zo min mogelijk veranderd, waardoor de prestaties hetzelfde blijven (Ericsson & Simon, 1993). Wanneer een hardopdenkprotocol geen optie is, is het rapporteren van gedachten onmiddellijk na de taak (*self-observation*) het beste alternatief (Ericsson, 2003). De kans dat de kandidaat zijn gedachtestroom accuraat kan reproduceren is dan redelijk groot.

Enkele nadelen van mondelinge rapporten, zoals het geven van sociaal geaccepteerde antwoorden, blijven ook dan bestaan, maar worden zoveel mogelijk beperkt. Cohen (2006) beschrijft daarnaast dat er vaak te weinig relevante data zijn, wanneer de proefpersonen mogen antwoorden zoals zij willen. In verschillende onderzoeken naar toetsstrategieën werd daarom een enquête over de gebruikte strategieën afgenomen. Op deze manier kan specifiek aangeduid worden welke strategie de proefpersoon inzet om tot een antwoord te komen. Teng (2013) onderzocht het strategiegebruik van Taiwanese studenten in een Engelse luistertoets en gebruikte hiervoor een enquête met 42 Likert-schaalitems. Deze enquête was voornamelijk gebaseerd op de enquêtes van Taguchi (2001) en Cohen en Upton (2007). De 42 items werden verdeeld in verschillende categorieën en de proefpersonen gaven op een schaal van 1 (sterk oneens) tot 5 (sterk eens) aan hoe vaak zij gebruikmaakten van de strategieën. In de enquête van Salehi (2011) werd een ietwat andere vorm gebruikt: de kandidaten gaven door middel van een checklist per item aan welke strategie(ën) zij gebruikten. Door middel van een enquête over strategiegebruik worden nietszeggende toelichtingen voorkomen. Een nadeel van een enquête is dat de proefpersonen bewust worden van deze strategieën en ze kunnen gaan inzetten, terwijl ze dat anders misschien niet hadden gedaan. Dit verstoort hun eigen proces. Het beste is daarom om de vragenlijst pas na de toetsvragen te geven.

2.6 Vaardige versus minder vaardige kandidaten

Taalvaardigheidsniveau heeft invloed op het strategiegebruik van kandidaten. Taguchi (2001) onderzocht het strategiegebruik van Japanse studenten in een Engelse luistertoets. Door middel van een enquête met 42 Likert-schaalitems gaven de proefpersonen aan in welke mate zij gebruikmaakten van bepaalde taalleerderstrategieën. De vaardige en minder vaardige leerders verschilden in hun waargenomen gebruik van *top-down* strategieën. Dit resultaat wordt bevestigd in een studie van Teng (2013). De vaardige luisteraars bleken significant meer *top-down* en emotionele strategieën te gebruiken dan de minder vaardige luisteraars. Ook wanneer kandidaten met een hoge en lage taalvaardigheid dezelfde strategieën gebruiken, kunnen de resultaten van het gebruik verschillend zijn (Purpura, 1999). Cohen en Upton (2006) observeerden dat hun proefpersonen meer toetsmanagementstrategieën gebruikten dan toetswijsheidstrategieën. De onderzoekers linkten dit aan het hoge taalvaardigheidsniveau van hun proefpersonen. Op basis hiervan kan herleid worden dat kandidaten met een lager taalvaardigheidsniveau geneigd zijn om meer toetswijsheidstrategieën dan toetsmanagementstrategieën te gebruiken (Assiri, 2011). Leerders met een lagere vaardigheid blijken anders naar een tekst te kijken dan meer vaardige leerders. Zij focussen zich meer op kleine taalonderdelen, terwijl vaardige leerders zich meer focussen op een passage of tekst als geheel (Gordon, 1987). In het onderzoek van Taguchi gebruikten de vaardige luisteraars een groter aantal strategieën. Uit een onderzoek van Yoshida-Morise (1998) bleek echter juist dat de proefpersonen met een lagere taalvaardigheid meer en een grotere variëteit aan strategieën inzetten om te compenseren voor hun ontoereikende kennis van de taal. Uit de studie van Taguchi bleek daarnaast dat er bij de minder vaardige leerders een groter gat zat tussen hun perceptie van compensatiestrategieën en hun daadwerkelijke toepassing van die strategieën. Vaardige leerders lijken zich dus meer bewust te zijn van de strategieën die ze toepassen.

2.7 Staatsexamens NT2

De proefpersonen in de huidige studie hebben ongeveer niveau B1, gelijk aan het benodigde niveau voor Programma I van de Staatsexamens NT2. De Staatsexamens NT2 zijn examens over de Nederlandse taal en maatschappij. Mensen die inburgeringsplichtig zijn, kunnen na het halen van een deze examens een Nederlands paspoort of permanente verblijfsvergunning aanvragen. De Staatsexamens bestaan uit de vier taalvaardigheidsexamens luisteren, spreken, lezen en schrijven. Het behalen van het Staatsexamen programma I is geeft toegang tot een mbo-opleiding. Programma II is op niveau B2 en geeft toegang tot een hbo- of wo-opleiding. In het jaarverslag van het jaar 2019 staat beschreven dat in dat jaar 12.187 kandidaten meededen aan de Staatsexamens NT2, van wie 6.395 aan programma I en 5.792 aan programma II (College voor Toetsen en Examens, 2019b). Er is ook een Inburgeringsexamen op niveau A2. Dit examen kan ook toegang tot de Nederlandse maatschappij verschaffen. Het bestaat, naast de taalvaardigheidsexamens, uit een examen Kennis van de Nederlandse Maatschappij, Oriëntatie op de Nederlandse Arbeidsmarkt en een participatieverklaring.

Eind 2018 kwamen complete beschrijvingen en uitwerkingen van opdrachten van alle onderdelen en beide programma's op straat te liggen. De afname van de Staatsexamens kwam daarom stil te liggen. In januari 2019 werden de afnames voor reeds ingeschreven kandidaten hervat (College voor Toetsen en Examens, 2019a). Vanaf eind maart 2019 konden nieuwe kandidaten zich pas weer inschrijven. Het nieuwe examensysteem dat toen is ingevoerd bestaat steeds uit nieuwe opgaven, zodat het onthouden en doorgeven van opgaven zinloos is (College voor Toetsen en Examens, 2018). Dat betekent wel dat kandidaten niet meer elke week examen kunnen doen, maar slechts enkele keren per jaar. Zoals eerder benoemd, verkleint een DOMC-toetsformat de kans op fraude, omdat de

kandidaten niet alle antwoordmogelijkheden te zien krijgen. Het is dus zeker relevant om te onderzoeken of dit format toegepast kan worden in de Staatsexamens NT2.

Het huidige stelsel voldoet op meer punten niet (Rijksoverheid, 2020). Per 1 januari 2022 wordt daarom een nieuw inburgeringsstelsel ingevoerd. Dit nieuwe stelsel moet ervoor zorgen dat inburgeraars sneller inburgeren en gestimuleerd worden om het hoogst mogelijke taalniveau te behalen. De norm voor het taalniveau wordt derhalve verhoogd. Het Inburgeringsexamen zal dan niet meer op niveau A2 zijn, maar op niveau B1. Mocht dit niet haalbaar zijn voor een cursist, is er wel de mogelijkheid om een bepaald examen op niveau A2 te maken. De meeste inburgeraars zullen deze route (de B1-route) volgen. De taallessen worden gecombineerd met (vrijwilligers)werk of een stage. De andere routes zijn de Onderwijsroute en de Z-route. De Onderwijsroute is gericht op jonge inburgeraars tot 28 jaar. Zij krijgen intensieve taallessen en volgen tegelijkertijd vakken als rekenen, Engels en leervaardigheden en studieloopbaanbegeleiding. Zij ronden dan in gemiddeld anderhalf jaar de inburgering af en stromen in het vervolgonderwijs in. Tot slot is er de Z-route. Voor een kleine groep inburgeraars zullen de andere twee routes niet haalbaar zijn. Zij gaan in het nieuwe stelsel meer tijd besteden aan het leren van de taal, zelfredzaamheid en participatie in de samenleving. Gemeenten krijgen een belangrijkere rol in dit nieuwe stelsel. Gemeenten staan het dichtst bij de inburgeraars, waardoor zij maatwerk kunnen leveren. Het huidige systeem laat ruimte voor kwaadwillenden om te frauderen met hun lening. Doordat de gemeenten straks de taallessen betalen, wordt dit probleem opgelost.

2.7.1 Luistervaardigheidsexamen

Het luisterexamen binnen het Staatsexamen programma I en II bestaat uit drie delen met daarin vijf of meer teksten (College voor Toetsen en Examens, z.d.-a). Dit kunnen ook video's zijn. Er mag telkens maar één keer geluisterd worden. Het examen duurt 90 minuten (College voor Toetsen en Examens, z.d.-b). De examenkandidaten maken het examen op een computer met een koptelefoon. De teksten worden gesproken door een of meer sprekers. Zij vertellen over een onderwerp of bespreken een onderwerp met elkaar. De onderwerpen beslaan situaties op het werk, tijdens een opleiding of in het dagelijks leven. Het examen bestaat uit ongeveer 40 vragen. Elke tekst heeft minimaal vijf en maximaal tien vragen (College voor Toetsen en Examens, z.d.-a). Dit zijn vragen die globale of selectieve luistervaardigheid toetsen. De verschillen tussen programma I en II zijn dat bij programma II moeilijkere woorden worden gebruikt, moeilijkere onderwerpen voorkomen en beschouwende teksten belangrijker zijn.

Hoofdstuk 3: Methode

3.1 Onderzoeksontwerp

Het onderzoek is uitgevoerd als een experiment met online vragenlijst. Het gebruikte onderzoeksdesign is een 2x1 tussen-personendesign met als onafhankelijke variabele toetsformat (DOMC versus standaard meerkeuze) en als voornaamste afhankelijke variabele het strategiegebruik van de kandidaat. Ook de afhankelijke variabelen tijdsduur en gemiddelde score werden onderzocht.

3.2 Materiaal

Het experiment werd voorafgegaan door een korte schriftelijke introductie en toestemmingsverklaring. In de introductie werd verteld dat onderzocht werd op welke manieren NT2-leerders een toets maken. Ook werden de verschillende onderdelen van het experiment toegelicht. Het experiment bestond globaal uit drie delen: achtergrondvragen, luistervragen en vragen over hoe de toets gemaakt en ervaren was. In de toestemmingsverklaring werd informatie gegeven over de vrijwilligheid van het onderzoek en de manier waarop de gegevens bewaard worden. Wanneer een proefpersoon akkoord ging met de toestemmingsverklaring, werd hij doorgestuurd naar het experiment. Indien een proefpersoon niet akkoord ging, werd de vragenlijst beëindigd.

Allereerst moesten de proefpersonen een aantal achtergrondvragen beantwoorden. Deze vragen gingen over het geslacht, de leeftijd, de herkomst, het Nederlandse taalniveau en het opleidingsniveau van de proefpersoon. Ook werd hier gevraagd of de proefpersonen gehoorproblemen hadden. Als een proefpersoon aangaf gehoorproblemen te hebben, werd de vragenlijst beëindigd.

Vervolgens werd de instructie van het experiment gegeven. Een deel van de instructie was in beide condities hetzelfde. In beide instructies werd bijvoorbeeld aangegeven dat de proefpersoon elk luisterfragment slechts één keer kon beluisteren. De instructie van het experiment was korter in de versie van de standaard meerkeuzetoets, omdat dit format geen verdere uitleg behoeft. De instructie in de versie van het DOMC-format was uitgebreider. In de DOMC-versie werd namelijk uitgelegd hoe een DOMC-vraag werkt. Na de instructie volgden drie oefenvragen. Winke en Lim (2014) benadrukken het belang van vertrouwd raken met het toetsformat. Zij onderzochten het effect van toetsvoorbereiding op toetsscores. Hieruit bleek dat uitgebreide voorbereiding niet nodig is, maar dat enige vorm van toetsvoorbereiding wel helpt om bekend te worden met de vraagtypen. Vooral bij onbekende vraagtypen, zoals in de huidige studie het geval was, is het vertrouwd raken van belang. Om deze reden kregen de proefpersonen drie vragen om te oefenen met het format. Om de twee versies zoveel mogelijk gelijk te houden, kregen ook de proefpersonen in de meerkeuzeversie drie oefenvragen. Na de oefenvragen kregen de proefpersonen te zien dat de echte toets nu zou beginnen. Het daadwerkelijke experiment bestond uit negentien vragen uit het Oefenexamen Luisteren uit 2017. Deze vragen, met antwoorden, staan in Bijlage I. In totaal waren er dus 22 vragen. Bij drie van die vragen kreeg de proefpersoon afbeeldingen te zien en moest hij aangeven welk plaatje het goede antwoord weergaf. De vragen waren afkomstig van drie verschillende teksten. De drie verschillende typen teksten toetsten alle drie globale luistervaardigheid. De tekst 'Zwembadmedewerker' was een persuasieve tekst, 'Verkoopmedewerker' een informatieve tekst en 'Een eigen klusbedrijf' een instructieve tekst. De eerste drie vragen van 'Zwembadmedewerker' werden gebruikt als oefenvragen. De mondelinge instructies uit het oefenexamen werden ook gebruikt, weliswaar ingekort. De voorbeeldopgave werd bijvoorbeeld verwijderd, omdat de proefpersonen al oefenvragen hadden gemaakt. De proefpersonen kregen de instructie om de vraag goed door te lezen. Vervolgens moesten zij op de startknop klikken om het audiofragment te horen. De oorspronkelijke geluidsfragmenten waren ingekort, omdat het genoemde vraagnummer niet overeenkwam met het vraagnummer in het experiment. Ook werd de stilte in het fragment, waarin een kandidaat normaal gesproken moet antwoorden en de volgende vraag moet lezen, verwijderd. Hierna kregen de proefpersonen de vraag en een of meerdere antwoordopties te zien (ofwel de eerste DOMC-vraag met één antwoordoptie, ofwel de eerste meerkeuzevraag met drie antwoordopties). Dit wijkt af van de procedure bij het Staatsexamen. Daar ziet de kandidaat de alternatieven al tijdens het luisteren. In dit onderzoek is ervoor gekozen dat niet te doen, omdat de proefpersonen in de DOMC-conditie zich teveel zouden kunnen gaan richten op het eerste alternatief, dat als enige zichtbaar is op dat moment. Om de twee versies zoveel mogelijk gelijk te houden, werden ook in de meerkeuzeconditie de alternatieven pas na het luisteren getoond. Na het beantwoorden van elke vraag moesten de proefpersonen in een tekstvak

toelichten waarom zij voor hun antwoord gekozen hadden. De gedachtegang achter hun keuze moest hiermee duidelijk worden gemaakt.

