College voor Toetsen en Examens

Ontwikkelingen in examinering van tweede taal

Een literatuuronderzoek

Kaatje Dalderop
in opdracht van het CvTE

10 februari 2026

Inhoud

Managementsamenvatting **4**

1 Introductie **5**

2 Onderzoeksvragen **6**

3 Methode **7**

4 Resultaten **9**

4.1 Hedendaagse inzichten en tendensen **9**
4.1.1 Bij de start 10
4.1.2 Validiteit 11
4.1.3 Technologie 13
4.1.4 Het gebruik van toetsen 14
4.1.5 De toekomst 15

4.2 De voor- en nadelen van geïntegreerd toetsen **15**
4.2.1 Het construct 17
4.2.2 Vergelijkbaarheid van geïntegreerde en onafhankelijke taken 17
4.2.3 De relatie tussen receptieve en productieve vaardigheden 19
4.2.4 Kenmerken van de producten 20
4.2.5 De kandidaat 20
4.2.6 Task-based language assessment 21
4.2.7 Betrouwbaarheid 22
4.2.8 Kritische reflectie 23

4.3 Alternatieven voor het toetsen van vier vaardigheden **24**
4.3.1 Cloze tests 25
4.3.2 Word Associates Format (WAF) 25
4.3.3 Context-gebonden assessment en portfolio 26
4.3.4 Toetsen die meerdere niveaus meten 27
4.3.5 Kritische reflectie 31

4.4 Alternatieven voor afname van het spreekexamen via een talenpracticum-model **31**
4.4.1 Het construct spreekvaardigheid 32
4.4.2 Afnameformats 33
4.4.3 Toetstaken 35
4.4.4 Voor- en nadelen van de modellen 35
4.4.5 Kritische reflectie 37

4.5 Lezen van een beeldscherm en van papier **37**
4.5.1 Het construct leesvaardigheid 38
4.5.2 De lezer 39
4.5.3 Tekstbegrip 40

4.6 Beoordeling van Schrijven en Spreken **42**
4.6.1 Het construct spreekvaardigheid 44
4.6.2 Het construct schrijfvaardigheid 44
4.6.3 Het ERK 44
4.6.4 Validiteit en betrouwbaarheid 45
4.6.5 Nieuwe taaktypen 45
4.6.6 Beoordelingsmodellen en beoordelaars 46
4.6.7 Beoordeling van IELTS en TOEFL 48
4.6.8 Kritische reflectie 49

4.7 Positieve beoordelingsmodellen **49**
4.7.1 CAF onderzoek 51
4.7.2 Hoe hangt CAF samen met tweede-taalontwikkeling? 52
4.7.3 Hoe hangt CAF samen met taakkenmerken? 54
4.7.4 Kritische reflectie 56

4.8 Dichotoom beoordelen **57**
4.8.1 Checklists 58
4.8.2 Beslisbomen 61
4.8.3 Onderzoek naar schaallengte 63
4.8.4 Kritische reflectie 63

4.9 Holistisch beoordelen **63**
4.9.1 Analytische en holistische modellen 65
4.9.2 Onderzoek dat holistische en analytische modellen vergelijkt 65
4.9.3 TOEFL iBT 68
4.9.4 Beoordelaars 70
4.9.5 Kritische reflectie 71

4.10 Andere beoordelingsmodellen **72**
4.10.1 Multiple trait scoring 73
4.10.2 Paarsgewijze vergelijking 73
4.10.3 Kritische reflectie 76

5 Conclusies en aanbevelingen **77**

5.1 Conclusies **77**

5.2 Aanbevelingen **78**

5.3 Beperkingen van het onderzoek **79**

Begrippenlijst **81**

Referenties **83**

Managementsamenvatting

Dit literatuuronderzoek heeft tot doel na te gaan in hoeverre het staatsexamen Nt2 in de huidige vorm aansluit bij recente inzichten op het gebied van taaltoetsing. Daartoe is een systematische literatuurreview uitgevoerd, waarbij relevante onderzoeksliteratuur is geselecteerd, geanalyseerd, samengevat en besproken. Het onderzoek is uitgevoerd in de periode oktober 2025 tot januari 2026.

Het staatsexamen Nt2 bestaat sinds 1992 en kent twee programma's: programma I is bedoeld voor anderstaligen die willen werken of studeren op mbo-niveau 3 of 4 en programma II is voor anderstaligen die willen studeren of werken op hbo- of wo-niveau. Het oorspronkelijke doel van beide programma's was kandidaten een startbewijs van taalvaardigheid te bieden voor onderwijs en werk. Sindsdien is de maatschappelijke context ingrijpend veranderd. Met name veranderde opvattingen over migratie hebben ertoe geleid dat het staatsexamen Nt2 tegenwoordig ook wordt ingezet in het kader van inburgering. Deze ontwikkeling is niet uniek voor Nederland maar zichtbaar in heel Europa. Hierdoor is het belang van validiteit en betrouwbaarheid van de examens verder toegenomen. Validiteit wordt daarbij niet langer opgevat als een eigenschap van de toets zelf, maar omvat ook de rechtvaardigheid van het toetsgebruik.

Een tweede belangrijke maatschappelijke ontwikkeling betreft digitalisering. Veel grote taalexamens worden inmiddels digitaal afgenomen, wat voordelen biedt op het gebied van efficiëntie en betrouwbaarheid, maar ook nieuwe uitdagingen met zich meebrengt, onder meer op het terrein van veiligheid. Daarnaast worden grote taalmodellen (AI) steeds vaker ingezet bij afname en beoordeling, een trend die naar verwachting zal doorzetten.

Wat betreft toetsinhoud is in dit onderzoek gekeken naar de integratie van vaardigheden en niveaus. Geïntegreerde vaardigheden¬toetsen – waarbij bijvoorbeeld schrijven plaatsvindt op basis van gelezen of beluisterde input – worden vaak gezien als relevanter en eerlijker dan toetsen die de vier vaardigheden afzonderlijk toetsen. Het beschikbare onderzoek kon deze veronderstelde voordelen echter niet overtuigend bevestigen. Integratie van niveaus, waarbij kandidaten een score en niveau-indicatie ontvangen in plaats van een zak/slaagbeslissing, lijkt voor kandidaten wel voordelen te bieden, omdat een score altijd informatieve waarde behoudt.

Onderzoek gericht op de afname van spreekexamens laat zien dat computerafname internationaal steeds gebruikelijker is. Face-to-face interviews, zoals bij het bekende Engelse examen IELTS, worden bekritiseerd vanwege mogelijke beoordelaarsbias en de sterke sturing door de examinator. Computerafname wordt daarom als een beter alternatief gezien.

Uit onderzoek naar de afname van leesexamens en de vergelijking van lezen op papier versus lezen van scherm blijkt dat lezen van papier nog altijd betere prestaties oplevert dan lezen van scherm, met name bij langere en complexere teksten. Het verschil is over de afgelopen decennia wel afgenomen.

Het tweede deel van het onderzoek richt zich op de beoordeling van productieve vaardigheden. Daarvoor worden verschillende typen beoordelingsmodellen gebruikt. Analytische beoordelingsmodellen zoals ook gebruikt worden bij het staatsexamen Nt2, blijken doorgaans betrouwbaarder dan holistische modellen. Automatische beoordeling is veelbelovend voor het scoren van aspecten van de taal, zoals complexiteit en vloeiendheid. Een innovatieve methode, paarsgewijze vergelijking, laat zeer hoge betrouwbaarheid zien, en is daarom mogelijk interessant.

De centrale conclusie luidt dat het staatsexamen Nt2 qua vorm en inhoud in grote lijnen nog goed aansluit bij hedendaagse inzichten in taaltoetsing. Een aantal verbeteringen is nog mogelijk, bijvoorbeeld door uitgebreidere feedback, verbeterde afnamecondities bij het spreekexamen en keuzemogelijkheden in taken. Verder onderzoek naar de efficiëntie van de beoordeling wordt aanbevolen, evenals een kritische evaluatie van de lees- en luisterexamens, die buiten de scope van dit onderzoek bleven. Tot slot benadrukt dit rapport het belang van een blijvend kritische houding ten aanzien van het politieke gebruik van taaltoetsen in het algemeen en van het staatsexamen Nt2 in het bijzonder.

1 Introductie

In 1992 werd het staatsexamen Nederlands als tweede taal (Nt2) voor het eerst afgenomen. Het staatsexamen Nt2 bestond en bestaat uit twee programma's: programma I en programma II. Programma I is bedoeld voor anderstaligen die willen werken of studeren op mbo-niveau 3 of 4 en programma II is voor anderstaligen die willen studeren of werken op hbo- of wo-niveau (Van der Zanden & Dirkx, 2024).

Sinds de start van het examen in 1992 zijn er uiteraard ontwikkelingen geweest: zo is na de publicatie van het Europees Referentiekader (Council of Europe, 2001) de relatie tussen de twee staatsexamenprogramma's en de ERK-niveaus onderzocht en zijn de examens uiteindelijk ook gekoppeld aan deze niveaus, zodat we nu weten dat programma I het niveau B1 meet, en programma II het niveau B2. Ook op het gebied van digitalisering is het nodige veranderd. Waar kandidaten hun antwoorden op de lees- en luisterexamens in de eerste jaren van het staatsexamen Nt2 aankruisten op een antwoordblad, dat vervolgens moest worden uitgelezen door een optische leesmachine (en in het geval van kreukels eerst moest worden gestreken met een strijkijzer!), is afname via de computer nu vanzelfsprekend.

Wat echter door de jaren in grote lijnen hetzelfde bleef, is de opzet van het examen. Gebaseerd op de communicatieve benadering van taalvaardigheid toetst elk programma vier vaardigheden: Luisteren, Spreken, Lezen en Schrijven. Op het gebied van taaktypen is er weinig veranderd en ook de beoordeling van de productieve vaardigheden verloopt in grote lijnen zoals in het begin (zij het met grote verbeteringen op het gebied van kwaliteit, efficiëntie, monitoring en controle).

Bij het College voor Toetsen en Examens (CvTE) was er zoveel jaar na de start de behoefte ontstaan aan een check: sluit het staatsexamen Nt2 nog aan bij de hedendaagse inzichten op het gebied van taaltoetsing? Die vraag ontstond met name bij de productieve vaardigheden Spreken en Schrijven rondom de beoordelingsmodellen. Om die vraag te beantwoorden werd de opdracht verstrekt tot het uitvoeren van een literatuuronderzoek, op basis van een tiental onderzoeksvragen (hoofdstuk 2). De gehanteerde methode wordt beschreven in hoofdstuk 3 en de resultaten in hoofdstuk 4. Het onderzoek valt in drie delen uiteen: na een brede hoofdvraag naar hedendaagse inzichten en tendensen (paragraaf 4.1) , bespreekt de tekst een eerste cluster deelvragen met betrekking tot de vorm en inhoud van de examens (paragraaf 4.2 t/m 4.5). De tweede set deelvragen zoomt diep in op beoordeling en beoordelingsmodellen voor de productieve vaardigheden Spreken en Schrijven (paragraaf 4.6 t/m 4.10). De tekst sluit af met conclusies en aanbevelingen.

2 Onderzoeksvragen

Het literatuuronderzoek heeft tot doel een groot aantal vragen te beantwoorden. Een deel van de vragen heeft betrekking op de inhoud en vorm van de examens, een ander deel richt zich op de beoordelingsmodellen.

Hoofdvraag 1:

Wat zegt wetenschappelijke onderzoek over ontwikkelingen met betrekking tot inhoud en de beoordeling van taalexamens in een tweede en vreemde taal in de afgelopen 30 jaar: wat zijn hedendaagse inzichten en tendensen? (Blijkend uit meta-analyses, handboeken over taaltoetsing, grootschalige onderzoeken naar betrouwbaarheid en validiteit van grootschalige highstakes examens.)

Deelvragen:

De deelvragen hebben betrekking op het format en de inhoud van de examens en op de beoordeling ervan. Ontwikkelingen op het gebied van psychometrie en de inzet van AI bij ontwikkeling en beoordeling van toetsen vallen buiten de scope van dit onderzoek. De relatie tussen het ERK Companion Volume en het staatsexamen Nt2 vallen eveneens buiten de scope van het onderzoek.

Inhoud van de examens

1. Welke alternatieven zijn er voor het toetsen van vier vaardigheden (luisteren, spreken, lezen en schrijven) en in welke mate worden deze formats gebruikt bij highstakes examens?
2. Wat zijn voor- en nadelen van integratief toetsen (combineren van meerdere deelvaardigheden in één examen)?
3. Welke alternatieven zijn er voor afname van het spreekexamen via een talenpracticum-model, wat zijn de voor-en nadelen van de alternatieven?
4. Is lezen van een beeldscherm anders dan lezen van papier? Is een leesexamen waarbij van scherm moet worden gelezen valide en betrouwbaar?

Beoordeling van de examens

1. Wat zegt wetenschappelijke onderzoek op het gebied van beoordeling van tweede- en of vreemde-taalexamens, met name op het gebied van de onderdelen Schrijven en Spreken?
2. Wat is er bekend over positieve beoordelingsmodellen (beoordelingsmodellen die productie honoreren en die minder focussen op fouten). Worden deze modellen ingezet in highstakes examens?
3. Wat is er bekend over dichotoom beoordelen? Zijn er examens die werken met dichotome beoordelingsmodellen voor het beoordelen van open opgaven Spreken en Schrijven? En hoe zitten die beoordelingsmodellen in elkaar?
4. Wat is er bekend over holistisch beoordelen? Zijn er andere examens die voor het beoordelen van Schrijven en Spreken werken met (gedeeltelijk) holistische beoordelingsmodellen? En hoe zitten die beoordelingsmodellen in elkaar?
5. Zijn er nog andere beoordelingsmodellen die worden gebruikt bij summatieve taalexamens? Wat zijn voor- en nadelen, wat zijn de ervaringen, hoe functioneren ze in de praktijk?

3 Methode

Om de onderzoeksvragen te beantwoorden is een systematische review van de literatuur uitgevoerd. Een systematische review heeft tot doel te komen tot een uitgebreid beeld van de beschikbare literatuur in een gespecificeerd onderzoeksveld (Bramer et al., 2017; Macaro & Cohen, 2007). De zoekprocedure maakt deel uit van de rapportage, zodat deze herhaalbaar en controleerbaar is.

De gebruikte database is Google Scholar. Voor elk van de deelvragen en voor de hoofdvraag afzonderlijk is de database doorzocht op relevante literatuur. De resultaten zijn geordend op relevantie; dit is een functie van de database. Daarna zijn de eerste honderd resultaten nader bestudeerd door de samenvatting te lezen, of als deze ontbraken de introductie. Vervolgens zijn op basis van deze inspectie de meest relevante 20 tot 25 teksten geselecteerd, grondig gelezen, samengevat en besproken in relatie tot de onderzoeksvraag. Bij het bepalen van de relevantie speelt uiteraard de mate van aansluiting tussen de tekst en de onderzoeksvraag een hoofdrol. Indien meerdere teksten goed pasten bij de vraag, is de voorkeur gegeven aan meta-analyses (waarin de resultaten van meerdere onderzoeken worden samengevat en geanalyseerd), literatuurreviews en handboeken. Scripties zijn niet meegenomen in de selectie.

Om vooral de aandacht te vestigen op inzichten die zijn ontstaan na de start van het staatsexamen Nt2 in 1992, is gezocht naar literatuur die sinds het jaar 2000 is verschenen. Daarmee richtten we ons op de literatuur van de afgelopen 25 jaar. Meer recente teksten kregen de voorkeur boven oudere teksten. Bij deelvraag 4, de vraag naar lezen van beeldscherm versus lezen van papier, is de zoekperiode verkleind, omdat ervaring met schermlezen van lezers van 25 jaar geleden waarschijnlijk afwijkt van die van de huidige lezers. Voor het beantwoorden van deze deelvraag is onderzoek van de afgelopen tien jaar geselecteerd. Bij het beantwoorden van vragen over integratief toetsen (het gebruik van gatenteksten) was het nodig de zoekperiode uit te breiden en werden ook enkele teksten van vóór het jaar 2000 geselecteerd.

De gebruikte zoektermen zijn ‘complexe zoektermen’, die gebruik maken van zogeheten Booleaanse operatoren. Hierbij worden zoektermen gecombineerd om zoekresultaten te verfijnen en te specificeren. Omdat de wetenschappelijke literatuur voor het overgrote deel in het Engels wordt gepubliceerd, is gebruik gemaakt van Engelse zoektermen. Tabel 1 geeft een overzicht van de gebruikte zoekopdrachten.

Tabel 1 Zoekopdrachten

	Onderzoeksvraag	**Zoekopdracht(en)**
Hoofdvraag		
	Wat zegt wetenschappelijke onderzoek over ontwikkelingen met betrekking tot inhoud en de beoordeling van taalexamens in een tweede/ vreemde taal in de afgelopen 30 jaar: wat zijn hedendaagse inzichten en tendensen?	“language testing” AND (“developments” OR “recent insights” OR “trends”)
Deelvragen		
1	Wat zijn voor- en nadelen van geïntegreerd toetsen?	(“language testing” OR “language tests”) AND (“integrative” OR “integrated” OR “discrete”)
2	Welke (andere) alternatieven zijn er voor het toetsen van vier vaardigheden (luisteren, spreken, lezen en schrijven) en in welke mate worden deze formats gebruikt bij *highstakes* examens?	(“language testing” OR “language tests”) AND (“format” OR “formats” OR “test design”)
3	Welke alternatieven zijn er voor afname van het spreekexamen via een talenpracticummodel, wat zijn de voor-en nadelen van de alternatieven?	(“testing oral skills” OR “testing speaking”) AND (“language laboratory” OR “computer-based” OR “face-to-face”)

	Onderzoeksvraag	Zoekopdracht(en)
4	Is lezen van een beeldscherm anders dan lezen van papier? Is een leesexamen waarbij van scherm moet worden gelezen valide en betrouwbaar?	“reading from paper” OR “reading from screen”
5	Wat zegt wetenschappelijke onderzoek op het gebied van beoordeling van tweede- en of vreemde-taalexamens, met name op het gebied van de onderdelen Schrijven en Spreken?	(“assessing” OR “testing”) AND (“speaking” OR “writing” OR “productive skills”)
6	Wat is er bekend over positieve beoordelingsmodellen. Worden deze modellen ingezet in *highstakes* examens?	“complexity” AND (“language testing” OR “writing” OR “speaking” OR “productive skills”)
7	Wat is er bekend over dichotoom beoordelen? Zijn er examens die werken met dichotome beoordelingsmodellen voor het beoordelen van open opgaven Spreken en Schrijven? En hoe zitten die beoordelingsmodellen in elkaar?	“rating scales” AND “language testing” AND “dichotomous”
8	Wat is er bekend over holistisch beoordelen? Zijn er andere examens die voor het beoordelen van Schrijven en Spreken werken met (gedeeltelijk) holistische beoordelingsmodellen? En hoe zitten die beoordelingsmodellen in elkaar?	“holistic rating" AND “language testing”
9	Zijn er nog andere beoordelingsmodellen die worden gebruikt bij summatieve taalexamens? Wat zijn voor- en nadelen, wat zijn de ervaringen, hoe functioneren ze in de praktijk?	“rating scales” AND (“assessing” OR “testing”) AND (“speaking” OR “writing” OR “productive skills”) AND (“advantages” OR “disadvantages”)

Om ervoor te zorgen dat de meest recente literatuur niet gemist wordt, zijn in aanvulling op het database-onderzoek de jaargangen van de wetenschappelijke tijdschriften *Language Testing* en *Language Assessment Quarterly* van 2023, 2024 en 2025 doorgenomen. Relevante artikelen zijn toegevoegd aan de leeslijst en in samenhang met de overige geselecteerde teksten besproken.

Daarnaast zijn in sommige gevallen artikelen die werden gevonden in één van de zoekopdrachten verplaatst naar een andere deelvraag of naar de hoofdvraag, indien ze daar beter pasten. Ook zijn enkele zeer relevant geachte teksten toegevoegd die ergens anders werden gevonden dan in de database, bijvoorbeeld in één van de geraadpleegde handboeken (Fulcher & Harding, 2022; Salaberry & Burch, 2021; Winke & Brunfaut, 2021). Bij enkele deelvragen is aanvullend gezocht met een meer specifieke opdracht. Waar dit gebeurd is, is dit vermeld bij de uitwerking van de betreffende deelvraag.

4 Resultaten

4.1 Hedendaagse inzichten en tendensen

Tussen de start van het staatsexamen Nt2 in 1992 en nu, 2025, is de wereld veranderd. Ten eerste hebben technologische ontwikkelingen nieuwe mogelijkheden gecreëerd voor het meten van taalvaardigheid, maar ook nieuwe vragen opgeroepen over validiteit, betrouwbaarheid en toegankelijkheid. Daarnaast hebben veranderingen op het gebied van migratie en veranderde politieke opvattingen over migratie geleid tot een veranderde rol voor veel taaltoetsen, inclusief het staatsexamen Nt2. De hoofdvraag van dit onderzoek luidt daarom: Wat zegt wetenschappelijk onderzoek over ontwikkelingen met betrekking tot inhoud en de beoordeling van taalexamens in een tweede/vreemde taal in de afgelopen 30 jaar: wat zijn hedendaagse inzichten en tendensen?

De literatuur voor de hoofdvraag van dit onderzoek werd gevonden met de zoekopdracht “language testing” AND (“developments” OR “recent insights” OR “trends”). Een overzicht van de gebruikte literatuur is te vinden in Tabel 2. De literatuur reflecteert de toegenomen aandacht voor de validiteit en rechtvaardigheid van het gebruik van toetsen en toetsscores e`n de toegenomen rol van technologie.