Na de negentien toetsvragen moesten de proefpersonen een vragenlijst over hun strategiegebruik invullen. De lijst van strategieën die werd aangehouden was gebaseerd op die van Winke en Lim (2014), die zich weer baseerden op Vandergrift (1997) voor de luisterstrategieën en Cohen en Upton (2007) voor de toetswijsheid- en toetsmanagementstrategieën.¹ Door middel van een factoranalyse beperkten Winke en Lim de lijst tot vijftien luisterstrategieën en zestien toetswijsheid- en toetsmanagementstrategieën. Deze lijst was echter niet geheel toereikend voor de huidige studie. Allereerst werd in de huidige studie wél een onderscheid gemaakt tussen toetswijsheid- en toetsmanagementstrategieën, iets wat Winke en Lim niet deden. Op basis van de definities van de categorieën werden de toetswijsheid- en toetsmanagementstrategieën in de lijst van Winke en Lim ingedeeld in een van de twee categorieën. Ten tweede moest de omvang van de lijst beperkt blijven om de tijdsduur van het experiment te beperken. Om deze reden is ervoor gekozen vijf strategieën per categorie in de lijst op te nemen. Er waren drie categorieën: taalleerderstrategieën, toetsmanagementstrategieën en toetswijsheidstrategieën. In totaal werden er dus vijftien strategieën bevestigd. De lijst met strategieën en de bijbehorende categorie is te vinden in Bijlage II. Een groot deel van die strategieën komt voor in de lijst van Winke en Lim. De keuze om strategieën wel of niet op te nemen is gebaseerd op enkele overwegingen. Er is bijvoorbeeld op gelet dat de strategieën begrijpelijk zijn voor de proefpersonen en dat er geen overlappende strategieën voorkomen. Daarnaast is ervoor gezorgd dat de strategieën die in de literatuur veel benoemd worden, maar niet in de lijst van Winke en Lim voorkwamen, opgenomen zijn in de vragenlijst. De proefpersonen gaven op een vijfpunts-Likertschaal (nauwelijks – weinig – soms – vaak – heel vaak) aan hoe vaak zij die strategieën hadden ingezet in de toets die ze zojuist gemaakt hadden. Zij mochten daarnaast zelf ook strategieën inbrengen, om te voorkomen dat de lijst te beperkt was (Salehi, 2011). Ten slotte kregen de proefpersonen een aantal stellingen over de toets. Op wederom een vijfpunts-Likertschaal (helemaal oneens – oneens – neutraal – eens – helemaal eens) gaven zij aan in hoeverre zij het eens waren met de stellingen. In de DOMC- en meerkeuzeversie werden drie dezelfde stellingen gegeven, maar in de DOMC-toets werd nog een extra stelling toegevoegd. Deze stelling ('Ik koos zo snel mogelijk een antwoord') betrof de terughoudendheid wat betreft het accepteren van een antwoord. Door het toevoegen van deze stelling kan een vergelijking gemaakt worden met de resultaten van Papenberg (2018).

3.2.1 Hardopdenkprotocol

Zes proefpersonen (5 vrouwen, 1 man) namen deel aan een hardopdenkprotocol. Zij maakten de luistertoets terwijl zij videobelden met de onderzoeker. Het experiment dat zij kregen, verschilde op enkele punten van het experiment zoals hierboven beschreven. Ten eerste was de toestemmingsverklaring uitgebreider, omdat ook toestemming voor video- en geluidsopnamen gegeven moest worden. Daarnaast kregen zij voorafgaand aan de toets een instructie over het hardopdenken tijdens het maken van de toets. Ten slotte hoefden zij de keuze voor een antwoord niet schriftelijk toe te lichten, omdat ze dit mondeling zouden doen.

¹ Winke en Lim verwijzen helaas niet naar dit artikel van Vandergrift in de literatuurlijst. In twee van zijn artikelen uit 1997 die ik gevonden heb, wordt niet gesproken over een vragenlijst. Het is daarom onduidelijk waar de vragenlijst precies op gebaseerd is. Mogelijk bedoelden zij de MALQ-vragenlijst van Vandergrift, Goh, Mareschal en Tafagodhari (2006), maar die vragenlijst heeft 21 vragen en niet 28 zoals Winke en Lim benoemen. Het blijft dus onduidelijk waar zij zich precies op baseerden.

3.2.2 Implementatie Qualtrics

Het experiment werd ingevoerd in Qualtrics, software voor enquêtes. Qualtrics heeft veel verschillende functies. Standaard kunnen audiobestanden in Qualtrics opnieuw afgespeeld, gepauzeerd en doorgespoeld worden. Voor deze studie was het van belang dat die mogelijkheden er niet waren, omdat dat tijdens het echte examen ook niet kan. Met behulp van een code in JavaScript kon ervoor gezorgd worden dat de proefpersonen het fragment eenmalig konden starten, maar niet konden pauzeren of doorspoelen. Qualtrics kent ook een 'Display Logic'-functie. Hiermee kan aangegeven worden wanneer een vraag wel of niet getoond wordt. Deze functie was nodig om de DOMC-vragen te implementeren. Een tweede alternatief, dat in Qualtrics als aparte vraag werd ingevoerd, kon op deze manier getoond worden wanneer de proefpersoon bij de eerste antwoordmogelijkheid 'Fout' had geantwoord, maar niet wanneer hij 'Goed' had geantwoord. Ten slotte werden extra timers toegevoegd aan het experiment. Dit is een functie van Qualtrics waardoor de afnemer kan zien hoe lang een proefpersoon op een bepaalde pagina is geweest. Hierdoor kon de tijdsduur in de twee versies niet alleen in zijn geheel vergeleken worden, maar ook per vraag of alleen voor de instructie. Toen het eerste hardopdenkprotocol werd afgenomen, viel op dat de betreffende proefpersoon de instructie niet goed doorlas en meteen verder klikte. Om die reden is daarna in Qualtrics binnen de Timing-functie toegevoegd dat de proefpersoon pas verder kon klikken na 20 seconden (in de MK-versie) of 30 seconden (in de DOMC-versie). Dit secondenaantal was gebaseerd op de tijd die de onderzoeker nodig had om de instructie goed te lezen. Aangezien de onderzoeker een moedertaalspreker van het Nederlands is en zichzelf als snelle lezer beschouwt, werd aangenomen dat een proefpersoon op B1-niveau de instructie nooit sneller dan de onderzoeker zorgvuldig zou kunnen doornemen. De twee versies (zonder of met hardopdenkprotocol) kunnen bekeken en gemaakt worden door middel van de links in de voetnoot².

3.3 Proefpersonen

De proefpersonen werden op verschillende manieren benaderd met een wervingsbericht. Zij werden veelal benaderd via hun docent. Docenten van veel verschillende taalscholen hebben een door de onderzoeker opgesteld wervingsbericht ontvangen. Velen hebben aangegeven het door te sturen of te bespreken met hun cursisten. Daarnaast hebben enkele docenten door middel van een reactie op een oproep in de Facebook-groep 'NT2-docenten' laten weten dat zij dit wervingsbericht door hebben gestuurd. Ten slotte werd het wervingsbericht geplaatst in de Facebook-groep 'Staatsexamens NT2' met NT2-leerders die zich voorbereiden op de Staatsexamens NT2 om de doelgroep direct te bereiken. In dit wervingsbericht werd verteld waar het onderzoek over ging, namelijk: de manier waarop NT2-leerders een luistertoets maken. In het wervingsbericht werd ook verteld dat de standaard meerkeuzetoets met een nieuwere meerkeuzetoets vergeleken zou worden. De proefpersonen kregen echter geen informatie over het toetsformat dat zij zouden gaan krijgen. Uit onderzoek van Finley en Benjamin (2012) bleek dat kandidaten hun voorbereiding aanpasten op basis van het toetsformat dat ze verwachtten (*cued recall* versus *free recall*). De kandidaten die het format kregen dat zij verwachtten, presteerden beter dan de kandidaten die een onverwacht format kregen. Om deze reden is ervoor gekozen de proefpersonen pas tijdens het experiment te vertellen welk format zij kregen.

In totaal waren er 87 aanmeldingen via de mail. Het experiment werd uiteindelijk begonnen door 69 tweedetaalleerders van het Nederlands. Niet alle proefpersonen hebben het experiment voltooid. Zij

² DOMC (hardopdenkprotocol): https://radboudletteren.eu.qualtrics.com/jfe/form/SV_3jFZloDC2vEBBuB
DOMC: https://radboudletteren.eu.qualtrics.com/jfe/form/SV_29RU7fA9ykmV1IN
MK (hardopdenkprotocol): https://radboudletteren.eu.qualtrics.com/jfe/form/SV_eDqsj7jcFZojd09
MK: https://radboudletteren.eu.qualtrics.com/jfe/form/SV_3fnGUBjBZY70vfD

worden daarom niet meegenomen in de resultaten. In de DOMC-conditie vielen om die reden 10 proefpersonen af en in de MK-conditie 6 proefpersonen. De uitval was dus iets groter bij de DOMC-toets. Daarnaast waren er 8 proefpersonen die om andere redenen afvielen. Zij gaven aan gehoorproblemen te hebben of toch niet meer mee te willen doen. Hierna bleven er 45 proefpersonen over (34 vrouwen, 9 mannen). 22 van hen maakten de DOMC-toets en 23 maakten de standaard meerkeuzetoets.

De proefpersonen waren tussen de 24 en 79 jaar oud ($M = 34.87$, $SD = 9.54$). Het was een heterogene groep wat betreft land van herkomst. De proefpersonen kwamen uit 29 verschillende landen. China was het meest vertegenwoordigd (4 proefpersonen), gevolgd door Syrië en Hongarije (3 proefpersonen). Het taalniveau van de proefpersonen was ongeveer B1-niveau. Proefpersonen konden deelnemen als zij een niveau tussen halverwege B1 en halverwege B2 hadden (niveau A2,5 tot niveau B1,5). Het aantal jaren dat zij al Nederlands leerden, lag tussen het half jaar en de 12 jaar ($M = 2.67$, $SD = 2.32$). De proefpersonen waren over het algemeen hoogopgeleid. De meesten hadden een hbo-, bachelor- of masterdiploma. Een klein aantal proefpersonen gaf aan over een mbo-diploma te beschikken. Een enkeling gaf nog aan geen diploma te hebben en een andere proefpersoon gaf aan alleen een middelbareschooldiploma te hebben.

3.4 Procedure

De proefpersonen die geïnteresseerd waren om deel te nemen, lieten dit weten door middel van een e-mail naar de onderzoeker. Hierna kregen zij een link naar het onderzoek doorgestuurd (willekeurig de DOMC- of de meerkeuzeversie, maar beide versies werden even vaak uitgezet) en eventueel tijdsloten voor het hardopdenkprotocol. De proefpersonen kregen de opdracht om de toets in één zitting en in een rustige ruimte met koptelefoon te maken. Dat eerste was van belang, omdat de duur van de toetsafname ook onderzocht werd. Dit is ook het meest realistisch, aangezien een toets meestal zonder pauzes wordt gemaakt. De proefpersonen maakten de vragenlijst individueel online met behulp van een computer. De vragenlijst is in principe compatibel voor mobiele telefoon, maar binnenkomende berichten zouden daar voor afleiding kunnen zorgen. Ook zou dit minder realistisch zijn, aangezien een luisterexamen normaliter op een computer wordt gemaakt.