Tabel 2 *Bestudeerde literatuur voor de hoofdvraag*

1. Bachman, L. F. (2024). What does language testing have to offer? 1. In: *The Writings of Lyle F. Bachman* (pp. 367-399). Routledge.
2. Brunfaut, T. (2023). Future challenges and opportunities in language testing and assessment: Basic questions and principles at the forefront. *Language Testing, 40*(1), 15-23.
3. Chapelle, C.A. (2021) Validity in Language Assessment. In: Winke, P. & Brunfaut, T. (2021) *The Routledge Handbook of Second Language Acquisition and Language Testing*. pp. 11-20. Routledge / Taylor & Francis.
4. Chapelle, C.A., Lee, H-W. (2022) Conceptions of validity. In: Fulcher, G. & Harding, L. (2022). *The Routledge Handbook of Language Testing*. pp. 17-31. Routledge / Taylor & Francis.
5. Deygers, B. (2017). Assessing highstakes assumptions. A longitudinal mixed-methods study of university entrance language tests, and of the policy that relies on them. KU Leuven.
6. Deygers, B., Van den Branden, K., & Van Gorp, K. (2018). University entrance language tests: A matter of justice. *Language Testing*, 35(4), 449-476.
7. Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing*, 26(1), 123-144.
8. Fulcher, G., Davidson, F., & Kemp, J. (2011) Effective rating scale development for speaking tests: Performance Decision Trees. Language Testing 28(1), 5-29.
9. Hamid, M. O., & Hoang, N. T. (2018). Humanising Language Testing. *Tesl-Ej*, 22(1), 1-20.
10. Harsch, C. & Malone, M.E. (2021). Language proficiency frameworks and scales. In: Winke, P. & Brunfaut, T. (eds.) *The Routledge Handbook of Second Language Acquisition and Language Testing*. Routledge / Taylor & Francis.
11. Heyworth, F. (2006). The common European framework. *ELT Journal*, 60(2), 181-183.
12. Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR)1. *Language Testing*, 29(2), 203-221.
13. Jenkins, J., & Leung, C. (2019). From mythical ‘standard’ to standard reality: The need for alternatives to standardized English language tests. *Language Teaching*, 52(1), 86-110.
14. Jin, Y. (2023). Test-taker insights for language assessment policies and practices. *Language Testing, 40*(1), 193-203.
15. Kane, M. T. (2022). Articulating a validity argument. In: Fulcher, G. & Harding, L. *The Routledge Handbook of Language Testing*. (pp. 32-47). Routledge/ Taylor & Francis.
16. Luxia, Q. (2005). Stakeholders’ conflicting aims undermine the washback function of a highstakes test. Language Testing, 22(2), 142-173.
17. McKay, T.H. & Plonsky, L. (2021) Reliability analyses. Estimating error. In: Winke, P. & Brunfaut, T. (eds.) *The Routledge Handbook of Second Language Acquisition and Language Testing*. pp 468-482. Routledge / Taylor & Francis.

18. McLeod, M. (2021). Investigating test purpose pluralism and test retrofitting in highstakes language proficiency testing. *Canadian Journal for New Scholars in Education/Revue canadienne des jeunes chercheures et chercheurs en éducation, 12*(1), 85-93.
19. Purpura, J. E. (2009). The impact of large-scale and classroom-based language assessments on the individual. *Language Testing Matters, 301*.
20. Rajendram, S., Larson, E., & Sinclair, J. (2019). International graduate students' perspectives on highstakes English tests and the language demands of higher education. *Language and Literacy, 21*(4), 68-92.
21. Salaberry, M. R., & Burch, A. R. (Eds.). (2021). *Assessing speaking in context: Expanding the construct and its applications* (Vol. 149). Multilingual Matters.
22. Sawaki, Y. (2013). Technology in language testing. In *The Routledge handbook of language testing* (pp. 440-451). Routledge.
23. Schildt, L., Deygers, B., & Weideman, A. (2024). Language testers and their place in the policy web. *Language Testing, 41*(2), 338-356.
24. Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Routledge.
25. Suvorov, R., & Hegelheimer, V. (2013). Computer-assisted language testing. *The companion to language assessment, 2*, 594-613.
26. Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281-300.
27. Xi, X., Norris, J. M., Ockey, G. J., Fulcher, G., & Purpura, J. E. (2021). Assessing academic speaking. *Assessing academic English for higher education admissions*, 152-199. Routledge

4.1.1 *Bij de start*

- Communicatieve taalvaardigheid was het dominante paradigma.
- Dat resulteerde in de ontwikkeling van communicatieve taaltoetsen.

Begin jaren '90 begon het veld van taaltoetsing uit te groeien tot een volwassen discipline, na een aantal cruciale ontwikkelingen in de jaren '80 (Bachman, 1991)[1]. Tien jaar eerder, rond 1980, was Oller's *unitary competence hypothesis*, de hypothese die taalvaardigheid beschouwt als één ondeelbaar construct, enige tijd de gangbare opvatting van taalvaardigheid. Deze opvatting werd in de jaren '80 ter discussie gesteld en vervangen door een communicatief paradigma, gebaseerd op inzichten van met name Canale en Swain. Zij introduceerden een meerdimensionaal model van communicatieve competentie, bestaande uit grammaticale, sociolinguïstische, strategische en discourse competentie. Een expliciet doel van de theoretische ontwikkeling was erop gericht om taalvaardigheid beter meetbaar te maken in realistische, communicatieve situaties (Salaberry & Burch, 2021; Xi et al., 2021).

Een tweede inzicht dat was ontstaan door onderzoek in de jaren '80, was dat de methode van toetsing invloed heeft op de meting van taalvaardigheid. Hetzelfde geldt voor het onderwerp van toetstaken. Er ging dus aandacht uit naar de vormgeving van instrumenten voor het meten van taalvaardigheid. Ook de beoordelaars en overeenstemming tussen beoordelaars was een onderwerp dat aandacht had gekregen in onderzoek in de jaren '80, evenals de eigenschappen van kandidaten en hoe die van invloed konden zijn op de performance op een toets. Daarnaast waren er stappen gemaakt op het gebied van methodologie, onder andere op het gebied van item-responstheorie (IRT) (Bachman, 1991).

Herziening van het construct taalvaardigheid en technische ontwikkelingen kwamen tot uiting in communicatieve taaltoetsen die eind jaren '80 en begin jaren '90 werden ontwikkeld. In tegenstelling tot hun voorgangers, die vaak enkel focusten op grammatica en woordenschat, richtten deze toetsen zich op meerdere aspecten, waaronder taalfuncties, samenhang en register. Communicatieve toetsen werken met een informatiekloof, integreren soms meerdere vaardigheden, en nodigen de kandidaat uit een communicatieve taak uit te voeren.

1 De tekst werd opnieuw gepubliceerd in een verzamelbundel in 2024 (en kwam zo terecht bij de zoekresultaten). De tekst beschrijft de inzichten in taaltoetsing die ten tijde van de eerste ontwikkeling van het staatsexamen Nt2 actueel waren en is daarom interessant. Ik gebruikte de oorspronkelijke tekst uit 1991.

4.1.2 *Validiteit*

- Validiteit is een eigenschap van toetsgebruik.
- Het ERK biedt een theoretische onderbouwing van het construct taalvaardigheid.
- Onderzoek naar validiteit van universitaire toelatingstoetsen maakt duidelijk dat de claims van de toetsen niet altijd gerechtvaardigd zijn.
- Als toetsen worden ingezet voor een ander doel dan waarvoor ze oorspronkelijk ontwikkeld zijn, moet de validiteit opnieuw worden onderbouwd.

Validiteit is het meest besproken onderwerp in de literatuur over het toetsen van taalvaardigheid. Het is een veelomvattend concept dat zowel het construct taalvaardigheid, het doel van taaltoetsing, de vorm ervan, en het gebruik van toetsen voor beoogde beslissingen omvat. Huidige opvattingen van validiteit gaan terug op het werk van Messick (1989, geciteerd in Chapelle, 2021). In plaats van validiteit als een eigenschap van de toets (meet de toets wat hij beoogt te meten?) introduceerde Messick het idee van validiteit als een eigenschap van het gebruik van de toets. Validering van een toets vereist zorgvuldige definitie van het te meten construct. Daarnaast omvat validering onderzoek naar en onderbouwing van het gebruik van de toets, de onderliggende waarden van de toets en de sociale implicaties ervan (Chapelle, 2021; Chapelle & Lee, 2022). Een toets en het toetsgebruik moeten worden onderbouwd met argumenten. Dit proces van *argument-based validity* vertrekt vanuit de onderbouwing van de claims die de toetsontwikkelaar op basis van de toets wil maken. Het validiteitsonderzoek heeft dus niet primair betrekking op de toets, als wel op de inferenties die gemaakt worden op basis van de scores (Chapelle, 2021; Kane, 2022).

Theoretische onderbouwing

Op basis van taaltoetsen doen we verschillende soorten uitspraken, bijvoorbeeld over het niveau van de kandidaat of over diens taalvaardigheid in verschillende contexten buiten de toetssituatie. Om de validiteit van het gebruik van testscores te onderzoeken en aan te tonen, is een theoretisch kader nodig dat de relatie legt tussen performance op de toets en het taalgebruik in doelsituaties. Bachman (1991) beschrijft taalvaardigheid als *"the capacity for using the knowledge of language in conjunction with the features of the language use context to create and interpret meaning"* (p 682). ['het vermogen om kennis van de taal te gebruiken, samen met de kenmerken van de situatie waarin de taal wordt gebruikt, om betekenis te creëren en te begrijpen']. Taalgebruik vereist zowel taalkennis als metacognitieve strategieën. Taalkennis valt uiteen in meerdere componenten, waaronder grammaticale, tekstuele, functionele, en sociolinguïstische kennis. Metacognitieve strategieën omvat assessment, doelen stellen, en planning (Bachman, 1991).

Een van de belangrijke ontwikkelingen die grote invloed heeft gehad op de definitie van het construct communicatieve competentie, is de publicatie van het Europees Referentiekader (ERK) (Council of Europe, 2001). Ontwikkeld met als doel communicatie en mobiliteit te bevorderen, biedt het ERK vanuit een sociolinguïstische oriëntatie een uitgebreide actiegerichte beschrijving van taalvaardigheid (Harsch & Malone, 2021; Heyworth, 2006). Het ERK is '*empirically derived*', dat wil zeggen: gebaseerd op expertoordelen over wat leerders op verschillende niveaus kunnen doen. Dat leidt tot kritiek van taalwetenschappers die vinden dat modellen over taalontwikkeling gebaseerd moeten zijn op data van leerders (onder andere Fulcher et al., 2011). In Nederland boden in elk geval Hulstijn et al. (2012) voor de niveaus B1 en B2 enige onderbouwing op basis van prestaties van taalleerders. Zie hierover ook paragraaf 4.7.

Hoewel ontwikkeld als een referentiedocument werd het ERK in de loop van de afgelopen decennia in toenemende mate gebruikt in normatieve contexten. Verschillende auteurs zijn kritisch over deze ontwikkeling (Harsch & Malone, 2021; Weir, 2005). De ERK-niveaus zijn te breed en te algemeen voor dat doel en laten allerlei cruciale zaken onvermeld, zoals de kwaliteit van de prestatie, de context waarin deze plaatsvindt en de eisen die daarmee gemoeid gaan, en de omvang van de woordenschat. Bij het ontwikkelen van highstakes toetsen zou daarom behoefte-onderzoek onder stakeholders moeten plaatsvinden om de validiteit van de toets te waarborgen (Weir, 2005).

Een voorbeeld van zo'n valideringsonderzoek werd uitgevoerd in Vlaanderen (Deygers, 2017; Deygers et al., 2018). Het onderzoek keek naar de validiteit van twee taaltoetsen die anderstalige aspirant-studenten in Vlaanderen konden afleggen om hun taalvaardigheid te bewijzen als eis voor toelating tot een universitaire studie. Het ene examen is de interuniversitaire taaltoets Nederlands voor anderstaligen (ITNA) en de andere het CNaVT-examen educatief-startbekwaam (STRT). Beide toetsen beogen het niveau B2 te meten. Deygers en collega's onderzochten drie claims die op basis van de examens worden gemaakt. De eerste claim is dat de toetsen representatief zijn voor de talige eisen van universitaire opleidingen in Vlaanderen; de tweede claim is dat kandidaten die de taaltoetsen succesvol

afleggen klaar zijn voor de talige eisen die in de opleidingen aan studenten worden gesteld en de derde claim is dat het gebruik van de taaltoetsen als toegangseis voor universitaire studies in Vlaanderen gerechtvaardigd is. De data waren afkomstig van 32 studenten en 24 stafleden. De studenten legden beide toetsen af, deden mee aan interviews en werden gevolgd tijdens het eerste jaar van hun studie. De stafleden van Vlaamse universiteiten namen deel aan een panelgesprek. Met betrekking tot de eerste claim lieten de resultaten zien dat de toetsen niet representatief zijn voor de eisen die in de academische opleidingen worden gesteld. STRT focust op communicatieve adequaatheid, en dat werd belangrijk gevonden door de stafleden. ITNA focust op linguïstische correctheid, wat niet erg relevant werd geacht. De studenten gaven aan dat de opleidingen vrijwel uitsluitend een beroep doen op de receptieve vaardigheden luisteren en lezen maar beide toetsen leggen nadruk op productie: ITNA vereist het slagen voor een spreektoets en bij de geïntegreerde toets STRT is de uitslag uitsluitend gebaseerd op taalproductie. De tweede claim werd weerlegd door aan te tonen dat de eisen die in de academische context werden gesteld aan de receptieve vaardigheden veel hoger waren dan geoperationaliseerd in de toetsen. De studenten werden verondersteld hun taalvaardigheid in het eerste jaar verder uit te breiden maar uit een hertest in het eerste jaar van de studie bleek dat niet te gebeuren. Ten aanzien van de derde claim: een deel van de studenten die voor één van de twee toelatingstoetsen waren gezakt bleken succesvol in de studie en omgekeerd bleken ook studenten die voor de twee toelatingstoetsen slaagden soms grote problemen te ervaren tijdens de opleiding. De auteurs concludeerden dan ook dat de claims waarop het universitaire toelatingsbeleid rust niet gevalideerd konden worden.

In een kwalitatief onderzoek met vijf participanten keken ook Rajendram et al. (2019) naar de perceptie van internationale studenten. Het onderzoek werd uitgevoerd aan een Canadese universiteit. Deze studenten hadden één of beide bekende taalexamens voor het Engels als vreemde taal, IELTS en TOEFL, afgelegd in de context van de toelatingseisen. De algemene aanname is dat de toetsscore op één van deze toetsen een voorspellende waarde heeft voor het functioneren in de studie, al verschillen Canadese universiteiten in de TOEFL of IELTS score die ze vereisen voor toelating. Beide toetsen rapporteren namelijk niet een eenvoudige zak-slaagbeslissing maar een score op een schaal. Uit analyse van de interviewdata bleek dat de toetsen niet goed aansluiten op de taaleisen die gepaard gaan met functioneren in een academische context. Zowel met betrekking tot vorm, als inhoud en taaklengte vonden de participanten de toetstaken niet representatief en ondanks toelating op basis van toetsscore ervoeren ze grote uitdagingen bij het functioneren in de studie.

Retrofit

Bij de ontwikkeling van een toets specificeert de ontwikkelaar zaken als het doel van de toets, de populatie voor wie de toets bestemd is en de kennis en vaardigheden die de toets beoogt te meten. Deze beschrijving is de onderbouwing van de beslissingen die op basis van de toetsscores worden genomen. Vaak wordt een toets in de loop van de tijd aangepast. Dit wordt aangeduid als **retrofit**[2] (Fulcher & Davidson, 2009; McLeod, 2021). Er zijn twee soorten retrofit te onderscheiden, de eerste is *upgrade retrofit*, waarbij de toets wordt verbeterd om beter aan te sluiten bij de oorspronkelijke doelen. De tweede wordt aangeduid als change retrofit, waarbij de toets wordt ingezet voor een ander doel dan dat waarvoor die oorspronkelijk is ontwikkeld. Aangezien toetsdoel en toetsontwerp sterk met elkaar verbonden zijn, betogen de onderzoekers dat retrofit gepaard moet gaan met een stevige evaluatie, waarbij het beoogde gebruik van de toets en de vorm, inhoud en onderliggende assumpties opnieuw worden onderbouwd.

Betrouwbaarheid

Taaltoetsen hebben tot doel de vaardigheid van leerders te schatten op basis van de uitvoering van taken. Daarbij kan **ruis** optreden in allerlei vormen: de vorm van toetstaken, de onderwerpen van de taken, de beoordelaars en de gebruikte beoordelingsmodellen kunnen een bron zijn van construct-irrelevantie variantie en daarmee de nauwkeurigheid van de schatting beïnvloeden. Om claims te kunnen maken over de vaardigheid van kandidaten moet de consistentie van toetsscores kunnen worden aangetoond. Betrouwbaarheid is daarmee een onderdeel van validiteit, maar betrouwbaarheid alleen is niet toereikend om validiteit te kunnen claimen (McKay & Plonsky, 2021). In het huidige onderzoek komt de invloed van taken op betrouwbaarheid aan de orde bij met name deelvraag 4 en de impact van beoordelingsmodellen en beoordelaars op scores komt ruim aan de orde bij de deelvragen 5 tot en met 9. Het onderwerp wordt hier daarom niet uitgebreider besproken.

2 Vetgedrukte begrippen worden toegelicht in de woordenlijst.

4.1.3 Technologie

- In de afgelopen decennia is technologie een grotere rol gaan spelen in de afname van toetsen en in de maatschappij.
- Toekomstige ontwikkelingen lijken vooral gericht op de inzet van AI.

Sinds begin jaren '90 zijn er grote ontwikkelingen geweest op het gebied van technologie. Sawaki (2022) onderscheidde drie fases in de geschiedenis van computer-based testing (**CBT**). CBT's werden ontwikkeld vanaf de jaren '80 en in die tijd werd de computer met name ingezet om de efficiëntie van het toetsen te bevorderen. Veel aandacht ging uit naar de ontwikkeling van adaptieve toetsen (**CAT**). Zo werd bijvoorbeeld de TOEFL als adaptieve toets ontwikkeld. Dit spoor werd echter ook weer verlaten, omdat het ontwikkelen van een CAT erg kostbaar is en het vaststellen van itemparameters lastig. Een tussenvorm die nu nog wel wordt ingezet in highstakes toetsing is een CBT met een min of meer adaptief karakter, waarbij de kandidaat de keuze heeft op welk niveau items worden aangeboden. Dit is het geval bij de taaltoets COPI. Een andere variant wordt ingezet bij DIALANG, waar de kandidaat een zelfbeoordeling uitvoert, die samen met een woordenschattoets bepaalt op welk niveau de toets wordt aangeboden. In de tweede fase die Sawaki onderscheidt, was de aandacht gericht op het benutten van de mogelijkheden van de computer om goede taken te ontwikkelen. Zo is het schrijven van een e-mail meer authentiek op de computer. Ook werden technische mogelijkheden benut om nieuwe taaktypen te ontwikkelen, zoals drag & drop. In deze fase werden CBT's op CD-rom verlaten en werden internet-gebaseerde afnames geïntroduceerd. Dit riep ook nieuwe vragen op, bijvoorbeeld met betrekking tot de veiligheid. Mogelijkheden die nog in de kinderschoenen staan betreffen bijvoorbeeld de inzet van *virtual reality*. De derde, huidige fase wordt gekenmerkt door aandacht voor het integreren van instructie, leren en assessment. Meer recente ontwikkelingen wijzen op het belang van geïntegreerde assessments, als integraal deel van het curriculum, en aangepast aan de behoeftes van individuele leerders. Voorbeelden van toepassing zijn echter nog schaars. Sawaki betoogt dat er meer aandacht nodig is voor het veranderd construct taalvaardigheid in de digitale samenleving. Daarbij zouden we moeten kijken hoe we mogelijkheden van multimedia en VR beter kunnen benutten. Ook moeten er meer inspanningen worden gericht op het versterken van de koppeling tussen leren en assessment.

Volgens Suvorov en Hegelheimer (2014) liggen er drie soorten motieven ten grondslag aan het toetsen via de computer: efficiency, equivalentie en innovatie. Efficiëntie van toetsen wordt verbeterd bijvoorbeeld via automatische scoring, equivalentie is gericht op de vergelijkbaarheid van computertoetsen met papieren toetsen, en innovatie richt zich op de herziening van het construct taalvaardigheid in het digitale tijdperk, waarin taalgebruik immers vaak plaatsvindt via een digitaal medium. Zij beschrijven een aantal highstakes toetsen die gebruik maken van de computer en/of het internet.

- TOEFL iBT, de meest grootschalige en bekende toets voor het Engels als tweede en vreemde taal. Het examen is internet-based, de taken zijn (deels) geïntegreerd en beoordeling vindt plaats door een combinatie van menselijke beoordeling en automatische scoring.
- Bulats online test. Deze toets is gericht op zakelijk Engels. De receptieve vaardigheden worden adaptief getest, het spreekexamen wordt afgenomen via de computer op basis van korte opdrachten en de schrijftoets bestaat uit twee taken. De productieve vaardigheden worden beoordeeld door een menselijke beoordelaar.
- Best Plus Computer-Adaptive Version: deze toets die wordt afgenomen in de VS wordt afgenomen via een standalone computer. De toets toetst luisteren en spreken. De prestaties van kandidaten worden tijdens de afname onmiddellijk beoordeeld en een nieuwe opgave wordt gekozen op basis van deze beoordeling.
- Versant English (voorheen Phone Pass, ontwikkeld door Pearson). De toets is gebaseerd op spraakherkenning en bepaalt het niveau spreekvaardigheid op basis van een serie voorspelbare taken.
- Pearson Test of English: deze toets meet vier vaardigheden, deels in geïntegreerde taken. De toets maakt gebruik van automatische scoring en beoordeling van alle vaardigheden.

In het huidige literatuuronderzoek werd het gebruik van grote taalmodellen (AI) expliciet benoemd als een thema buiten de scope van het onderzoek. In de context van ontwikkelingen op het gebied van technologie moet echter toch -kort- genoemd worden dat grote taalmodellen in toenemende mate een rol spelen bij de beoordeling en afname van toetsen. Zo worden zoals gezegd de onderdelen spreken en schrijven van de bekende toets TOEFL iBT beoordeeld door een menselijke beoordelaar en een automatisch scoringssysteem (Alderson, 2009; Suvorov & Hegelheimer, 2014; Xi, 2007) (Alderson, 2009; Xi, 2007) en biedt de Duolingo English Test, op basis van AI, aan de kandidaten vervolgtaken aan die gebaseerd zijn op de inhoud van diens respons op de eerste taak, waarmee een element van interactie wordt geïntroduceerd in de taak (Runge et al., 2025).

4.1.4 *Het gebruik van toetsen*

- Het terugslageffect van toetsen kan negatief of positief zijn, en kan worden beïnvloed door het belang van de toets.
- Kandidaten ervaren de beslissingen op basis van highstakes taaltoetsen niet altijd als zinvol en fair.
- Er is grote kritiek op het gebruik van toetsen in de context van highstakes beslissingen over verblijfsrecht. Dit type toetsgebruik maakt toetsen tot politieke instrumenten.