3.4.1 Hardopdenkprotocol

De proefpersonen in het hardopdenkprotocol werd verteld dat zij het eerste deel van het onderzoek (de toestemmingsverklaring en de achtergrondvragen) individueel konden maken. In de Qualtrics-vragenlijst werd aangegeven wanneer de proefpersoon op de via e-mail verkregen Zoom-link moest klikken. Hoewel in de introductie vermeld stond welk deel met videobellen zou zijn én deze boodschap later in het experiment dikgedrukt stond, vonden niet alle proefpersonen het duidelijk wanneer zij moesten beginnen met videobellen. Om deze reden begon het videobellen soms aan het begin van het experiment, soms aan het begin van de oefenvragen en soms aan het begin van de toets. Ook maakten twee proefpersonen die eigenlijk zouden meedoen aan het hardopdenkprotocol de toets alleen, omdat zij in de veronderstelling waren dat het videobellen na de toets plaatsvond. Later is de boodschap om op de Zoom-link te klikken ook nog onderstreept. Een reden voor de verwarring is mogelijk de hoeveelheid aan instructie en informatie die de proefpersonen tot zich moesten nemen. Tijdens het maken van de toets deelden de proefpersonen in het hardopdenkprotocol hun scherm met de onderzoeker en dachten hardop na bij het beantwoorden van de vragen. Bij de oefenvragen konden zij hiermee oefenen. De onderzoeker keek en luisterde mee. Als een proefpersoon voor een lange periode stil was of een vraag beantwoordde zonder toelichting te geven, vroeg de onderzoeker naar

de gedachten en de redenering van de proefpersoon. De mondelinge rapporten hadden ofwel de vorm van een *self-revelation* (hardopdenkprotocol), ofwel van een introspectieve *self-observation* waarbij de proefpersoon zijn gedachten en gedrag beschreef binnen 20 seconden nadat het gebeurde. Er werd een opname gemaakt van het videogesprek. Deze opname werd later getranscribeerd door de onderzoeker.

3.5 Data-analyse

De data werden kwantitatief en kwalitatief geanalyseerd.

Kwantitatieve analyse

Het meest voorkomende strategiegebruik per conditie werd bepaald door de gemiddelde waarden van de drie categorieën te vergelijken. Dit werd gedaan met behulp van *paired samples t-tests*. Om eventuele verschillen in strategiegebruik tussen de twee condities te duiden, werd een *independent t-test* uitgevoerd. Het DOMC-format en het meerkeuzeformat werden ook vergeleken op de gemiddelde tijdsduur en de gemiddelde score. Hiervoor werden *independent samples t-tests* en *Mann-Whitney U tests* gebruikt. Daarnaast werden de reacties op de stellingen in beide condities vergeleken met behulp van *Mann-Whitney U tests*. Ook werd gekeken of er een verband was tussen terughoudendheid en behaalde score door de Spearman's correlatie te berekenen. Ten slotte werd berekend wat het gemiddeld aantal alternatieven was dat getoond werd.

Kwalitatieve analyse

Er werden ook kwalitatieve analyses van het strategiegebruik uitgevoerd. Dit werd gedaan aan de hand van de toelichtingen bij de vragen en de hardopdenkprotocollen. Er werd gekeken welke strategieën opvallend, veelvoorkomend of construct-(ir)relevant waren en welke overeenkomsten en verschillen er tussen de twee condities waren.

Hoofdstuk 4: Resultaten

4.1. Analyse vragenlijst over strategiegebruik

In totaal werden er vijftien strategieën onderzocht, vijf per categorie. De betrouwbaarheid van de drie schalen werd onderzocht door de Cronbachs alpha te berekenen. De betrouwbaarheid van de schaal taalleerderstrategieën was .585. Deze waarde laat zien dat de interne consistentie van de schaal slecht is. De betrouwbaarheid van deze schaal werd iets, maar niet veel, groter als de stelling 'Ik bedacht voorafgaand aan het luisteren voor welk doel ik ging luisteren' verwijderd werd, namelijk .608. De Cronbachs alpha van de schaal toetsmanagementstrategieën was .636 en werd .681 als de stelling 'Ik probeerde de vraag anders te formuleren' verwijderd zou worden. De interne consistentie van deze schaal is daarom twijfelachtig. Ten slotte was de Cronbachs alpha van de schaal toetswijsheidstrategieën: .518. Ook deze waarde laat zien dat de interne consistentie van de schaal slecht is. De betrouwbaarheid van de schaal kon maximaal verhoogd worden tot .524 als 'Ik lette op hints in de vraag, in de antwoorden en in andere vragen' verwijderd zou worden.

4.1.2 Strategiegebruik (kwantitatief)

De centrale vraag in deze scriptie is: **Hoe verschillen de strategieën die kandidaten inzetten in een DOMC-toets van de strategieën die kandidaten inzetten in een standaard meerkeuzetoets?**

De tweede deelvraag betrof de strategieën die kandidaten inzetten tijdens een DOMC-toets. Zij gaven aan vooral toetsmanagementstrategieën te gebruiken ($M = 3.56$, $SD = .77$), gevolgd door toetswijsheidstrategieën ($M = 3.48$, $SD = .69$). Zij gaven aan taalleerderstrategieën het minst te gebruiken ($M = 3.16$, $SD = .78$). Er werd een *paired samples t-test* uitgevoerd om te bepalen of deze verschillen significant zijn. Deze test wordt normaliter toegepast op variabelen op interval- of rationiveau. De strategieën zijn op ordinaal niveau, maar de vijfpunts-Likertschaal benadert wel een continue schaal. Daarom werd hier toch voor een *paired samples t-test* gekozen. De data waren normaal verdeeld. Omdat er drie significantietoetsen werden gedaan, is de Bonferroni-correctie toegepast. De p-waarde die nodig was voor significantie was daarom $.05/3 = .0167$. De verschillen tussen de waarden voor de drie categorieën bleken niet significant te zijn.

Om een duidelijker beeld te krijgen van de meest en minst gebruikte strategieën zijn de gemiddelden van de losse strategieën vergeleken. In Tabel 1 staat beschreven hoe vaak proefpersonen in beide condities aangaven een strategie te gebruiken. Uit de posities van de strategieën is af te leiden welke strategieën het meest en minst werden ingezet tijdens de DOMC-toets. Van de vier strategieën die het meest ingezet werden, zijn er drie toetsmanagementstrategieën. De overige strategie is een toetswijsheidstrategie. De strategie die het minst werd gebruikt was tevens een toetsmanagementstrategie. De strategieën die daarna het minst ingezet werden, zijn taalleerderstrategieën. De proefpersonen mochten zelf ook strategieën inbrengen, om te voorkomen dat de lijst te beperkt was. Eén proefpersoon in de DOMC-conditie gaf aan nog een andere strategie in te zetten, namelijk: het letten op werkwoorden.

Tabel 1: De waardes van alle strategieën in de DOMC- en MK-conditie en de positie binnen de eigen conditie op basis van gemiddelde.

Strategie	Categorie	Conditie	M	SD	Positie
Ik probeerde tijdens het luisteren onbelangrijke informatie te negeren.	TL	DOMC	3.41	1.22	8
		MK	2.96	1.11	11
Ik bedacht van tevoren welke woorden ik misschien zou gaan horen.	TL	DOMC	3.23	1.11	12
		MK	2.78	1.24	12
Ik gebruikte kennis van de talen die ik spreek.	TL	DOMC	2.73	1.67	14
		MK	2.26	.96	14
Ik bedacht tijdens het luisteren hoe de tekst verder zou gaan.	TL	DOMC	2.95	1.29	13
		MK	2.65	1.34	13
Ik bedacht voorafgaand aan het luisteren voor welk doel ik ging luisteren.	TL	DOMC	3.50	1.23	7
		MK	3.57	1.08	3
Ik voorspelde wat het antwoord zou zijn en keek of mijn antwoord erbij stond.	TM	DOMC	3.55	1.22	6
		MK	3.13	1.42	8
Ik controleerde of mijn antwoord logisch was.	TM	DOMC	3.91	1.23	2
		MK	3.83	1.11	1
Ik koos een antwoord op basis van de algemene betekenis van de tekst.	TM	DOMC	4.05	.95	1
		MK	3.74	.75	2
Ik koos een antwoord, omdat ik wist dat de andere antwoorden fout waren.	TM	DOMC	3.82	1.01	3
		MK	3.52	.85	4
Ik probeerde de vraag anders te formuleren.	TM	DOMC	2.50	1.10	15
		MK	2.26	.96	14
Ik koos een antwoord, omdat het een woord bevatte dat ook in de tekst of in de vraag genoemd werd.	TW	DOMC	3.32	1.25	10
		MK	3.30	1.02	6

Ik lette op hints in de vraag, in de antwoorden en in andere vragen.	TW	DOMC	3.82	1.05	3
		MK	3.13	1.06	8
Ik koos een antwoord, ook al wist ik het niet zeker.	TW	DOMC	3.32	1.25	10
		MK	3.09	1.08	10
Ik vergeleek de antwoorden met elkaar en probeerde zo antwoorden te elimineren.	TW	DOMC	3.36	1.33	9
		MK	3.52	.79	4
Ik gebruikte algemene kennis om de vraag te beantwoorden.	TW	DOMC	3.59	1.22	5
		MK	3.22	1.09	7

De derde deelvraag ging over de strategieën die kandidaten inzetten tijdens een standaard meerkeuzetoets. De kandidaten in deze conditie gaven aan voornamelijk toetsmanagement- ($M = 3.29$, $SD = .58$) en toetswijsheidstrategieën ($M = 3.25$, $SD = .62$) in te zetten. Net als de kandidaten in de DOMC-conditie, werden taalleerderstrategieën het minst vaak ingezet tijdens de toets ($M = 2.84$, $SD = .72$). Er werd wederom een *paired samples t-test* uitgevoerd om te onderzoeken of deze verschillen significant zijn. De data waren normaal verdeeld. De waarden voor taalleerder- en toetsmanagementstrategieën bleken significant te verschillen ($t(22) = -2.504$, $p = .020$), eveneens de waarden voor taalleerder- en toetswijsheidstrategieën ($t(22) = -2.527$, $p = .019$) wanneer een p-waarde van .05 werd gehanteerd. Na de Bonferroni-correctie, waardoor de p-waarde voor significantie .0167 werd, bleken deze verschillen niet langer significant. De Bonferroni-correctie wordt algemeen beschouwd als een strenge correctie en zorgt ervoor dat het onderscheidingsvermogen (te) sterk daalt. Het is dus wellicht wel het geval dat taalleerderstrategieën minder ingezet worden dan toetsmanagement- en toetswijsheidstrategieën. Ook in de conditie van de standaard meerkeuzetoets werden de gemiddelden van de losse strategieën bekeken. Deze gegevens staan in Tabel 1. De top drie strategieën waren twee toetsmanagementstrategieën en een taalleerderstrategie. De drie strategieën die het minst gebruikt werden, waren twee taalleerderstrategieën en één toetsmanagementstrategie. Eén proefpersoon gaf aan een strategie te gebruiken die niet in de vragenlijst voorkwam, namelijk: het zoeken naar synoniemen.

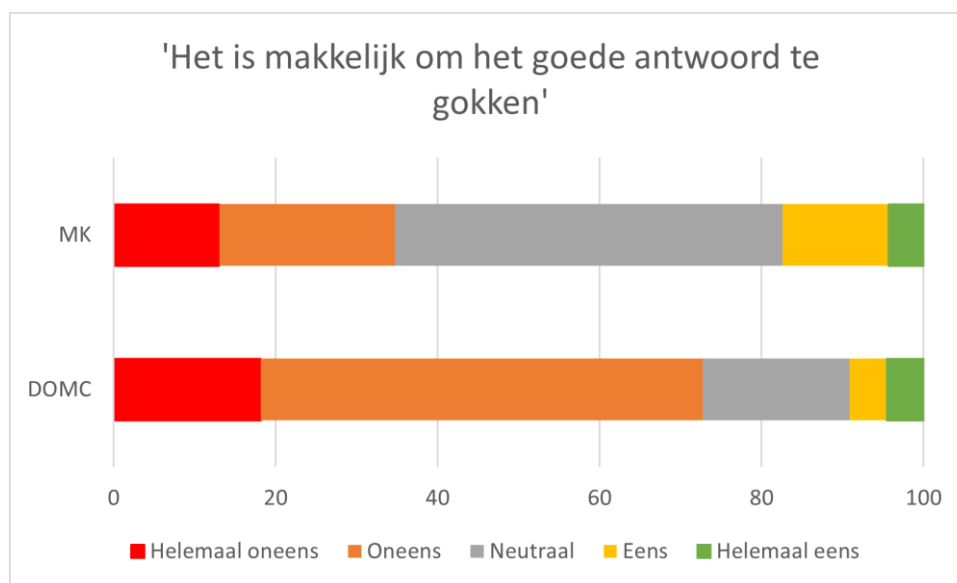
De vierde deelvraag betrof de verschillen tussen het strategiegebruik van de kandidaten in de DOMC-toets en het strategiegebruik van de kandidaten in de standaard meerkeuzetoets. De kandidaten in de DOMC-conditie gaven aan alle categorieën vaker in te zetten dan de kandidaten in de MK-conditie. Om te onderzoeken of deze waardes tussen de condities significant verschilden, werd een *independent t-test* uitgevoerd. Aan de assumptie van homogeniteit van varianties was voldaan. Daarnaast waren de data normaal verdeeld. De verschillen tussen de condities bleken niet significant te zijn. Er was geen significant verschil in de mate waarin taalleerderstrategieën ingezet werden ($t(43) = 1.433$, $p = .159$). Hetzelfde geldt voor de mate waarin toetsmanagementstrategieën ingezet werden ($t(43) = 1.322$, $p = .193$) en de mate waarin toetswijsheidstrategieën ingezet werden ($t(43) = 1.178$, $p = .245$). Ter overzicht staan deze gegevens in Tabel 2.