Terugslageffect

Toetsen kunnen grote impact hebben op alle betrokkenen. Eén van de vormen van impact is het terugslageffect van de toets. Terugslag is de mate waarin de toets in positieve of negatieve zin bijdraagt aan het behalen van onderwijsdoelen. Naarmate de belangen verbonden met de toets groter zijn, zal ook het terugslageffect van de toets groter zijn. Individuen kunnen op verschillende manieren worden beïnvloed door een toets: ten eerste kan de inhoud van de toets de voorbereiding en het afleggen van het examen beïnvloeden. Ten tweede kan de feedback die kandidaten krijgen op grond van hun prestaties van invloed zijn op hun leerproces. Ten slotte worden zij beïnvloed door de beslissingen die over hen worden genomen op basis van de interpretaties van de testscores (Purpura, 2009). Purpura bepleit daarom het inzetten van formatief assessment, omdat het de regie bij docenten laat en mogelijkheden biedt voor sterkere verbinding van leren en beoordeling.

Terugslageffect kan doelbewust worden ingezet om onderwijsvernieuwing te bewerkstellingen. Dat gebeurde bijvoorbeeld in China, waar grote aantallen kandidaten een universitair toelatingsexamen Engels afleggen, genaamd NMET. De toets is een selectie-instrument voor toelating tot universitair onderwijs. In de jaren ’80 werd de toets herzien met als doel de toen gangbare aanpak van taalonderwijs, gericht op geïsoleerde linguïstische vaardigheden zoals grammatica en woordenschat, te vervangen door een communicatieve aanpak. Onderzoek van Luxia (2005) onder 1388 stakeholders, gebaseerd op vragenlijsten, interviews en lesobservaties, maakte duidelijk dat het beoogde terugslageffect niet werd bereikt. Vanwege de grote belangen die gepaard gingen met het slagen voor de toets moesten toetsontwikkelaars zorgdragen voor hoge betrouwbaarheid, wat onder andere resulteerde in de inzet van een beperkt repertoire van meerkeuze-items. Uit oogpunt van transparantie mochten ze daarbij ook alleen gebruik maken van de woorden uit een gepubliceerde woordenlijst met 2000 woorden, wat gebruik van authentieke teksten uitsloot. Vanwege het grote belang van het (slagen voor het) examen bestonden de lessen voor 70% uit ‘NMET-drilling’. De selectiefunctie van de toets was in deze context dus zo prominent dat er geen ruimte was voor de beoogde sturende functie in termen van onderwijsvernieuwing.

Het perspectief van de kandidaat

Bij het onderzoeken van de validiteit van het toetsgebruik, moet ook de kandidaat worden betrokken (Jin, 2023). Deze groep stakeholders wordt vaak over het hoofd gezien. Jin bepleit meer betrokkenheid van kandidaten bij de ontwikkeling van toetsen en bij de beslissingen die op basis van de toets worden gemaakt. Hamid en Hoang (2018) onderzochten de ervaringen van in totaal 948 kandidaten die IELTS of TOEFL hebben afgelegd. Participanten in dit onderzoek vulden een vragenlijst in en een subgroep deed daarnaast mee aan een interview. Uit de resultaten bleek dat veel kandidaten het afleggen van de toets zagen als een bureaucratische plichtpleging. Ze waren kritisch op het afnemen van een taaltoets als voorwaarde voor studie of verblijf en ze ervoeren dit als onredelijk. Ze hadden het gevoel dat ze in de toets niet hun volledig potentieel kunnen demonstreren en ervoeren de score vaak als een onderschatting van hun vaardigheid. Daar kwam bij dat veel participanten de vragen en taken als erg Westers ervoeren en als niet aansluitend bij hun kennis en interesses. Kandidaten gaven aan dat het afleggen van de toets veel tijd, geld en stress kost. Ook de afname werd als stressvol ervaren, omdat veel deelnemers het gevoel hadden dat ze als potentiële fraudeurs werden bejegend.

Een kritisch perspectief

Beslissingen die worden genomen op basis van toetsing en assessment kunnen grote impact hebben op alle betrokkenen: op individuen, groepen, instituties en op de maatschappij in zijn geheel. Deze impact wordt nog groter als grootschalige toetsen worden gebruikt als de sleutel die deuren kan openen of gesloten houdt, zoals in de context van verblijfsrecht en naturalisatie, of om te beslissen over toelating tot hoger onderwijs. Overheden kunnen toetsen gebruiken als instrument om verandering te bewerkstelligen, bijvoorbeeld om vaardigheden te verhogen (Luxia, 2005; Purpura, 2009). Daarmee zijn toetsontwikkelaars politieke actoren (Schildt et al., 2024; Shohamy, 2006).

Kritiek op gebruik van taaltoetsen als ‘gatekeeper’ is niet alleen recent. Al in 2001 publiceerde Shohamy haar boek ‘The power of tests’ en legde daarmee de basis voor een beweging die de ontwikkeling van grootschalige en

highstakes taaltoetsen zeer kritisch volgde. Verwijzend naar Messick's opvatting van validiteit wees Shohamy (2001) erop dat constructvaliditeit niet ophoudt bij de ontwikkeling van een toets, maar ook het gebruik ervan omvat. Het is de verantwoordelijkheid van de toetsontwikkelaar om de interpretatie en het gebruik van toetsscores te volgen. Shohamy's betoog is gebaseerd op drie onderzoeken uitgevoerd in Israël. De data lieten zien hoe toetsen een sterk en onbedoeld terugslageffect hadden op het onderwijs, hoe op basis van toetsresultaten leerders werden uitgesloten van vervolgonderwijs en hoe toetsen werden gebruikt voor andere beslissingen dan waarvoor ze waren bedoeld. Toetsen zijn op die manier instrumenten waarmee beleid, zoals taalbeleid of migratiebeleid, wordt vormgegeven. Onder verwijzing naar het werk van Shohamy wijzen ook Rajendram et al. (2019) en Jenkins en Leung (2019) erop dat validiteit onder andere gebaseerd is op de rechtvaardigheid van het toetsgebruik. Toetsontwikkelaars bepalen de standaarden, de inhoud, de vorm, de wijze van scoring, de kosten en de rapportage van de uitslag van taaltoetsen, nagenoeg zonder input van kandidaten. In het proces van highstakes beslissingen is de macht niet evenredig verdeeld (Rajendram et al., 2019) en kunnen toetsen fungeren als 'gatekeeper' en daarmee mensen uitsluiten (Jenkins & Leung, 2019).

Highstakes taaltoetsen kunnen dus worden beschouwd als politieke instrumenten. Beleidsmakers geven de toets waarde in de echte wereld door de scores te verbinden met toelating tot opleiding, werk, verblijfsdocumenten of een paspoort (Schildt et al., 2024). In een onderzoek onder toetsontwikkelaars (N=26) uit twintig Europese landen gingen Schildt en collega's na hoe taaltoetsers hun relatie met beleidsmakers en hun impact op het beleid zelf zagen. Er stonden drie vragen centraal in het onderzoek. De eerste vraag was hoe taaltoetsers aankeken tegen het gebruik van toetsen in de context van migratiebeleid. Het onderzoek liet zien dat taaltoetsers zich ervan bewust zijn dat toetsscores gebruikt worden voor politieke doeleinden en daar overwegend genuanceerd over denken, ze proberen het perspectief van de beleidsmakers te begrijpen. Wel maken ze zich zorgen als toetsen worden ingezet in de context van migratie terwijl ze niet primair voor dat doel ontwikkeld waren. De tweede vraag was hoe taaltoetsers hun relatie met beleidsmakers omschrijven. De meeste toetsontwikkelaars hebben geregeld contact met beleidsmakers en vinden een goede relatie essentieel. Ze beïnvloeden beleidsmakers overwegend als hen om advies gevraagd wordt en geven niet graag ongevraagd advies. De derde vraag luidde: hoe zien taaltoetsers hun invloed op de politiek? Taaltoetsers hadden uiteenlopende opvattingen over hun impact, deze verschilde van land tot land.

4.1.5 *De toekomst*

- De basale vragen wat, hoe en waarom zouden leidend moeten zijn in de verdere ontwikkelingen op het gebied van toetsing en examinering.

Brunfaut (2023), tot slot, in een reflectie op de toekomst van taaltoetsing, betoogt dat de toekomstige uitdagingen nog altijd betrekking hebben op drie simpele vragen: wat testen we, hoe testen we en waarom testen we? Ze bepleit meer inspanningen om de validiteit van toetsen te onderbouwen. Met name is meer onderzoek nodig naar de vraag hoe taal wordt gebruikt in de contexten waarvoor we toetsen. Met betrekking tot de hoe-vraag wijst ze vooral op de snel toenemende ontwikkelingen op het gebied van technologie, die mogelijk relevant kunnen zijn voor taaltoetsing. De waarom-vraag wordt traditioneel beantwoord door onderscheid te maken tussen formatieve en summatieve toetsen. Maar op een dieper niveau kunnen aan toetsing politieke overwegingen ten grondslag liggen. Als deel van het validiteitsargument is het belangrijk dat toetsontwikkelaars ethisch gebruik van de toetsen garanderen door inzichtelijk te maken hoe de scores op toetsen zich verhouden tot de beslissingen die op basis van de toets worden genomen. Zie over het proces van valideren bijvoorbeeld ook paragraaf 4.6.4 (p. 67).

4.2 De voor- en nadelen van geïntegreerd toetsen

De eerste deelvraag waarop dit onderzoek focust, luidt: Wat zijn voor- en nadelen van **geïntegreerd toetsen**?
Voor het zoeken naar literatuur over geïntegreerde toetsen is de volgende zoekterm gebruikt: ("Language testing" OR "language tests") AND ("integrative" OR "integrated" OR "discrete"). De zoekopdracht leverde veel relevante informatie op over geïntegreerd toetsen van taalvaardigheid. Er is met name veel onderzoek gedaan naar de geïntegreerde taken in de TOEFL-toets. Een specifieke vorm van geïntegreerde toetsing is task-based assessment (Van den Branden, 2021). Met betrekking tot task-based assessment leverde de zoekopdracht weinig informatie op. Er is daarom aanvullend gezocht met de volgende zoekterm: ("Language testing" OR "language assessment") AND ("task-based") (zoekdatum 31 oktober) en twee teksten zijn toegevoegd (Bachman, 2002; Van Gorp & Deygers, 2013). De teksten waarop dit hoofdstuk gebaseerd is, zijn opgenomen in Tabel 3.

Deze paragraaf bespreekt achtereenvolgens het construct onder geïntegreerde taaltoetsen, de vergelijkbaarheid van geïntegreerde en geïsoleerde taken, en de relatie tussen receptieve en productieve vaardigheden in geïntegreerd assessment. Daarna volgt een paragraaf over de kenmerken van de producten die het resultaat zijn van een geïntegreerde taak, vervolgens komen de kandidaat diens strategieën aan de orde, gevolgd door een paragraaf over task-based language assessment. Het hoofdstuk eindigt met een paragraaf over betrouwbaarheid.

Tabel 3 *Bestudeerde literatuur deelvraag 1*

1. Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453-476.
2. Van den Branden, K. (2022) Measuring task-based performance. In: Winke, P. & Brunfaut, T. (Eds). *The Routledge Handbook of Second Language Acquisition and language Testing*. pp 316-325. Routledge / Taylor & Francis.
3. Cho, Y., Rijmen, F., & Novák, J. (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBT™ integrated writing tasks. *Language Testing, 30*(4), 513-534.
4. Fischer, J., Chouissa, C., Dugovičová, S., & Virkkunen-Fullenwider, A. (2011). *Guidelines for task-based university language testing*. Council of Europe Pub.
5. Frost, K., Clothier, J., Huisman, A., & Wigglesworth, G. (2020). Responding to a TOEFL iBT integrated speaking task: Mapping task demands and test takers' use of stimulus content. *Language Testing, 37*(1), 133-155.
6. Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing, 29*(3), 345-369.
7. Fulcher, G. (2000). The 'communicative' legacy in language testing. *System, 28*(4), 483-497.
8. Garcia Laborda, J., & Fernandez Alvarez, M. (2021). Multilevel language tests: Walking into the land of the unexplored.
9. Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*(4), 507-531.
10. Van Gorp, K., & Deygers, B. (2013). Task-based language assessment. In *The companion to language assessment* (pp. 578-593).
11. Huang, H. T. D., & Hung, S. T. A. (2018). Investigating the strategic behaviors in integrated speaking assessment. *System, 78*, 201-212.
12. Huang, H. T. D., Hung, S. T. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing, 35*(1), 27-49.
13. Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing, 33*(3), 319-340.
14. Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*(2), 131-166.
15. Li, J. (2014). The role of reading and writing in summarization as an integrated task. *Language Testing in Asia, 4*(1), 3.
16. Lewkowicz, J. A. (2000). Authenticity in language testing: some outstanding questions. *Language Testing, 17*(1), 43-64.
17. Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing, 26*(4), 561-587.
18. Plakans, L. (2013). Writing integrated items. In: *The Routledge Handbook of Language Testing* (pp. 263-275). Routledge Taylor & Francis.
19. Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing, 36*(2), 161-179.
20. Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing, 37*(1), 31-53.
21. Shin, S. Y., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores? *Language Testing, 32*(2), 259-281.
22. Wei, W., & Zheng, Y. (2017). An investigation of integrative and independent listening test tasks in a computerised academic English test. *Computer Assisted Language Learning, 30*(8), 864-883.
23. Yang, H. C., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *Tesol Quarterly, 46*(1), 80-103.
24. Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? *Language Testing, 25*(4), 521-551.
25. Zhu, X., Li, X., Yu, G., Cheong, C. M., & Liao, X. (2016). Exploring the relationships between independent listening and listening-reading-writing tasks in Chinese language testing: Toward a better understanding of the construct underlying integrated writing tasks. *Language Assessment Quarterly, 13*(3), 167-185.

4.2.1 *Het construct*

- Geïntegreerde taaltoetsen komen voort uit de communicatieve benadering van taalvaardigheid.
- Ze beogen toetstaken meer authentiek en meer betrouwbaar te maken, met een positief terugslageffect.

Geïntegreerde taaltoetsen zijn toetsen die meer dan één vaardigheid meten. Integratie van vaardigheden past bij moderne benaderingen van taalleren en taalonderwijs, zoals *task-based language teaching* (TBLT) en *content and language integrated learning* (CLIL). Deze benaderingen beschouwen taal holistisch en niet als vier verschillende vaardigheden (Plakans, 2022; Van den Branden, 2021). Beide benaderingen komen voort uit de communicatieve benadering van taalvaardigheid (Fischer et al., 2011). De communicatieve benadering legt sterke nadruk op face-validiteit. Het wordt belangrijk gevonden dat toetstaken authentiek zijn, dus lijken op de situaties waarin taalleerders hun taalvaardigheid nodig hebben (Fulcher, 2000; Lewkowicz, 2000). Task-based assessment legt de nadruk op de uitvoering van de taak, terwijl andere communicatieve benaderingen het construct communicatieve competentie in kaart proberen te brengen (Bachman, 2002; Fischer et al., 2011).

Geïntegreerd toetsen wordt bepleit om drie redenen. De eerste reden is authenticiteit van de toetstaak. Omdat we 'in het echte leven', en zeker in academische settings, vaak meerdere vaardigheden combineren, is het toetsen van gecombineerde vaardigheden meer authentiek (Plakans, 2022; Rukthong & Brunfaut, 2020; Shin & Ewert, 2015). Een tweede argument voor het geïntegreerd toetsen, dat hiermee verband houdt, is het terugslageffect. Geïntegreerde toetsen worden verondersteld te leiden tot een geïntegreerde, en daarmee meer authentieke en nuttigere aanpak van het taalonderwijs (Van den Branden, 2021). Dit zou ook de motivatie van taalleerders versterken (Lewkowicz, 2000). Een derde argument heeft met name betrekking op betrouwbaarheid. In een geïsoleerde spreek- of schrijftaak wordt gebruik gemaakt van een korte prompt en er wordt weinig input gegeven die de kandidaat kan verwerken in de taakuitvoering. De taak doet daarom een groot beroep op de algemene kennis van de kandidaten en die kennis kan de resultaten op de spreek- of schrijftoets sterk bepalen. De geïntegreerde taak wordt daarom beschouwd als meer fair: alle kandidaten hebben immers dezelfde input om mee te werken (H. T. D. Huang et al., 2018; Shin & Ewert, 2015).

Onderzoek naar de validiteit van geïntegreerde taken richt zich op het te meten of gemeten construct, met name op de vraag of de score op de taak alle geïntegreerde vaardigheden reflecteert. Plakans (2022) stelt dat de performance een betere afspiegeling is van de productieve vaardigheid dan van de onderliggende receptieve vaardigheden, al verschillen onderzoeksbevindingen enigszins. Een andere benadering is om niet te kijken naar hoe de verschillende vaardigheden gemeten worden maar in plaats daarvan te focussen op het geïntegreerde construct. Onderzoek van Plakans en collega's laat zien dat een geïntegreerde schrijftaak op basis van leesteksten een beroep doet op overlappende processen lezen en schrijven, zoals het activeren van de woordenschat, gebruik maken van achtergrondkennis, monitoring, herlezen en samenvatten (Plakans, 2022).

Plakans (2022) betoogt dat geïntegreerd toetsen van taalvaardigheid uitdagend is voor toetsontwikkelaars. Ze formuleert de volgende eisen aan geïntegreerde taken:

- De taken zijn gebaseerd op talig bronmateriaal (geen plaatjes of grafiek);
- Het bronmateriaal moet worden verwerkt in geschreven of gesproken tekst;
- De productie moet worden beoordeeld met een geïntegreerde schaal;
- Het oordeel is gebaseerd op de taakuitvoering.

Waar sommige geïntegreerde taaltoetsen focussen op beoordeling van het geïntegreerde construct, zoals de samenvatting die de kandidaat oplevert, gaat een *task-based language assessment* (TBLA) nog een stap verder. Van Gorp en Deygers (2013) definiëren een taak als een functionele activiteit in een specifieke setting waarin een individu taal gebruikt om een doel te bereiken. Van den Branden (2021) wijst erop dat het communicatieve doel ook een niet-linguïstisch doel kan zijn. Een voorbeeld van een taak zou kunnen zijn: soep koken (op basis van een recept, en dus op basis van leesvaardigheid). De kandidaat wordt beoordeeld op het resultaat: de soep.

4.2.2 *Vergelijkbaarheid van geïntegreerde en onafhankelijke taken*

- Scores op geïntegreerde en onafhankelijke schrijf- en spreektaken correleren.
- Onderzoek vindt sterke taakeffecten: moeilijkheid van een taak is een interactie tussen kandidaat en taak, niet een eigenschap van de taak.
- Verschillende taaktypes vullen elkaar aan.

Een relevante vraag in de context van geïntegreerd toetsen is hoe de scores op geïntegreerde taken zich verhouden tot die op **onafhankelijke taken**. Onderzoek heeft laten zien dat de scores over het algemeen hoog correleren, maar onderzoekers vonden wel kwalitatieve verschillen. Zo waren de schrijfproducten in een onafhankelijke schrijftoets langer dan in een geïntegreerde taak en vond ander onderzoek dat in een geïntegreerde taak meer onderwerpen werden uitgewerkt. Onderzoekers vonden geen verschillen in grammaticale accuraatheid (Plakans, 2022).

Een onderzoek naar de validiteit van items van TOEFL iBT werd uitgevoerd door Kyle et al. (2016) en richtte zich op spreekvaardigheid. TOEFL iBT bevat zowel geïntegreerde als onafhankelijke (ook wel aangeduid als geïsoleerde) spreektaken. In de onafhankelijke taak moet de kandidaat spreken op basis van diens eigen ideeën. Een voorbeeld, afkomstig van de TOEFL-website met voorbeeldopgaven[3], is het volgende:

> *Some people think it is more fun to spend time with friends in restaurants or cafes. Others think it is more fun to spend time with friends at home. Which do you think is better? Explain why.*

Bij de geïntegreerde taak moet de kandidaat spreken op basis van gesproken input of op basis van gesproken en geschreven input. De kandidaat leest bijvoorbeeld een korte opiniërende tekst over een onderwerp en luistert vervolgens naar een audiofragment waarin een tegengesteld standpunt aan bod komt. In de spreekuiting moet de kandidaat de standpunten vergelijken en erop reflecteren.

Het doel van het onderzoek was na te gaan of de performance van kandidaten op de geïntegreerde en onafhankelijke taken systematisch verschilt. Het onderzoek, gebaseerd op 240 spreekproducten op basis van onafhankelijke en geïntegreerde taken, vond kleine categorische verschillen op het gebied van lexicon en cohesie tussen de spreekproducten behorend bij de twee taaktypen. Er werden echter grotere verschillen gevonden tussen de individuele taken, aangeduid als **taakeffecten**. Het begrip taakeffecten verwijst naar het verschijnsel dat kenmerken van een toetstaak de prestaties van kandidaten beïnvloeden, los van hun daadwerkelijke taalvaardigheid. Met andere woorden: verschillen in scores kunnen worden veroorzaakt door hoe een taak is ontworpen, en niet alleen door hoe goed iemand de taal beheerst. Er is bijvoorbeeld veel invloed van het onderwerp van de taak, waardoor kandidaten die vertrouwd zijn met het onderwerp beter presteren dan kandidaten die niet veel affiniteit hebben met het onderwerp. Het onderzoek onderstreept het belang om meerdere verschillende taaktypen in te zetten, die betrekking hebben op verschillende domeinen (Kyle et al., 2016).

Onderzoek van Shin en Ewert (2015) keek naar een geïntegreerde lees-schrijftaak, met als doel zicht te krijgen op de relatieve bijdrage van leesvaardigheid, schrijfvaardigheid en beoordelaar op de scores van de kandidaten. De participanten waren 83 studenten Engels als tweede taal, die een toelatingsexamen deden voor een universiteit in de VS. De toelatingstoets bestaat uit de onderdelen lezen (meerkeuze), luisteren (meerkeuze), essay schrijven en interview. Aan deze toets werd een geïntegreerde lees-schrijftaak toegevoegd om de betrouwbaarheid van het taaktype te onderzoeken. Vergelijking tussen de lees-schrijftaak en de afzonderlijke lees- en schrijftoets liet significante, hoge correlaties zien. De onderzoekers concluderen dat de lees-schrijftaak een beroep doet op zowel lezen als schrijven.

Een onderzoek naar taakeffecten was gericht op schrijven. Een schrijftoets heeft over het algemeen weinig taken. Daarom is het van groot belang dat die taken vergelijkbaar zijn. Bij TOEFL iBT bleken de taakeffecten op de geïntegreerde taak sterker dan die op de onafhankelijke taken. Cho et al. (2013) probeerden daarom grip te krijgen op de factoren die moeilijkheid van de geïntegreerde lees-luister-schrijftaak (**RLW**) bepalen. De participanten in dit onderzoek waren 30 native sprekers van het Engels die studeerden aan een universiteit. Hier werd voor gekozen vanwege geheimhouding van de items: de items moesten geheim blijven voor potentiële kandidaten. 107 geïntegreerde RLW items werden geanalyseerd door de participanten. Ze beoordeelden de moeilijkheid van de taken aan de hand van een vragenlijst. Uit de resultaten bleek dat de moeilijkheid van items voornamelijk wordt bepaald door de mate waarin ideeën in de inputteksten helder te onderscheiden zijn en door de complexiteit van de ideeën in de inputteksten.