Tabel 2: De waardes van de verschillende categorieën in de DOMC- en MK-conditie.

Categorie	Conditie	M	SD	Toetsing
Taalleerderstrategieën (TL)	DOMC	3.16	.78	$t(43) = 1.433$, $p = .159$
	MK	2.84	.72	
Toetsmanagementstrategieën (TM)	DOMC	3.56	.77	$t(43) = 1.322$, $p = .193$
	MK	3.29	.58	
Toetswijsheidstrategieën (TW)	DOMC	3.48	.69	$t(43) = 1.178$, $p = .245$.
	MK	3.25	.62	

Dankzij Tabel 1 kan de volgorde waarin de losse strategieën ingezet worden, gemakkelijk vergeleken worden voor de twee condities. Opvallend is dat er veel overeenkomsten zijn. De twee meest gebruikte strategieën zijn namelijk dezelfde, maar omgekeerd in volgorde. Daarnaast is de volledige top 3 minst gebruikte strategieën hetzelfde. De strategieën die qua positionering het verst uit elkaar liggen in de twee condities, zijn: 'Ik lette op hints in de vraag, in de antwoorden en in andere vragen' (DOMC plek 3, MK plek 8) en 'Ik vergeleek de antwoorden met elkaar en probeerde zo antwoorden te elimineren' (DOMC plek 9, MK plek 4). Die eerste strategie is ook de strategie waarbij de gemiddelden van de twee condities het verst uit elkaar lagen. Dit verschil was bij de laatstgenoemde strategie maar klein.

In beide condities moesten de proefpersonen hun mening over de volgende stelling geven: 'Het is makkelijk om het goede antwoord te gokken.' De proefpersonen in beide condities bleken het over het algemeen oneens te zijn met deze stelling (DOMC $M = 2.23$, $SD = .97$; MK $M = 2.74$, $SD = 1.01$). Om te onderzoeken of hun meningen over deze stelling significant verschilden, werd een *Mann-Whitney U test* uitgevoerd. De data waren namelijk niet normaal verdeeld. De reacties op deze stelling verschilden significant ($U = 169$, $z = -2.005$, $p = .043$) tussen de proefpersonen in de DOMC-conditie en de proefpersonen in de MK-conditie. Dit bleek een klein negatief effect te zijn ($r = -.30$). Grafiek 1 maakt duidelijk dat in de DOMC-conditie vaak 'Oneens' als antwoord werd gegeven, terwijl in de MK-conditie voornamelijk 'Neutraal' werd gekozen. Het lijkt er dus op dat de proefpersonen in de DOMC-conditie meer moeite hebben om het goede antwoord te gokken.



Grafiek 1: Antwoorden op de stelling 'Het is makkelijk om het goede antwoord te gokken' in de twee condities (in percentages).

4.1.3 Strategiegebruik (kwalitatief)

Naast kwantitatieve analyses van het strategiegebruik zijn er ook kwalitatieve analyses uitgevoerd. De toelichtingen van de proefpersonen bij hun gekozen antwoord (zowel schriftelijk als mondeling in het hardopdenkprotocol) werden geclassificeerd als construct-relevant of -irrelevant (indien dit duidelijk werd door de toelichting). De meest opvallende en meest voorkomende toelichtingen worden in deze paragraaf besproken. Over het algemeen bleek uit deze analyse dat de proefpersonen de luisterprocessen doorliepen die de ontwerpers beoogd hadden en dat de antwoorden dus construct-relevant gedrag weerspiegelden. Dit gold voor beide groepen. De reden dat proefpersonen het correcte antwoord kozen, was veruit in de meeste gevallen dat ze het antwoord in de tekst hadden gehoord. Ze citeerden dan bijvoorbeeld een zin uit de luistertekst waarin duidelijk werd wat het juiste

antwoord moest zijn. Bij Vraag 1 'Parttime werken betekent dat je niet de hele week werkt. Waarom wil het zwembad graag mensen die parttime werken?' was het goede antwoord 'Omdat ze dan meer mensen tegelijk kunnen inzetten als dat nodig was.' De proefpersonen gaven als toelichting voor hun correcte antwoord onder andere het volgende:

PP9³: 'Want zij kunnen met parttime meedwerekres een better rooster maken.'

PP10: 'Zij kunnen meer mensen inroosteren.'

PP19: 'Tekst zegt dat als het druk is, is beter om vier parttimer te hebben dan een fulltimer.'

Sommige proefpersonen kwamen tot het juiste antwoord door middel van eliminatie van de andere antwoordmogelijkheden. Dit wordt in de huidige studie geclassificeerd als een toetsmanagementstrategie. Het kwam voornamelijk voor in de standaard meerkeuzetoets. Dit is in overeenstemming met de relatief hoge notering van deze strategie in de MK-conditie ten opzichte van de DOMC-conditie. Twee proefpersonen in de DOMC-conditie gaven aan dat het niet kunnen vergelijken van de antwoorden vooral een probleem vormde bij de vragen waar het juiste plaatje gekozen moest worden. Proefpersonen die eliminatie gebruikten om tot het correcte antwoord te komen, gaven als argument voor hun keuze:

PP23: 'want de andrer antwoorden zijn fout' (Vraag 2)

PP27: 'Omdat de andere zinnen zeker niet juist zijn, want ik heb gehoord dat er zijn weinig vervelende jongeren zijn en ik heb niet gehoord dat hij dat moeilijk vindt.' (Vraag 2)

PP35: 'A en B horen niet bij de kantoor klussen.' (Vraag 15)

Uit de toelichtingen viel ook af te leiden dat sommige proefpersonen bij het kiezen van een antwoord veel focus legden op sleutelwoorden uit de vraag of de antwoorden in combinatie met wat zij (niet) hebben gehoord in de tekst. Dit bleek bijvoorbeeld uit toelichtingen van proefpersonen 3 en 18.

PP18: 'Het woord 'taken' werd vaak gezegd.' (Vraag 4)

PP18: 'Ik heb het woord 'grote' niet gehoord.' (Vraag 5)

PP18: 'Ik heb het woord 'gestolen' gehoord' (Vraag 11)

PP3: 'doors and schilderen, i hier main words and make a conclusion' (Vraag 18)

PP3: 'key word collega' (Vraag 19)

In de meeste gevallen leidde dit tot het goede antwoord. De focus op sleutelwoorden leidt echter niet altijd tot het goede antwoord. Dit kan komen doordat het sleutelwoord weliswaar in de tekst voorkomt, maar de context anders is dan in het antwoord (de afleider). Het is ook mogelijk dat een sleutelwoord niet genoemd is in de tekst, maar een synoniem of omschrijving van het woord wel.

De reden dat proefpersonen een incorrect antwoord kozen, was vaak omdat ze de tekst niet goed hadden begrepen. Ze citeerden dan een deel van een zin uit de luistertekst waarin naar een van de afleiders werd verwezen, ze interpreteerden de tekst verkeerd of ze gaven aan het antwoord niet te weten. Wanneer een proefpersoon de tekst niet goed heeft begrepen, wil je dat hij het verkeerde antwoord kiest. Dit pleit voor de validiteit van de toets.

PP19: 'In de tekst zegt dat jij kan nog een opleiding te hebben en niet alleen werken dus er zijn heel veel keuzes' (Vraag 5)

PP4: 'jij kan samenwerken met een collega.' (Vraag 8)

PP5: 'Je mag wat afstand houden en zeggen wat je er van vindt, maar Maaïke zegt niks over moet eerlijk te zijn.' (Vraag 9)

³ Proefpersonen 1 tot en met 22 zaten in de DOMC-conditie, proefpersonen 23 tot en met 45 in de MK-conditie.

PP21: 'Ja ik hoor dit. Paar grote kasten.' (Vraag 17)

Er was uit de toelichtingen van de proefpersonen echter ook construct-irrelevante variantie af te leiden. In dat geval koos een proefpersoon het correcte antwoord voor een incorrecte reden of koos een proefpersoon een incorrect antwoord, hoewel hij de zinsnede citeerde die bij het juiste antwoord hoorde. Een proefpersoon kon om verschillende incorrecte redenen het juiste antwoord kiezen. Wanneer een proefpersoon aangaf het antwoord niet te weten, maar toch het correcte antwoord koos, had hij waarschijnlijk goed gegokt. Het lijkt op basis van de toelichtingen niet zo te zijn dat in een van de condities meer of meer goed wordt gegokt dan in de andere. Sommige proefpersonen gaven zelf aan het antwoord te hebben gegokt. Proefpersoon 31 gaf dit zelfs driemaal aan.

PP31: 'het was een gok' (Vraag 1)

PP10: 'Ik geeft niet aandacht aan de tweede vraag, dus ik raad de antwoord.' (Vraag 5)

PP31: 'allen gok' (Vraag 6 en 10)

PP5: 'Hij vertelde best veel. Ik kan de helft niet onthouden. "Grote materiaal investeren zijn zonder van je geld. " Dus ik heb het antwoord gegokt.' (Vraag 17)

Daarnaast kwam het 'Het is logisch'-argument veel voor. Proefpersonen gaven bijvoorbeeld aan:

PP34: 'Het is logisch'. (Vraag 2)

PP30: 'De opties zijn niet duidelijk voor me. Ik heb dat gekozen omdat ik het zou doen.' (Vraag 10)

PP44: 'Omdat misschien is de beste idee. Om te vragen 'Hé?' Misschien weet je als dit product heeft een korting. Misschien jij wist dit niet.' (Vraag 10)

PP13: 'Het is logisch.' (Vraag 12 en 16)

Opvallend is dat dit argument vaak werd gebruikt wanneer een fout antwoord werd gegeven. Het antwoord dat de proefpersoon het meest logisch vond, was dus vaak niet het juiste antwoord. Dit pleit voor de validiteit van deze toets. Op basis van alleen algemene kennis of logisch denken wordt over het algemeen niet het correcte antwoord gekozen. Bij Vraag 3 moesten de proefpersonen het juiste plaatje kiezen bij de vraag 'Welke oplossing geeft Jochem zodat je niet zo veel hoeft te schreeuwen?'. Bij deze vraag werd in beide condities vaak uitgegaan van eigen algemene kennis. Het juiste antwoord was het omhooghouden van een plankje, maar meerdere personen kozen voor de fluit of de microfoon.

PP32: 'je kunt met de fluit makkelijk "schreeuwen"'

PP13: 'Ik begrijp niet alles maar de foto is logisch'

PP19: 'Ja omdat ik zie hier een figuur met een foon. Dus ik denk dat is makkelijk. Dat is een beetje logisch. Ik heb de tekst niet zo begrijpt maar ja ik zie .. logisch.'

In sommige gevallen koos de proefpersoon een incorrect antwoord, hoewel hij in zijn toelichting verwees naar de woorden of zinnen die nodig waren om tot het correcte antwoord te komen. Het is mogelijk dat de proefpersoon de zinsnede niet goed begreep of interpreteerde. In dat geval is er geen sprake van construct-irrelevante variantie. Het geven van de benodigde informatie, hoewel het verkeerde antwoord gekozen werd, gebeurde voornamelijk in de DOMC-conditie. De onderstaande toelichtingen zijn van proefpersonen die een incorrecte antwoordoptie kozen, maar laten blijken dat zij wel de woorden of zinnen hebben gehoord die nodig zijn om de vraag juist te beantwoorden.

PP44: 'Ik denk daarmee heb je vaak te maken. Omdat het meevalt weet je. Hij heeft gezegd dat is ja is gewoon een uurtje.' (Vraag 2)

PP3: 'Omdat hij praat dat het is better vier medewerkers te hebben om hetzelfde tijd en dat is niet mogelijk met voeltijd werkers.' (Vraag 3)

PP6: 'er is een praktijdsbegeleider beschikbaar wat kennis heeft over dit soort bedrijf wat kan leiding geeft' (Vraag 6)

PP18: 'Ik heb gehoord dat Moud de woorden 'klanten' en 'familie' zei.' (Vraag 19)

In de DOMC-toets konden de proefpersonen geen antwoord kiezen dat zij al hadden afgewezen. Enkele proefpersonen in de DOMC-conditie aan dat zij dachten dat een eerder antwoord het juiste was. Soms waren zij daarin correct, zoals in de onderstaande voorbeelden.

PP8: 'ik denk misschien tweede optie was goed maar ik dacht dat misschien ik wil krijg een optie in relatie met rooster' (Vraag 1)

PP3: 'I think that the first answer was better' (Vraag 16)

In het onderstaande voorbeeld had de proefpersoon het correcte antwoord gekozen, maar gaf aan liever het eerste antwoord te hebben gekozen. Dit draagt bij aan construct-irrelevante variantie, aangezien de proefpersoon de vraag juist heeft beantwoord terwijl hij denkt dat zijn antwoord onjuist was.