3 https://www.ets.org/toefl/test-takers/ibt/transcript/speaking-independent-question-1.html

Onderzoek van Wei en Zheng (2017) keek naar de effectiviteit van verschillende taaktypen voor het meten van luistervaardigheid in de Pearson's Test of English Academic (PTE Academic). De test kent elf verschillende luistertaaktypen, waarbij sommige geïntegreerd zijn en sommige onafhankelijk (geïsoleerd). De taaktypen zijn: dictee; zin herhalen; ontbrekend woord selecteren; gesproken tekst samenvatten; incorrect woord markeren; korte vraag beantwoorden; meerkeuze met meerdere antwoorden; meerkeuze met één goed antwoord; lezing navertellen; correcte samenvatting kiezen; gatentekst. De onderzoeksvraag luidde: meten de geïntegreerde en geïsoleerde luistertaken hetzelfde. De data kwamen van 5697 kandidaten. De onderzoekers keken naar itemmoeilijkheid, naar de vraag welke taaktypen de beste voorspellers waren van luistervaardigheid, en naar efficiëntie van de taaktypen. Uit het onderzoek bleek dat alle elf onderzochte itemtypen één onderliggend construct meten. Geïntegreerde items (dictee en zinnen herhalen) waren betere voorspellers van luistervaardigheid dan onafhankelijke luistertaken, waartoe de auteurs meerkeuze-items en gatenteksten rekenen. Samenvatten van gesproken tekst en meerkeuze met meerdere antwoorden (waarbij bijvoorbeeld twee of drie correcte antwoorden moeten worden aangekruist) waren het moeilijkst. Uit combineren van iteminformatie met toetstijd bleek dat dictee het meest efficiënt was: het leverde de meeste informatie per seconde. Ook efficiënt waren het markeren van incorrecte woorden, gatentekst en zinnen herhalen. Meerkeuzevragen met één correct antwoord behoorden tot de minst efficiënte taaktypen.

Ook Zhu et al. (2016) keken naar de rol van luisteren. Ze vergeleken scores op een voor het onderzoeksdoel ontwikkelde onafhankelijke luistertaak met een geïntegreerde luister-lees-schrijftaak. De participanten waren 226 middelbare scholieren in Hong Kong. Uit de resultaten bleek dat de scores op de twee toetsen hoog correleerden. Echter, uit een factoranalyse bleek dat de twee toetsen niet hetzelfde construct meten. Zhu en collega's betogen daarom dat de twee taaktypen elkaar aanvullen.

4.2.3 De relatie tussen receptieve en productieve vaardigheden

- Verschillen in vaardigheid tussen kandidaten komen tot uiting in de verwerking van bronmateriaal.
- Ook hier vindt onderzoek sterke taakeffecten.

In geïntegreerde taken moeten kandidaten ideeën uit de bronteksten op passende wijze verwerken in een geschreven of gesproken tekst. Frost en collega's (2012) onderzochten de manier waarop kandidaten dit bronmateriaal verwerkten in een geïntegreerde luister-spreektaak. Het onderzoek beoogde de vraag te beantwoorden of de verschillen tussen kandidaten in het gebruik van bronmateriaal gerelateerd zijn aan de vaardigheid van de kandidaat. Data waren afkomstig van 50 leerders van het Engels in het VK en de data werden verzameld met behulp van een geïntegreerde spreek-luistertaak ontwikkeld door Oxford University Press. Kandidaten moesten luisteren naar een radiofragment van ongeveer twee minuten, dat vervolgens moest worden samengevat voor een vriend. De onderzoekers gebruikten discourse-analyse om de kwaliteit van de uitingen van de kandidaten te evalueren. Ze vergeleken deze analyse met de beoordeling op basis van het scoringsvoorschrift van de toets. Het beoordelingsvoorschrift kijkt sterk naar de manier waarop de kandidaat hoofd- en bijzaken verwerkt in de uiting. Daarnaast wordt beoordeeld op woordenschat, grammatica, vloeiendheid en uitspraak. Uit de vergelijking van de beoordeling volgens het model met de discourse-analyse bleek dat het model goed onderscheid kan maken tussen vaardige en minder vaardige kandidaten. Het onderzoekt ondersteunt de ontwikkeling van geïntegreerde taken.

Een latere replicatie (Frost et al., 2020) van het onderzoek van Frost et al., (2012) gebruikte twee geïntegreerde taken lees-luister-spreektaken uit TOEFL iBT. Dit onderzoek keek naar de relatie tussen brontekst, de opdracht, en het spreekproduct. Er werden 120 spreekproducten geanalyseerd van kandidaten op drie niveaus: laag, midden en hoog. Uit het onderzoek bleek een duidelijke relatie tussen het taalniveau en de verwerking van de ideeën uit het bronmateriaal in de spreekuiting, waarbij kandidaten met hoge scores meer ideeën in hun spreekproduct wisten te integreren in vergelijking met de lager vaardige kandidaten. Er was echter ook verschil in de verwerking van de twee teksten, met name in de structuur die kandidaten aanbrachten in hun uiting. Dit wijst op taakeffecten en roept daarom vragen op over de generaliseerbaarheid van de scores.

Onderzoek van Rukthong en Brunfaut (2020) had tot doel meer inzicht te krijgen in de rol van luistervaardigheid in een geïntegreerde luister-samenvattaak. In dit onderzoek voerden 72 leerders van het Engels, studerend op een universiteit in het VK, een luister-samenvattaak uit. In een deel van de taken moest de kandidaat een mondelinge samenvatting geven en in een ander deel moest informatie schriftelijk worden samengevat. De studenten vulden na afloop een vragenlijst in, en een subgroep (n=12) nam ook deel aan een naverteltaak. Het onderzoek liet zien dat studenten een groot aantal cognitieve processen en metacognitieve strategieën gebruiken om de auditieve input

te verwerken. De data lieten ook zien dat vaardige en minder vaardige leerders dat op deels verschillende manieren deden en dat dit zichtbaar werd in de mondelinge en geschreven samenvattingen. De onderzoekers concluderen dat daarom luistervaardigheid een cruciale rol speelt in een samenvattaak en impact heeft op de kwaliteit van het spreek- of schrijfproduct.

Li (2014) keek naar de relatieve bijdrage van lezen en schrijven op samenvatten. Participanten waren 64 studenten Engels als vreemde taal op een Chinese universiteit. De studenten legden een onafhankelijke lees- en schrijftoets af en vatten twee teksten samen. Ook vulden ze een vragenlijst in en namen ze deel in hardop denk-sessies. De samenvattingen werden beoordeeld aan de hand van een beoordelingsmodel. De onderzoekers keken naar de strategieën die de participanten gebruikten bij het lezen en schrijven ten behoeve van de samenvatting. In totaal werden zeventig strategieën geïdentificeerd. De gevonden leesstrategieën waren met name bottom-up strategieën voor het decoderen en het begrijpen van de tekst op woord- en zinsniveau. Bij het schrijven van de samenvatting leunden de studenten sterk op de brontekst. De scores op de samenvattingstaak correleerden hoog met die op de schrijftoets maar minder met die op de leestoets. De conclusie van dit onderzoek luidt dat zowel lezen als schrijven belangrijk is bij het samenvatten, waarbij schrijven de grootste bijdrage levert.

Onderzoek van G. Yu (2008) keek naar de impact van de keuze voor eerste en tweede taal op de samenvatting. Participanten waren 157 gevorderde studenten Engels in China. Voor de samenvattingstaak werden drie teksten gekozen. Studenten vatten elk één van die teksten samen, in het Engels en in het Chinees. De samenvattingen werden beoordeeld aan de hand van een beoordelingsmodel, op basis van expertsamenvattingen door native speakers. Ook werd een vragenlijst afgenomen. Uit het onderzoek bleek dat de kandidaten hogere scores behaalden bij het samenvatten in het Engels. De samenvattingen in het Chinees waren beduidend langer. Studenten gaven de voorkeur aan het samenvatten in het Engels. Ze konden op die manier woorden en zinnen uit de tekst gebruiken in hun samenvatting. Ook vonden ze niet alle concepten uit de teksten goed vertaalbaar.

4.2.4 Kenmerken van de producten

- CAF-maten kunnen de scores op een geïntegreerde taak voorspellen.

Bij de toename van geïntegreerd toetsen was onderzoek nodig naar de validiteitsclaim: wat betekent de score op de toets? Onderzoek van Plakans et al. (2019) had tot doel de relatie te verkennen tussen scores op de TOEFL iBT geïntegreerde lees-luister-schrijftaak (RLW) en de linguïstische kenmerken van het schrijfproduct. Ze gebruikten daarvoor **CAF**-maten. CAF is een afkorting van *Complexity, Accuracy, Fluency*. Deze maten worden gebruikt om taalontwikkeling en de kwaliteit van productie te beschrijven in linguïstische termen. Zie hierover ook paragraaf 4.7. De onderzoeksvragen luidden: voorspelt CAF de score op een geïntegreerde taak, en welke CAF-maten zijn het meest voorspellend? CAF werd als volgt geoperationaliseerd: voor complexiteit werd de gemiddelde lengte van **T-units** (vergelijkbaar met zinnen) bepaald, voor accuraatheid werden grammaticale fouten geteld, verdeeld in syntactische en morfologische fouten. Als maat voor vloeiendheid werd de tekstlengte genomen. Uit de resultaten bleek dat vloeiendheid en complexiteit toenamen met de score, terwijl het aantal grammaticale fouten afnam met de score. Essays van hoger scorende kandidaten zijn dus langer en complexer en met minder fouten. Het onderzoek liet ook zien dat CAF-maten de score op de taak kunnen voorspellen, met vloeiendheid (tekstlengte) als beste voorspeller.

4.2.5 De kandidaat

- Kandidaten gebruiken een groot aantal strategieën bij het uitvoeren van een geïntegreerde taak.

Eén van de aannames van geïntegreerd toetsen is dat het motivatie-verhogend is als de toetstaken lijken op taken die kandidaten in de doelsituatie (de situatie waarvoor de leerder de nieuwe taal leert) moeten vervullen (Plakans, 2022). Deze taken worden meer authentiek geacht. Lewkowicz (2000) onderzocht het concept authenticiteit en ging na hoe belangrijk dit is voor taalleerders, in een kwalitatief onderzoek onder 72 studenten Engels als vreemde taal in Hong Kong. De kandidaten legden achtereenvolgens een niet-authentieke toets af met 90 meerkeuzevragen, gericht op grammatica, woordenschat, lezen en schrijven, en een geïntegreerde performance test. Daarna vulden ze een vragenlijst in met betrekking tot hun perceptie van de twee toetsen. De vragen waren gericht op de perceptie van toetsdoelen, zelfbeoordeling, de mate waarin de taakuitvoering de gepercipieerde vaardigheid reflecteert, en de voor- en nadelen van de taaktypen. Ook werd gevraagd welke toets de beste indicatie gaf van de vaardigheid van de kandidaat en welke beter aansloot op het gevolgde onderwijs. Uit de resultaten bleek dat studenten niet veel waarde hechtten aan authenticiteit. Met name de schrijflast in de authentieke taak werd negatief beoordeeld. Meer vaardige

kandidaten waren wel positiever over de geïntegreerde taak. De onderzoeker vermoedde dat de niet-authentieke toets beter aansloot op het gevolgde onderwijs.