PP2: 'Ik denk dat ik de verkeerde foto heb gekozen, maar ik zie niet alle opties meteen en daarom is het een beetje moeilijk om nu na het zien van alle foto's het juiste antwoord te weten, de eerste foto is het goede antwoord.' (Vraag 15)

Zoals benoemd in paragraaf 2.3.1, blijft in een DOMC-toets een gokkans aanwezig, wanneer de kandidaat niet overtuigd is van het laatste alternatief, maar dat antwoord desondanks als correct aanwijst, omdat het de laatste mogelijkheid is. Dit zorgt voor construct-irrelevante variantie. In de huidige studie kozen veel proefpersonen in de DOMC-conditie bij het derde alternatief alsnog 'Fout', als zij dachten dat dit niet het juiste antwoord was. Dit zal waarschijnlijk niet het geval zijn wanneer het een *high stakes* toets is. Een proefpersoon koos het derde alternatief, wat het juiste antwoord was, om de volgende reden:

PP20: 'Ik heb ook niet gehoord. Maar ik denk ja dat is derde dus moet goed zijn.' (Vraag 15)

Wat tijdens het analyseren van de toelichtingen al opviel, maar wat nog duidelijker werd tijdens het analyseren van de hardopdenkprotocollen, is dat proefpersonen een distinctieve eigen stijl lijken te hebben wat betreft het maken van toetsen. Proefpersonen 3 en 18 focusten op sleutelwoorden. Proefpersoon 31 gaf relatief vaak aan het antwoord te hebben gegokt. Hoewel er dus overeenkomsten zijn tussen de strategieën die proefpersonen inzetten, lijken proefpersonen een voorkeur te hebben voor een of meerdere strategieën. Door middel van uitingen van twee proefpersonen in het hardopdenkprotocol (beide in de MK-conditie) wordt dit geïllustreerd. Proefpersoon 42 had een duidelijk idee van hoe een correct antwoord er meestal uitziet en baseerde haar antwoorden regelmatig op algemene kennis. Enkele uitspraken van Proefpersoon 42 waren:

'Dus ik gok nummer twee. Ja want 'iedere' is heel sterk.' (Vraag 5)

'Ik denk dat in *reality* klopt ook niet. In mijn ervaring dat is ook altijd ze vragen naar jou hoe jij denk en dan ze praten alleen met jou eens.' (Vraag 9)

'Ik vind alle de eerste en twee is een beetje raar. *While* jij kunt niet altijd naar jouw chef vragen: 'Kun jij even komen?'' (Vraag 10)

'Ik gok nummer C, plaatje C. In mijn ervaring allebei B en C kan zijn in mijn kantoor. Dus ik weet niet wat klopt.' (Vraag 15)

'Oké ik gok plaatje C. Geen goede argument. Alleen ik denk dat nummer B is wat ik gehoord heb, maar misschien dat is te gemakkelijk. Misschien heb ik iets anders niet goed gehoord.' (Vraag 18)

Proefpersoon 44 gebruikte vaak de strategie van het elimineren van foute antwoorden. Dit deed ze ofwel terwijl ze de antwoordopties voor de eerste keer las, ofwel door middel van het herlezen en vergelijken van de antwoordmogelijkheden. Enkele uitspraken van Proefpersoon 44 waren:

'Omdat het werk veel te zwaar is, nee' (Vraag 1)

'Nee de moeilijk niet' (Vraag 2)

'Nee ik heb niet gehoord dat ze zegt dat de jurk haar mooie staat' (Vraag 9)

'Hmm oké. De eerste niet.' (Vraag 17)

4.2 Tijdsduur

De proefpersonen werd verteld dat zij de toets in één zitting moesten maken. Uit de data bleek echter dat niet alle proefpersonen dit hadden gedaan. De starttijd en eindtijd van de proefpersonen die langer dan 10.000 seconden (circa twee uur en drie kwartier) over het onderzoek hadden gedaan, werden gecontroleerd. Daaruit leek het aannemelijk dat zij de toets niet in een zitting hadden gemaakt, maar verspreid over een dag of meerdere dagen. Deze proefpersonen (3 in de DOMC-conditie, 4 in de MK-conditie) werden daarom niet meegenomen bij het vergelijken van de tijdsduur van het gehele onderzoek. Daarnaast was er in de MK-conditie een uitbijter van 8984 seconden waarvan het onduidelijk was of de toets in een zitting was gemaakt. Die is daarom ook niet meegenomen in de berekening. Ten slotte werden de proefpersonen in het hardopdenkprotocol niet meegenomen in de vergelijking van de tijdsduur, omdat zij vermoedelijk sneller waren geweest als zij de toets alleen hadden gemaakt. Hun data zijn daarom niet goed te vergelijken met de proefpersonen die de toets individueel maakten. Om de tijdsduur te vergeleken werd een *independent samples t-test* uitgevoerd. Aan de assumptie van homogeniteit van varianties was voldaan. Ook waren de data normaal verdeeld. De proefpersonen in de DOMC-conditie ($M = 4943.87$, $SD = 1616.60$) deden er gemiddeld langer over om het onderzoek af te ronden dan de proefpersonen in de MK-conditie ($M = 4385.00$, $SD = 1021.34$). Dit verschil bleek echter niet significant ($t(27) = 1.104$, $p = .280$).

De tijdsduur die hier werd vergeleken is de gehele tijdsduur van het onderzoek. De gemeten tijdsduur representeert daarom niet alleen de duur van de toets, maar ook de duur van het beantwoorden van de overige vragen. Om die reden werden van tevoren timers toegevoegd. Hierdoor kon gemeten worden hoe lang de proefpersonen deden over het beantwoorden van de toetsvragen en het lezen van de instructie.

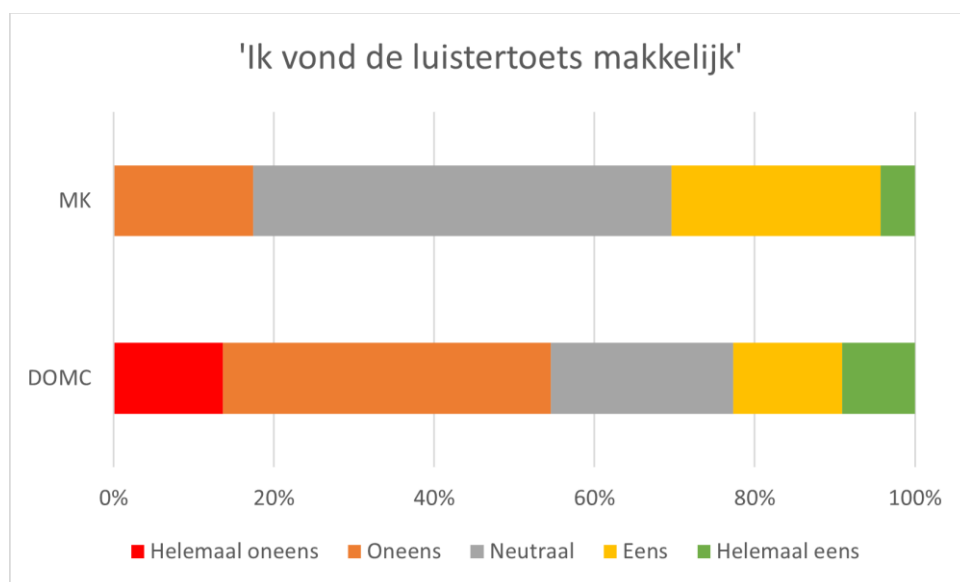
Voordat de duur van het lezen van de instructie vergeleken werd, werden drie uitbijters verwijderd (één in de DOMC-conditie, twee in de MK-conditie). De proefpersonen in de DOMC-conditie deden gemiddeld langer ($M = 128.91$, $SD = 117.6$) over het lezen van de instructie dan de proefpersonen in de MK-conditie ($M = 45.85$, $SD = 18.60$). Omdat de data in de DOMC-conditie niet normaal verdeeld waren, werd een *Mann-Whitney U test* gedaan. Hieruit bleek dat het verschil tussen de twee condities significant is ($U = 65$, $z = -2.738$, $p = .005$) en dat dit een klein negatief effect was ($r = -.47$). De tijd die het duurde om de instructie te lezen, verschilde binnen de DOMC-conditie sterk. Er waren bijvoorbeeld drie proefpersonen die circa 36 seconden deden over het lezen van de instructie, maar er waren ook drie proefpersonen die tien keer zo lang (ongeveer 360 seconden) deden over het lezen van de instructie.

Alle toetsvragen werden getimed. Het gaat hier niet om het luisterfragment, maar puur om de vragen en de antwoordopties. In de DOMC-conditie werd een uitbijter verwijderd. Zij bleek over een vraag 20 minuten te hebben gedaan. Waarschijnlijk heeft zij toen pauze genomen, waardoor de daadwerkelijke tijdsduur niet berekend kan worden. Daarnaast werden er twee uitbijters in de MK-conditie verwijderd. De proefpersonen die de standaard meerkeuzetoets maakten, deden langer over de

toetsvragen ($M = 622.85$, $SD = 201.39$) dan de proefpersonen die de DOMC-toets maakten ($M = 540.83$, $SD = 337.48$). Er werd een *independent t-test* uitgevoerd om te kijken of dit verschil significant was. De data waren normaal verdeeld en aan de assumptie van homogeniteit van varianties was voldaan. Het verschil in tijdsduur bleek niet significant te zijn ($t(32) = -.861$, $p = .396$).

4.3 Gemiddelde score

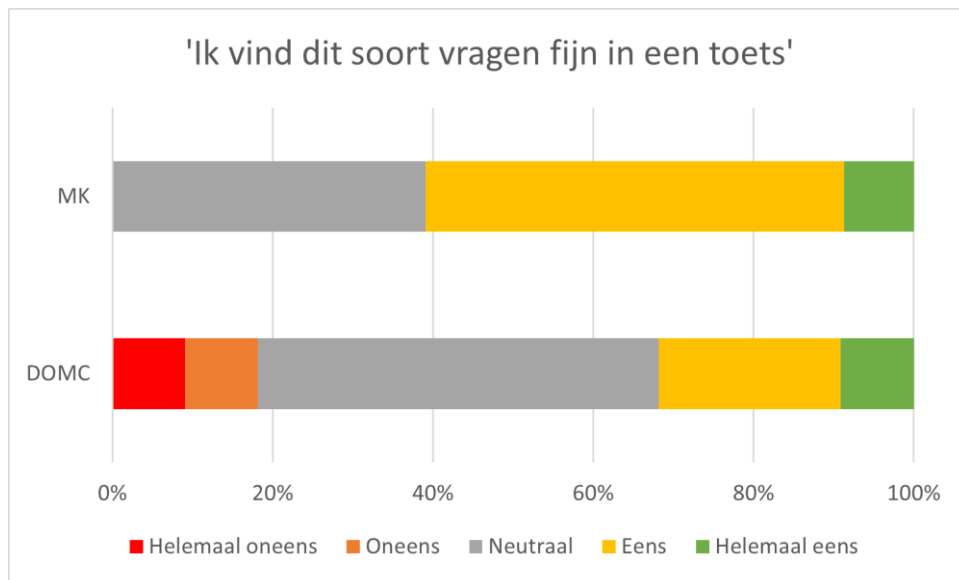
Ook werd onderzocht of de gemiddelde score op de toets verschilde tussen de kandidaten die een DOMC-toets maakten en de kandidaten die een standaard meerkeuzetoets maakten. Er waren 19 toetsvragen en voor elk juist antwoord werd één punt toegekend. Daaruit volgt dat de maximale score 19 punten was. De gemiddelde score behaald op de DOMC-toets was 10.9 ($SD = 2.94$). De gemiddelde score behaald op de standaard meerkeuzetoets was 15.3 ($SD = 2.07$). Om te toetsen of dit verschil significant was, werd een *independent samples t-test* uitgevoerd. Aan de assumptie van homogeniteit van varianties was voldaan. Daarnaast was de gemiddelde score in beide groepen normaal verdeeld. Uit de *independent samples t-test* bleek dat het verschil in gemiddelde score tussen de twee groepen significant was, $t(43) = -5.757$, $p = .000$, en dat dit een groot negatief effect was ($r = -.86$). De toets lijkt dus moeilijker te zijn geweest voor de proefpersonen in de DOMC-conditie. Om te onderzoeken of de proefpersonen dit zelf ook zo ervoeren, werd hun mening over de stelling 'Ik vond de luistertoets makkelijk' geanalyseerd. De proefpersonen in de DOMC-conditie waren het gemiddeld vaker oneens met deze stelling ($M = 2.64$, $SD = 1.18$) dan de proefpersonen in de MK-conditie ($M = 3.17$, $SD = 0.78$). Omdat de data niet normaal verdeeld waren, werd met behulp van de *Mann-Whitney U test* berekend of dit verschil significant was. Dat bleek niet het geval te zijn ($U = 170$, $z = -1.971$, $p = .051$)⁴. Grafiek 2 maakt duidelijk hoe de reacties op deze stelling verdeeld waren. De proefpersonen in de MK-conditie waren veelal neutraal. De proefpersonen in de DOMC-conditie waren het relatief vaak (helemaal) oneens met de stelling.



Grafiek 2: Antwoorden op de stelling 'Ik vond de luistertoets makkelijk' in de twee condities (in percentages).

⁴ De benadering via de z-waarde duidt op een significant verschil, hoewel de exacte berekening laat zien dat er geen significant verschil is.

In het onderzoek kwam ook de stelling 'Ik vind dit soort vragen fijn in een toets' voor. Naar verwachting zou de toets waarop lager gescoord werd (de DOMC-toets) als minder fijn beoordeeld worden dan de toets waarop hoger gescoord werd (de MK-toets). Dit bleek ook het geval te zijn, al gaven de proefpersonen in de DOMC-conditie gemiddeld genomen niet aan het oneens te zijn met deze stelling ($M = 3.14, SD = 1.04$). De proefpersonen in de MK-conditie waren het veelal eens met deze stelling ($M = 3.70, SD = .64$). Ook hier bleken de data niet normaal verdeeld te zijn, waarna de *Mann-Whitney U test* werd berekend. Hieruit volgde dat de gemiddelde reactie op deze stelling tussen de twee groepen significant verschillend was ($U = 168,5, z = -2.072, p = .038$) met een klein negatief effect ($r = -.31$). In Grafiek 3 is te zien hoe de reacties op de stelling verdeeld waren. In de standaard meerkeuzeconditie was er geen enkele proefpersoon die aangaf dit soort vragen niet fijn te vinden. De meesten vonden dit soort vragen juist fijn. In de DOMC-conditie stonden de meeste proefpersonen neutraal tegenover dat type vraag.



Grafiek 3: Antwoorden op de stelling 'Ik vind dit soort vragen fijn in een toets' in de twee condities (in percentages).

4.4 Terughoudendheid

Uit een eerder onderzoek bleek dat een hogere terughoudendheid wat betreft het accepteren van een antwoord als correct leidde tot een hogere score op de toets. In de huidige studie werd in de DOMC-conditie de stelling 'Ik koos zo snel mogelijk een antwoord' toegevoegd. De proefpersonen gaven aan in hoeverre zij het eens waren met die stelling (helemaal oneens – oneens – neutraal – eens – helemaal eens). De reactie van één proefpersoon werd niet meegenomen in de analyse. Zij gaf namelijk aan niet zo snel mogelijk een antwoord te kiezen, hoewel zij bijna elke keer antwoord A als goed aanwees. Om te kijken of er een relatie was tussen het antwoord op die stelling en de uiteindelijke score werd de Spearmans correlatie berekend. Er bleek geen correlatie te zijn ($r_s(21) = .347, p = .123$).

De proefpersonen gaven relatief vaak aan niet zo snel mogelijk een antwoord te kiezen (42,9%, ten opzichte van 15,9% die dat wel deed). Het gemiddelde aantal getoonde alternatieven in de DOMC-conditie was 1.72 ($SD = 0.25$) van de 3.

Hoofdstuk 5: Discussie en conclusie

5.1 Bespreking resultaten

De centrale vraag in deze studie, is: **Hoe verschillen de strategieën die kandidaten inzetten in een DOMC-toets van de strategieën die kandidaten inzetten in een standaard meerkeuzetoets?**

Bij deelvraag 1 'Wat zijn construct-relevante en -irrelevante strategieën?' werd gedefinieerd welke typen strategieën wenselijk of onwenselijk zijn. Op basis van de literatuur en overwegingen van de onderzoeker werd geconcludeerd dat taalleerderstrategieën altijd construct-relevant zijn, toetswijsheidstrategieën nooit en toetsmanagementstrategieën wanneer logistieke zaken buiten beschouwing worden gelaten.

Bij de beantwoording van deelvraag 2 'Welke strategieën zetten kandidaten in tijdens de DOMC-toets?' en deelvraag 3 'Welke strategieën zetten kandidaten in tijdens de standaard meerkeuzetoets?' bleek dat in beide condities meer toetsmanagement- en toetswijsheidstrategieën werden ingezet dan taalleerderstrategieën. Dit verschil was echter niet significant in de DOMC-conditie na de Bonferroni-correctie ook niet in de MK-conditie. Deze resultaten gaan in tegen mijn verwachting dat taalleerderstrategieën in beide condities het meest gebruikt zouden worden. Uit de kwalitatieve analyse van het strategiegebruik bleek dat in beide condities de meeste antwoorden gebaseerd waren op het al dan niet correct interpreteren van de luistertekst. Dit wijst er juist op dat er wel veel gebruik is gemaakt van taalleerderstrategieën, wat impliceert dat er voornamelijk construct-relevant taalgedrag werd uitgelokt in beide condities. Ook wanneer de meest en minst gebruikte strategieën in beide condities vergeleken werden, vielen de gelijkenissen op.

Deelvraag 4 'Wat zijn de verschillen tussen het strategiegebruik van de kandidaten in de DOMC-toets en het strategiegebruik van de kandidaten in de standaard meerkeuzetoets?' richtte zich op het bepalen van de verschillen in strategiegebruik tussen de twee condities. Er zijn een aantal verschillen gevonden. Ten eerste gaven de proefpersonen in de DOMC-conditie vaker aan het moeilijk te vinden om het goede antwoord te gokken. Het maken van een beredeneerde gok is lastiger, omdat de antwoordopties niet vergeleken kunnen worden. Proefpersonen in de MK-conditie gebruikten relatief vaak het proces van eliminatie. Deze strategie was minder goed toe te passen in de DOMC-conditie, terwijl hij in deze studie beschouwd wordt als een valide manier om tot het goede antwoord te komen. Daarnaast leek het in de DOMC-conditie vaker voor te komen dat een proefpersoon wel de goede zinnen of woorden had gehoord, maar toch het verkeerde antwoord koos. Een reden hiervoor zou kunnen zijn dat de proefpersonen de antwoordopties niet kunnen vergelijken en niet weten of er een optie komt die beter past bij wat zij hebben gehoord en daarom een minder passend antwoord selecteren. In Rossa (2016) kwam het ook voor dat proefpersonen zich de benodigde informatie herinnerden, maar toch het verkeerde antwoord kozen. Volgens hem doet dat af aan de validiteit van een toets. De vraag is of het construct-irrelevante variantie vermindert, omdat de proefpersonen blijkbaar niet overtuigd waren van wat het goede antwoord moest zijn, of dat het er juist aan bijdraagt, omdat de proefpersonen niet durven te wachten op het meest passende antwoord. Mijn verwachting was ook dat het gebruik van toetswijsheidstrategieën zou verminderen in de DOMC-toets. Uit de studie van Papenberg, Willing en Musch (2017) bleek namelijk dat de proefpersonen die een DOMC-toets maakten minder gemakkelijk gebruik konden maken van externe hints dan de proefpersonen die een standaard meerkeuzetoets maakten. Dit bleek in de huidige studie niet het geval te zijn. De proefpersonen in de DOMC-conditie leken zelfs meer te zoeken naar externe hints dan de proefpersonen in de MK-conditie.

Er zijn ook een aantal verschillen gevonden die geen betrekking hebben op het strategiegebruik van de proefpersonen. Er zat bijvoorbeeld een verschil in de tijd die het kostte om de instructie te lezen. Dit duurde significant langer in de DOMC-conditie. Er was daarentegen geen verschil in de tijdsduur van het beantwoorden van de toetsvragen, in tegenstelling tot wat in eerdere onderzoeken werd aangetoond (Foster & Miller, 2009; Willing, Ostapczuk & Musch, 2015). Daarnaast was de gemiddelde toetsscore in de DOMC-conditie significant lager dan in de MK-conditie. De DOMC-toets leek over het algemeen ook moeilijker gevonden te worden. De gemiddelde waarden op de stelling verschilden weliswaar niet significant, maar uit de verdeling van de antwoorden op de stelling 'Ik vond de luistertoets makkelijk' bleek dat relatief veel proefpersonen in de MK-conditie neutraal ten opzichte van de stelling stonden, terwijl veel proefpersonen in de DOMC-conditie het (helemaal) oneens waren met de stelling. Het is mogelijk dat de proefpersonen in de DOMC-conditie meer op zoek gingen naar externe hints dan proefpersonen in de MK-conditie, omdat de toets te moeilijk was. Ze probeerden op die manier wellicht te compenseren voor hun tekortschietende luistervaardigheid. De proefpersonen in de DOMC-conditie stonden relatief neutraal tegenover DOMC-vragen in een toets; ze vonden dit type vraag niet overwegend prettig of vervelend. Proefpersonen in de MK-conditie waren voornamelijk positief over de standaard meerkeuzevraag in een toets. Ook uit eerder onderzoek bleek dat het DOMC-format niet de voorkeur heeft ten opzichte van het standaard meerkeuzeformat (Foster & Miller, 2009). Mogelijk staan kandidaten positiever tegenover het DOMC-format wanneer zij meer gewend zijn aan dit type vraag. Papenberg (2018) ondervond dat proefpersonen beter presteerden in hun tweede DOMC-toets. Mogelijk dragen betere prestaties bij aan een positievere mening over het format.

5.2 Verklaringen resultaten

Een verklaring voor het feit dat er weinig verschillen in strategiegebruik tussen de twee condities zijn gevonden, kan allereerst zijn dat er weinig verschillen bestaan. De DOMC-toets en de standaard meerkeuzetoets zouden in dat geval soortgelijk strategiegebruik eliciteren. Er zijn echter ook andere mogelijke verklaringen voor de gevonden resultaten. Die verklaringen worden in deze paragraaf besproken.

De huidige studie kent een aantal tekortkomingen die er mogelijk voor zorgen dat de resultaten niet te generaliseren zijn naar situaties buiten dit onderzoek. Allereerst waren er niveauverschillen tussen de cursisten. Cursisten met een taalvaardigheidsniveau tussen A2,5 en B1,5 konden deelnemen aan het onderzoek. In de huidige studie zijn alle proefpersonen willekeurig aan een van de twee condities toegewezen, wat het aannemelijk maakt dat de niveaus in beide groepen vergelijkbaar zijn. Het is echter mogelijk dat er relatief veel cursisten met een lager of hoger taalniveau in een van de twee condities vielen. Het is namelijk onbekend wat het exacte niveau van elke proefpersoon was. Dit exacte niveau is ook lastig vast te stellen. Een eventueel verschil in taalvaardigheidsniveau zou de resultaten kunnen beïnvloeden. Met een grotere steekproef zou je nog zekerder van een gelijke verdeling kunnen zijn.

Waar bij het interpreteren van de resultaten ook rekening mee gehouden dient te worden is het feit dat de proefpersonen in dit onderzoek vaardige gebruikers van het Nederlands waren. Het strategiegebruik van vaardige en minder vaardige leerders lijkt namelijk te verschillen (o.a. Cohen en Upton, 2006; Taguchi, 2001). De resultaten kunnen daarom niet gegeneraliseerd worden naar NT2-leerders van een ander vaardigheidsniveau. Het zou interessant zijn om een soortgelijk onderzoek bij andere taalvaardigheidsniveaus uit te voeren.

Ook de vragenlijst over het strategiegebruik wil ik vanuit een kritisch oogpunt bespreken. In de huidige studie werd een selectie van vijftien strategieën bevestigd, vijf in iedere categorie (namelijk: taalleerder-, toetsmanagement- en toetwijsheidstrategieën). De betrouwbaarheid van de drie schalen bleek laag te zijn. De gekozen strategieën representeerden daarom de betreffende categorie strategieën wellicht niet goed. Conclusies op basis van overeenkomsten en verschillen tussen het gebruik van de drie categorieën zijn daardoor minder gerechtvaardigd. Dit geldt echter niet voor conclusies op basis van de individuele strategieën. Het heeft de voorkeur om meer strategieën mee te nemen, zodat de betrouwbaarheid van de schalen vergroot wordt. Om de tijdsduur van het experiment te beperken, was er in dit geval echter voor gekozen om dat niet te doen. Een andere opmerking bij de vragenlijst over het strategiegebruik is de mogelijkheid dat de kandidaten de vragenlijst hebben ingevuld vanuit een meer algemeen perspectief. Hun antwoorden op de vragenlijst zouden dan reflecteren hoe zij *in het algemeen* een toets maken en minder zeggen over hoe zij deze specifieke toets maakten. Dit zou de resultaten in grote mate beïnvloed kunnen hebben. De antwoorden van de proefpersonen in de DOMC-conditie zouden dan mogelijk reflecteren hoe zij een MK-toets maken (ervan uitgaand dat dit een toets is waarmee zij veel te maken krijgen).

Ik wil ook graag een kanttekening zetten bij de uitvoering van het hardopdenkprotocol in dit onderzoek. Volgens Cohen (1994) kan een hardopdenkprotocol het beste gedaan worden in de taal van de toets. Er kan volgens hem informatie verloren gaan wanneer de proefpersoon de processen in de tweede taal moet gaan vertalen naar hun eerste taal. In dit onderzoek waren alle hardopdenkprotocollen in het Nederlands, al kwamen ook Engelse uitingen voor. De toelichtingen van de proefpersonen die de toets individueel maakten, waren voornamelijk in het Nederlands en soms in het Engels. Hoewel dit is wat Cohen voorschrijft, bestaat de kans dat het hardopdenkprotocol hierdoor cognitief overbelastend was of dat de proefpersonen niet goed onder woorden konden brengen wat hun gedachten en redeneringen waren.

De vraag is of iedereen in het onderzoek de instructie goed gelezen heeft. De verschillen tussen de proefpersonen, met name in de DOMC-conditie, waren tamelijk groot: er waren drie proefpersonen die tien keer zo lang deden over het lezen van de instructie van het onderzoek dan drie andere proefpersonen. Dit kan te wijten zijn aan de leesvaardigheid van de proefpersonen, maar de proefpersonen zouden een vergelijkbaar taalvaardigheidsniveau moeten hebben. De timer op de instructie zorgde er weliswaar voor dat de proefpersonen minstens 20 of 30 seconden op de pagina moesten blijven, maar dit wil niet zeggen dat de instructie ook zorgvuldig gelezen werd. In het experiment kwamen echter ook drie oefenvragen voor, dus indien de instructie niet zorgvuldig gelezen was, kregen de proefpersonen hier nogmaals de mogelijkheid om te ontdekken hoe een DOMC-vraag werkt.

Ten slotte was er een tekortkoming die een volgende keer gemakkelijk voorkomen zou kunnen worden. De proefpersonen konden in het experiment niet terug als zij naar de volgende pagina hadden geklikt. Wanneer een proefpersoon te snel doorklikte, kon dit ertoe leiden dat de proefpersoon een luisterfragment niet had afgespeeld en ook niet meer kon afspelen. De proefpersoon moest het antwoord dan gokken. Om te voorkomen dat proefpersonen te snel doorklikken, zou een timer kunnen worden ingesteld, zoals in de huidige studie ook bij de instructie was gedaan.

5.3 Generaliseerbaarheid naar Staatsexamen

Het luisterexamen in deze studie verschilde op een aantal punten van het luisterexamen in het Staatsexamen Programma I. Het is daarom onduidelijk of de resultaten gegeneraliseerd kunnen worden naar dat volledige examen en of het DOMC-format in deze vorm toegepast kan worden in het

Staatsexamen. Het luisterexamen in het Staatsexamen kent namelijk ook vragen waarbij een video getoond wordt. Wanneer er videobeelden beschikbaar zijn, kan de kandidaat gebruikmaken van die extra informatie, zoals gebaren, gezichtsuitdrukkingen en begeleidende tekst of foto's. Het ligt daarom voor de hand dat andere strategieën ingezet kunnen worden. Li (2013) onderzocht het gebruik van strategieën in een luistertoets met of zonder video's. De toets bestond uit vier minicolleges, waarvan twee met video en twee zonder video. Uit zijn analyse van de resultaten bleek dat er wat betreft strategiegebruik merkbare verschillen waren tussen de verschillende modi. Li maakte een onderscheid tussen bron-van-kennisstrategieën (de directe reden van de kandidaat voor zijn keuze) en benadering-van-het-probleemstrategieën (de initiële strategieën die een kandidaat inzet om een probleem op te lossen). In vergelijking met de audiotaken, werden minder benadering-van-het-probleemstrategieën en meer bron-van-kennisstrategieën ingezet in de videotaken. De visuele informatie werd voornamelijk gebruikt om het begrip van de proefpersonen te verfijnen, een hypothese op te stellen of te bevestigen. Het kon echter ook voor verwarring zorgen, waardoor het luisterbegrip van de proefpersonen gehinderd werd. Een ander verschil met het luisterexamen uit het Staatsexamen is dat de antwoordmogelijkheden pas na het luisteren zichtbaar waren. Enkele proefpersonen gaven aan hier moeite mee te hebben. Dat is logisch, omdat ze minder gericht naar woorden uit de antwoordopties kunnen luisteren dan normaal. Dit is wellicht juist iets goeds, aangezien de proefpersonen niet geholpen worden door de antwoordopties. Ze moeten in eerste instantie zelf proberen een antwoord te formuleren op de vraag, die ze wel voorafgaand aan het luisteren konden lezen. Ook hier is meer onderzoek nodig. Het laatste relatief grote verschil met het echte luisterexamen is dat de proefpersonen meer tijd kregen om te antwoorden. Het echte luisterexamen moet binnen 90 minuten gemaakt worden. In dit experiment kregen de proefpersonen onbeperkt de tijd. Dit kan invloed hebben gehad op hun antwoorden.

5.4 Meer onderzoek naar DOMC-toetsen

Voor dit experiment zijn vragen uit het Oefenexamen Luisteren 2017 gebruikt. Dit is een standaard meerkeuzetoets. Het format is in de DOMC-conditie door de onderzoeker zelf omgezet, terwijl de vragen en antwoordopties gelijk werden gehouden. De toets is dus niet speciaal voor het DOMC-format ontworpen. De vraag is of een DOMC-toets anders ontworpen zou moeten worden dan een standaard meerkeuzetoets. Wellicht zouden de alternatieven in een DOMC-toets inhoudelijk verder uit elkaar moeten liggen dan in een meerkeuzetoets, en misschien is het in een DOMC-toets belangrijker dat een fout antwoord ook echt fout is en een goed antwoord echt goed. Vragen waarbij het beste antwoord moet worden gekozen, zijn niet geschikt voor het DOMC-format. Volgens Foster (2016) eist het DOMC-format op dit gebied meer van de toetsontwikkelaar. Daarnaast gaven enkele proefpersonen in de DOMC-conditie aan moeite te hebben met de toetsvragen waarbij het juiste plaatje gekozen moest worden. Het was volgens de proefpersonen niet altijd direct duidelijk wat de plaatjes toonden. Dit is mogelijk extra belangrijk bij een DOMC-vraag. Het is ook denkbaar dat dit soort vragen niet geschikt zijn voor het DOMC-format. Er is nog weinig bekend over welke vragen en antwoordmogelijkheden geschikt zijn voor het DOMC-format. Het zou ook interessant zijn om te onderzoeken of videotaken geschikt zijn voor het DOMC-format en of het strategiegebruik dan verandert. Ten slotte moet nagedacht worden over de volgorde waarin de alternatieven gepresenteerd worden. Er is nog geen onderzoek gedaan naar de manier waarop de resultaten beïnvloed worden als het goede antwoord relatief vaak het eerste of het laatste antwoord is. In het huidige format is er bijvoorbeeld een gokkans aanwezig, wanneer de kandidaat niet overtuigd is van het laatste alternatief, maar dat antwoord desondanks als correct aanwijst, omdat het de laatste mogelijkheid is. Dit zijn aspecten waarover nagedacht moet worden bij het ontwerpen van een DOMC-toets. Op dit moment zijn er nog geen richtlijnen voor het ontwerpen van DOMC-items die

taalvaardigheid toetsen. Er zijn evenmin richtlijnen voor het voorbereiden van kandidaten op een DOMC-toets. Hoewel de proefpersonen in de huidige studie niet negatief tegenover het format stonden, zou het format onzekerheid en faalangst in de hand kunnen werken. Er waren namelijk gevallen waarin de proefpersonen liever een eerder antwoord hadden geselecteerd. Dit kan stress opleveren, met name wanneer de toetsuitslag grote gevolgen heeft. Kandidaten zouden daarom ter voorbereiding in de lessen getraind moeten worden met DOMC-vragen. Een gevaar daarvan is dat de focus in de lessen teveel op het format komt te liggen in plaats van de daadwerkelijke taalvaardigheid. Dergelijke negatieve *backwash* dient voorkomen te worden.

In beide condities werden relatief veel toetswijsheidstrategieën ingezet, al is onduidelijk in hoeverre het inzetten van die strategieën tot het kiezen van het goede antwoord heeft geleid. Dit was lastig te achterhalen, omdat de classificatie van een strategie niet altijd te bepalen was op basis van de toelichtingen. Als bijvoorbeeld iemand 'Eliminatie' als toelichting gaf, bleef onduidelijk op basis waarvan die persoon antwoorden geëlimineerd had: op basis van externe hints of op basis van wat (niet) gehoord was? Dat maakt in dit geval het verschil tussen een toetswijsheid- of een toetsmanagementstrategie. Het classificeren van strategieën was gemakkelijker in het hardopdenkprotocol. Uit een analyse van de hardopdenkprotocollen bleken op dit gebied geen beduidende verschillen tussen de twee condities. Proefpersonen in de DOMC-conditie gaven wel aan het moeilijker te vinden om het goede antwoord te gokken dan de proefpersonen die de standaard meerkeuzetoets maakten. Er is meer onderzoek nodig, met een grotere steekproef, om te bepalen of de mening van de proefpersonen op dit gebied gereflecteerd wordt in het aantal goede gokken dat zijn doen tijdens een toets. Daarnaast is er meer onderzoek nodig om te bepalen of de dingen die opvielen tijdens de kwalitatieve analyse van het strategiegebruik (zoals het hebben gehoord van de juiste informatie, maar toch het verkeerde antwoord kiezen) toevalligheden zijn of een daadwerkelijk verschil tussen de twee soorten toetsen reflecteren.

5.5 Conclusie

Uit de huidige studie is gebleken dat het strategiegebruik in beide toetsen vergelijkbaar is. Op basis van het strategiegebruik is er dus niet een type toets te prefereren boven het andere. Ik ben terughoudend met het generaliseren van de resultaten van deze studie wegens de beschreven tekortkomingen en de conflicterende resultaten met andere studies.

Er is op het gebied van de DOMC-vraag nog veel te onderzoeken. Ook al bleken uit deze studie geen grote voordelen van het DOMC-format, bleken die er in andere onderzoeken wel te zijn. Die voordelen maken het de moeite waard om verder te onderzoeken of en wanneer het DOMC-format tot valide meningen leidt.

Bij een volgend soortgelijk onderzoek zou ik aanraden om de tekortkomingen en overwegingen van deze studie mee te nemen. Er moet bijvoorbeeld nagedacht worden over het aantal strategieën dat bevraagd wordt en over de manier waarop het hardopdenkprotocol het beste afgenomen kan worden.

Literatuurlijst

- Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing*, 9(2), 101-119.
- Assiri, M. S. (2011). *Test-taking strategy use on the reading section of the TOEFL iBT: A study of Arab ESL learners* (Doctoral dissertation, Oklahoma State University).
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, VK: Oxford University Press.
- Beare, S. (2000). *Differences in content generating and planning processes of adult L1 and L2 proficient writers*. University of Ottawa, Canada.
- Berman, R. (1994). Learner's transfer of writing skills between languages. *TESL Canada Journal* 12, 29-46.
- Buck, G. (2001). *Assessing listening*. Cambridge, VK: Cambridge University Press.
- Bügel, K. & Sanders, P.F. (1998). *Richtlijnen voor de ontwikkeling van onpartijdige toetsen*. Arnhem, Nederland: Cito.
- Cohen, A.D (1984). On taking tests: what the students report. *Language Testing*, 1, 70-81.
- Cohen, A.D. (1994). Verbal reports on learning strategies. *TESOL Quarterly*, 28, 678-682
- Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, 3(4), 307-331.
- Cohen, A. D. (2011). *Strategies in learning and using a second language*. Harlow, VK: Longman Applied Linguistics/Pearson Education.
- Cohen, A. D. (2014). Using Test-Wiseness Strategy Research in Task Development. In A.J. Kunnan (Red.), *The Companion to Language Assessment*, 2, (pp. 893-905). Hoboken, VS: John Wiley & Sons.
- Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks*. *ETS Research Report Series*, 2006(1), i-162.
- Cohen, A. D., & Upton, T. A. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL. *Language Testing*, 24(2), 209-250
- College voor Toetsen en Examens. (z.d.-a). *Staatsexamens NT2: Examenonderdelen*. Geraadpleegd van <https://www.staatsexamensNT2.nl/item/examenonderdelen>
- College voor Toetsen en Examens. (z.d.-b). *Staatsexamens NT2: Inhoud van het examen*. Geraadpleegd van <https://www.staatsexamensNT2.nl/item/inhoud-van-het-examen>
- College voor Toetsen en Examens. (2018, 5 december). *Geen examens meer in 2018: nieuw systeem in 2019*. Geraadpleegd van <https://www.staatsexamensnt2.nl/nieuws/20181205/geen-examens-meer-in-2018-nieuw>
- College voor Toetsen en Examens. (2019a, 25 maart). *Aanmelden voor Staatsexamens Nt2 weer mogelijk*. Geraadpleegd van <https://www.cvte.nl/actueel/nieuws/2019/03/25/aanmelden-voor-staatsexamens-nt2-weer-mogelijk>
- College voor Toetsen en Examens. (2019b). *Jaarverslag 2019 College voor Toetsen en Examen*.
- Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing*, 27(4), 471-491.
- Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as SAT. *Journal of Experimental Psychology*, 130(2), 208-223.
- Douglas, D., & Hegelheimer, V. (2005). *Cognitive processes and use of knowledge in performing new TOEFL listening tasks* (2nd Interim Report to Educational Testing Service). Ames: Iowa State University.
- Ericsson, A. (2003). Valid and non-reactive verbalization of thoughts during performance of tasks towards a solution to the central problems of introspection as a source of scientific data. *Journal of consciousness studies*, 10(9-10), 1-18

- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Herziene editie). Cambridge, VS: The MIT Press.
- Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(3), 632.
- Foster, D. (2016). *The Discrete Option Multiple Choice Test Item: An Important Evolution in Test and Item Design*.
- Foster, D., & Miller Jr., H. L. (2009). A new format for multiple-choice testing: Discrete-Option Multiple-Choice. Results from early studies. *Psychological Test and Assessment Modeling*, 51(4), 355.
- Gordon, C. M. (1987). *The effect of testing method on achievement in reading comprehension tests in English as a foreign language*. Masterscriptie, Tel-Aviv University.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3), 309-333.
- Hughes, A. (2002). *Testing for language teachers*. Cambridge, VK: Cambridge University Press.
- Katz, S., Lautenschlager, G. J., Blackburn, A. B., & Harris, F. H. (1990). Answering reading comprehension items without passages on the SAT. *Psychological Science*, 1(2), 122-127.
- Kingston, N. M., Tiemann, G. C., Miller Jr, H. L., & Foster, D. (2012). An analysis of the discrete-option multiple-choice item type. *Psychological Test and Assessment Modeling*, 54(1), 3.
- Li, Z. (2013). An empirical study of test-taking strategies in a video listening placement test. *Language and Communication Quarterly*, 1(2), 103-123.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, 18(2), 5-11.
- Papenberg, M. (2018). *On how testwiseness and acceptance reluctance influence the validity of sequential knowledge tests* (Proefschrift, Universitäts-und Landesbibliothek der Heinrich-Heine-Universität Düsseldorf).
- Papenberg, M., Diederhofen, B., & Musch, J. (2019). An Experimental Validation of Sequential Multiple-Choice Tests. *The Journal of Experimental Education*, 1-20.
- Papenberg, M., Willing, S., & Musch, J. (2017). Sequentially presented response options prevent the use of testwiseness cues in multiple-choice testing. *Psychological Test and Assessment Modeling*, 59(2), 245.
- Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge, VK: Cambridge University Press.
- Radboud in'to Languages. (z.d.). 5 grandioze geheimen om elke meerkeuzetoets te halen! (En hoe je dit als docent kunt voorkomen). Geraadpleegd van <https://www.ru.nl/radboudintolanguages/over-ons/onze-blogs/5-grandioze-geheimen-elke-meerkeuzetoets-halen/>
- Rijksoverheid (2020, 2 juli). *Nieuwe Wet inburgering*. Geraadpleegd van <https://www.rijksoverheid.nl/actueel/nieuws/2020/07/02/nieuwe-wet-inburgering-aangenomen>
- Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5), 1155.
- Rossa, H. (2016). The cognitive processes elicited by L2 listening test tasks—A validation study. *Developing, Modelling and Assessing Second Languages*, 207.
- Salehi, M. (2011). Test Taking Strategies: Implications for Test Validation. *Journal of Language Teaching & Research*, 2(4).
- Taguchi, N. (2001). L2 learners' strategic mental processes during a listening test. *JALT Journal*, 23(2), 176-201.
- Tavakoli, E., & Samian, S.H. (2014). Test-wisness Strategies in PBTs and IBTs: The Case of EFL Test Takers, Who Benefits More?. *Procedia-Social and Behavioral Sciences*, 98, 1876-1884.
- Teng, H. C. (2013). A Study of Test-Taking Strategies Used by EFL Listeners. *The 2013 PanSIG Proceedings*, 27.

- Tsagari, C. (1994). *Method effects on testing reading comprehension: How far can we go?* Masterscriptie, University of Lancaster, UK.
- Vandergrift, L., Goh, C. C., Mareschal, C. J., & Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire: Development and validation. *Language learning, 56*(3), 431-462.
- Vandergrift, L., & Tafaghodtari, M. H. (2010). Teaching L2 learners how to listen does make a difference: An empirical study. *Language learning, 60*(2), 470-497.
- Willing, S., Ostapczuk, M., & Musch, J. (2015). Do sequentially-presented answer options prevent the use of testwiseness cues on continuing medical education tests?. *Advances in Health Sciences Education, 20*(1), 247-263.
- Winke, P., & Lim, H. (2014). The effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation. *IELTS Research Reports Online Series, 30*.
- Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *The Modern Language Journal, 77*(iv), 473–489.
- Wu, A. D., & Stone, J. E. (2016). Validation through understanding test-taking strategies: An illustration with the CELPIP-General reading pilot test using structural equation modeling. *Journal of Psychoeducational Assessment, 34*(4), 362-379.
- Yoshida-Morise, Y. (1998). The use of communication strategies in language proficiency interviews. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 205–238). Amsterdam, NL: John Benjamins.

Bijlage I: Alle toetsvragen en antwoorden

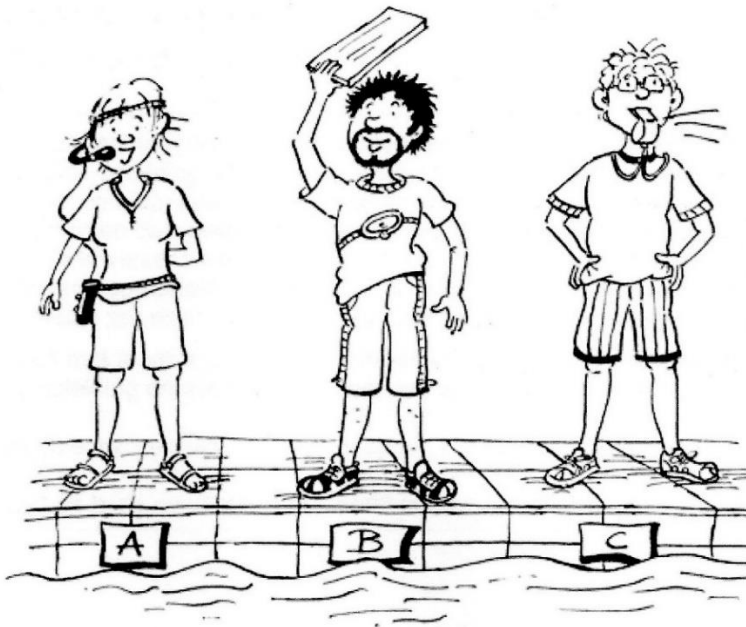
Vraag 1. Parttime werken betekent dat je niet de hele week werkt. Waarom wil het zwembad graag mensen die parttime werken?

- A Omdat het werk veel te zwaar is om de hele week te doen.
- B Omdat ze dan meer mensen tegelijk kunnen inzetten als dat nodig is.**
- C Omdat ze dan niet zoveel geld aan personeel hoeven uit te geven.

Vraag 2. Wat vertelt Jochem over omgaan met vervelende mensen in het zwembad?

- A Daarmee heb je vaak te maken.
- B Dat leer je in de opleiding.**
- C Dat vindt hij erg moeilijk.

Vraag 3. Welke oplossing geeft Jochem zodat je niet zoveel hoeft te schreeuwen? (**Antwoord B**)



Vraag 4. Wat vertelt Jochem over zwembadmedewerkers die al wat ouder zijn?

- A Die gaan meestal ander werk doen in het zwembad.**
- B Die worden regelmatig onderzocht op hun gezondheid.
- C Die zoeken toch vaak ander werk buiten het zwembad.

Vraag 5. Wat zegt Maike over het diploma Verkoopmedewerker? Met dit diploma kun je

- A alleen bij grote warenhuizen werken.
- B gemakkelijk naar een vervolgopleiding.
- C in iedere winkel werken.**

Vraag 6. Wat zegt Maaïke hier over de begeleiding tijdens de opleiding?

- A Gedurende de hele opleiding krijg je begeleiding vanuit de school.
- B In het begin van de opleiding krijg je de meeste begeleiding.
- C Je hebt een vaste begeleider op de afdeling van het warenhuis.**

Vraag 7. Wat zegt Maaïke over de werknemers op de afdeling elektronica?

- A Zij hebben voor hun werk een speciale opleiding gevolgd.
- B Zij hoeven alleen maar te weten hoe eenvoudige elektrische apparaten werken.**
- C Zij vinden vaak een nieuwe baan in een gespecialiseerde elektronikawinkel.

Vraag 8. Wat adviseert Maaïke hier over het lezen van de gebruiksaanwijzing van een apparaat?

- A Doe dat samen met een collega.
- B Doe dat thuis.
- C Doe dat zoveel mogelijk onder werktijd.**

Vraag 9. Wat adviseert Maaïke als een klant vraagt of een jurk mooi staat?

- A Geef eerlijk antwoord.**
- B Vraag of ze de jurk zelf mooi vindt staan.
- C Zeg dat de jurk haar mooi staat.

Vraag 10. Wat moet je volgens Maaïke doen als een klant korting wil hebben?

- A Vragen hoeveel korting de klant wil hebben.
- B Vragen of je chef even wil komen.**
- C Vragen waarom de klant korting wil hebben.

Vraag 11. Wat wil Maaike met de voorbeelden duidelijk maken?

- A **Waarom je dieven kunt herkennen.**
- B Waarom sommige klanten stelen.
- C Wat voor spullen er gestolen worden in het warehouse.

Vraag 12. Waarop moet je volgens Moud vooral letten als je gaat samenwerken met een ander bedrijf?

- A **Dat heel duidelijk is wie welk werk doet in de klus.**
- B Dat het werk eerlijk verdeeld wordt tussen de bedrijven.
- C Dat iedereen op tijd werkt, zodat je niet op elkaar hoeft te wachten.

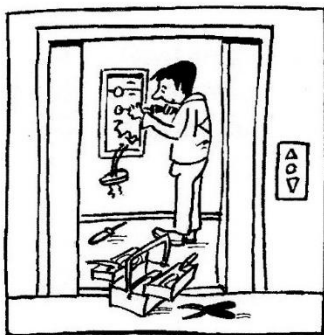
Vraag 13. Wat zegt Moud over het berekenen van de prijs?

- A De meeste klanten hebben het liefst een vaste prijs voor de hele klus.
- B **Je moet zorgen dat de klant vooraf weet hoe je de prijs berekent.**
- C Klanten willen de uren om dingen uit te zoeken meestal niet betalen.

Vraag 14. Wat zegt Moud over klanten die niet willen betalen?

- A **Als het werk goed is, moeten ze op den duur toch betalen.**
- B Dit probleem heeft Moud gelukkig nog nooit gehad.
- C Je moet dan toch blijven samenwerken met deze klanten.

Vraag 15. Op welk plaatje staat een klus die Moud noemt als iets wat nodig is in een kantoor?
(Antwoord C)



A



B



C

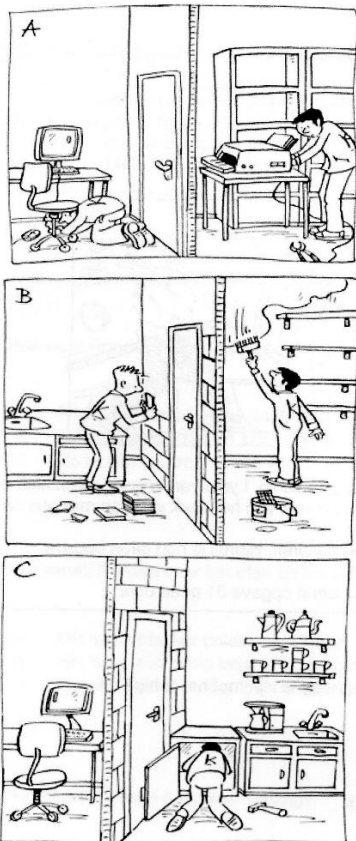
Vraag 16. Wat is volgens Moud het belangrijkste voordeel van werken met vaste klanten?

- A **Met vaste klanten is het overleg veel gemakkelijker.**
- B Vaste klanten kunnen soms vrienden van je worden.
- C Veel vaste klanten zijn sneller tevreden over je werk.

Vraag 17. Wat zegt Moud over de dingen die je nodig hebt?

- A Heel veel materialen die je bewaart, zul je nooit meer gebruiken.
- B Je kunt het beste je materialen bewaren in een paar grote kasten.
- C **Je moet proberen om zelf zo min mogelijk materialen te hebben.**

Vraag 18. Op welk plaatje staat wat Moud de komende week gaat maken? **(Antwoord B)**



Vraag 19. Hoe krijgt Moud alle dingen uit de auto op de juiste plaats?

- A De klant moet helpen het materiaal binnen te brengen.
- B **Er is altijd wel iemand bereid om hem even te helpen.**
- C Iemand die bij de bouwwinkel werkt, helpt hem daarbij.

Bijlage II: Lijst met strategieën en bijbehorende categorie

Strategie	Categorie
Ik probeerde tijdens het luisteren onbelangrijke informatie te negeren.	Taalleerderstrategie
Ik bedacht van tevoren welke woorden ik misschien zou gaan horen.	Taalleerderstrategie
Ik gebruikte kennis van de talen die ik spreek.	Taalleerderstrategie
Ik bedacht tijdens het luisteren hoe de tekst verder zou gaan.	Taalleerderstrategie
Ik bedacht voorafgaand aan het luisteren voor welk doel ik ging luisteren.	Taalleerderstrategie
Ik voorspelde wat het antwoord zou zijn en keek of mijn antwoord erbij stond.	Toetsmanagementstrategie
Ik controleerde of mijn antwoord logisch was.	Toetsmanagementstrategie
Ik koos een antwoord op basis van de algemene betekenis van de tekst.	Toetsmanagementstrategie
Ik koos een antwoord, omdat ik wist dat de andere antwoorden fout waren.	Toetsmanagementstrategie
Ik probeerde de vraag anders te formuleren.	Toetsmanagementstrategie
Ik koos een antwoord, omdat het een woord bevatte dat ook in de tekst of in de vraag genoemd werd.	Toetswijsheidstrategie
Ik lette op hints in de vraag, in de antwoorden en in andere vragen.	Toetswijsheidstrategie
Ik koos een antwoord, ook al wist ik het niet zeker.	Toetswijsheidstrategie
Ik vergeleek de antwoorden met elkaar en probeerde zo antwoorden te elimineren.	Toetswijsheidstrategie
Ik gebruikte algemene kennis om de vraag te beantwoorden.	Toetswijsheidstrategie