Yang en Plakans (2012) onderzochten de strategieën die kandidaten gebruiken bij het uitvoeren van een geïntegreerde luister-lees-schrijftaak en de relatie van strategiegebruik met de scores op de toets. De participanten in dit onderzoek waren 161 leerders van het Engels op een Amerikaanse universiteit. De studenten vulden een voor dit onderzoek ontwikkelde vragenlijst in over hun aanpak van een geïntegreerde schrijftaak. De vragenlijst boorde drie type strategieën aan: zelfregulatiestrategieën, discoursesynthesestrategieën, en toetsstrategieën. De zelfregulatiestrategieën hadden een positieve relatie met de beide andere type strategieën. De discoursesynthesestrategieën hadden een positieve relatie met de toetsresultaten, terwijl de toets-strategieën (zoals het inprenten van verschillende templates voorafgaand aan de toets of het overschrijven van zinnen uit de brontekst) een negatieve relatie hadden met de scores. De resultaten laten zien dat een geïntegreerde schrijftaak een meervoudig concept is, dat verschillende strategieën activeert.

Huang en Hung (2018) keken naar de strategieën die kandidaten gebruikten bij het uitvoeren van een geïntegreerde spreektoets. Ze gebruikten twee voor dit doel ontwikkelde geïntegreerde spreektaken en legden die voor aan 177 leerders van het Engels in Taiwan. Uit het onderzoek bleek dat strategische competentie deel uitmaakt van taalvaardigheid. Factoranalyse liet drie typen strategieën zien, te weten discourse synthese, cognitieve strategieën en communicatiestrategieën. De onderzoekers concludeerden dat een geïntegreerd assessment een uniek construct meet en dat het onderzoek de validiteit van de geïntegreerde spraaktaak onderbouwt.

4.2.6 *Task-based language assessment*

- Task-based assessment stelt uitvoeren van taken in contexten centraal.
- Contextgebondenheid versterkt mogelijk taakeffecten.

In **task-based language assessment (TBLA)** is de taak de focus van toetsing. De taak focust op taalgebruik en interactie in communicatieve contexten en beoogt studenten zo voor te bereiden op de toekomst (Fischer et al., 2011). De principes van TBLA liggen ten grondslag aan het CNaVT (Van Gorp & Deygers, 2013). Het CNaVT beoogt te voorzien in een valide en betrouwbare meting van taalvaardigheid en wil tevens een positief terugslageffect bewerkstelligen. Het CNaVT is gebaseerd op taken in specifieke contexten. Relevante contexten werden geïdentificeerd op basis van een behoefteanalyse onder kandidaten. Er werden zes profielen onderscheiden

(Van Gorp & Deygers, 2013)[4]. De beoordelingsvoorschriften zijn afgeleid van de context-specifieke eisen en zowel inhoudelijk als linguïstisch van aard.

In een kritische reflectie op task-based assessment stelde Bachman (2002) dat TBLA geen heldere antwoorden geeft op de vragen: wat willen we toetsen en hoe rechtvaardigen we claims over de vaardigheid van de kandidaat op basis van het assessment. Volgens de task-based benadering omvat taalvaardigheid zowel de vaardigheid als het vermogen tot het gebruik van de taal in contexten. Aangezien taken als context-specifiek worden beschouwd, is het moeilijk inferenties te maken van één taak naar de andere. Met andere woorden: als de kandidaat taak x kan uitvoeren, dan kunnen we daaraan geen conclusies verbinden over diens vermogen om ook taak y tot een goed einde te brengen. Sommige TBLA-experts zeggen ook dat het maken van inferenties niet de bedoeling is: het assessment meet of de kandidaat de assessmenttaak kan uitvoeren en conclusies die verder reiken dan de taakuitvoering zijn niet gerechtvaardigd. Uiteraard is dat in de ogen van Bachman problematisch: een taaltoets wordt normaal gesproken afgenomen met het oog op het trekken van conclusies die verder reiken dan de toets, bijvoorbeeld dat de kandidaat voldoende taalvaardig is voor het volgen van een opleiding. Bachman stelt dat TBLA in elk geval moet proberen te komen tot heldere definities van het gemeten construct. Een tweede probleem dat Bachman aankaart is de moeilijkheid van taken. Hij wijst erop dat moeilijkheid niet een inherent kenmerk is van een taak, maar een interactie tussen kandidaat en taak. Vaak zijn taken in een TBLA niet categorisch moeilijk of makkelijk maar is er sprake van taakeffecten. Onderzoek dat gericht is op het identificeren van taakkenmerken die moeilijkheid voorspellen is weinig succesvol geweest.

4 Een korte inspectie van de CNaVT website leert dat er in oktober 2025 sprake is van vijf examens (maatschappelijk informeel, maatschappelijk formeel, zakelijk professioneel, educatief startbekwaam en educatief professioneel), op vier niveaus (A2-C1) en in drie domeinen (maatschappelijk, professioneel, educatief). Bron: www.cnavt.org

4.2.7 Betrouwbaarheid

- Scores op onafhankelijke en geïntegreerde taken correleren hoog.
- Betrouwbaarheid kan het meest effectief worden verhoogd door inzetten van meer taken.
- Kennis van het onderwerp heeft impact op de prestaties op een spreektaak in zowel geïntegreerde als onafhankelijke taken.

Eén van de problemen met geïntegreerd toetsen is dat het niet altijd duidelijk is hoe een score tot stand komt. Het product van een geïntegreerde taak is een spreek- of schrijfproduct. Er is een menselijke beoordelaar nodig om het product te beoordelen, en de beoordelaar is een bron van ruis: de score kan immers het gevolg zijn van mild- of strengheid of inconsistentie van de beoordelaar (Lee, 2006). Ook de taak zelf kan een bron van ruis zijn omdat kandidaten niet consistent zijn in hun interactie met taken: niet elke taak is voor elke kandidaat even makkelijk of moeilijk (Gebril, 2009; Lee, 2006). Dat kan komen doordat het taaktype of het onderwerp van de taak de kandidaat beter of minder goed ligt. Fair toetsen vergt daarom inzicht in de bronnen van meetfout, of ruis. **Generaliseerbaarheidstheorie** (ook: **G-theorie**) wordt veel gebruikt bij dit type onderzoek. G-theorie stelt de onderzoeker in staat verschillende bronnen van meetfout te schatten. Het geeft een antwoord op de vraag of verschillen tussen kandidaten gebaseerd zijn op verschillen in vaardigheid of bijvoorbeeld het gevolg van de taak of de beoordelaar. Door taken of beoordelaars toe te voegen kan de meetfout worden verkleind tot een niveau dat acceptabel wordt geacht. **Dependability**, in het Nederlands ook wel **generaliseerbaarheid**, is vergelijkbaar met betrouwbaarheid in klassieke toets-theorie; het verwijst naar de mate waarin de score generaliseerbaar is, aangenomen dat er altijd sprake is van een meetfout.

In de context van de introductie van nieuwe taaktypen in de TOEFL iBT spreektoets onderzocht Lee (2006) de generaliseerbaarheid van de scores. Het doel was te komen tot een optimaal aantal taken, de optimale samenstelling van de spreektoets, en een optimaal (ook in economische zin) aantal beoordelaars. De onderzoeksvragen luidden:

- Wat is de impact van toename van het aantal taken?
- Wat is de impact van vergroten van het aantal beoordelaars?
- Zijn de scores van de verschillende spreekonderdelen voldoende samenhangend om het verstrekken van één score te rechtvaardigen?
- Welke combinatie van taaktypen is optimaal?

De participanten waren 261 leerders van het Engels als tweede en vreemde taal die deelnamen aan een proefafname. Er waren in totaal 13 taken bij het onderzoek betrokken, vier geïntegreerde luister-spreektaken, twee geïntegreerde lees-spreektaken en vijf onafhankelijke spreektaken. Elke prestatie werd beoordeeld door twee beoordelaars. Uit de resultaten bleek dat de meeste ruis (error variantie) het gevolg was van de taak; er was weinig ruis als gevolg van de beoordelaar (minder dan 2%). De tweede beoordelaar droeg dus weinig bij aan de betrouwbaarheid van de beoordeling. Het vergroten van het aantal taken had relatief veel impact op de generaliseerbaarheid van de score. Vijf taken waren optimaal, en toevoegen van meer taken had weinig impact op de betrouwbaarheid. De verschillende taaktypen correleerden hoog, wat het verstrekken van een gecombineerde score rechtvaardigt. Het toevoegen van een geïntegreerde luister-spreektaak leidde tot iets hogere betrouwbaarheid in vergelijking met toevoegen van de andere taaktypen.

Ook Gebril (2009) keek naar de generaliseerbaarheid van geïntegreerde taken. In dit onderzoek betrof het geïntegreerde lees-schrijftaken. In deze studie werden twee geïntegreerde taken en twee onafhankelijke taken ingezet bij 115 leerders van het Engels in Egypte. De schrijfproducten werden gescoord door drie getrainde beoordelaars, met behulp van de TOEFL-beoordelingsmodellen voor schrijfvaardigheid. Ook dit onderzoek keek naar het percentage variantie dat kan worden toegeschreven aan leerder, taak en beoordelaar, en daarnaast naar het effect van toevoegen van taken en beoordelaars. De resultaten lieten sterke taakeffecten zien: kandidaten reageerden verschillend op de verschillende taken en niet voor iedereen was dezelfde taak makkelijk of juist moeilijk. Om de meetfout te verkleinen had het toevoegen van meer taken (hier: twee in plaats van één) een groot effect. Er werd ook effect gevonden van het toevoegen van een beoordelaar, maar dit effect was veel kleiner dan het effect van het toevoegen van een extra taak. Er werd geen verschil gevonden in de generaliseerbaarheid van de geïntegreerde en onafhankelijke taken. De generaliseerbaarheid (betrouwbaarheid) was vergelijkbaar.

Het eerdergenoemde onderzoek van Shin en Ewert (2015), dat keek naar een geïntegreerde lees-schrijftaak, onderzocht met behulp van G-theorie ook de impact van de beoordelaar op de scores van de kandidaten. De verschillen tussen beoordelaars veroorzaakten in dit onderzoek relatief weinig ruis. Met inzet van twee beoordelaars werd betrouwbaarheid boven .80 bereikt. Een goed, op kandidaatprestaties gebaseerd beoordelingsvoorschrift werd daarbij belangrijk geacht.

Een betrouwbare toets meet dus taalvaardigheid van de kandidaat en meet niet onbedoeld andere zaken mee. Eén van de mogelijke bronnen van ruis is, en een mogelijke verklaring voor taakeffecten, is voorkennis. Een beoogd voordeel van geïntegreerd toetsen is dat het de verschillen tussen kandidaten in voorkennis zou wegnemen of sterk verkleinen; de inputtekst geeft immers de informatie die moet worden verwerkt in het schrijf- of spreekproduct. Onderzoek van Huang et al. (2018) keek naar de rol van voorkennis bij geïntegreerde en onafhankelijke spreektaken. De onderzoeksvragen luidden: wat is de relatie tussen onderwerpkennis en performance op een geïntegreerde spreektaak? En is er verschil in impact van onderwerpkennis in een geïntegreerde of onafhankelijke spreektaak? De onderzoekers gebruikten vier geïntegreerde taken uit de TOEFL iBT voorbeeldopgaven. De geïntegreerde TOEFL iBT-taken bestaan uit een leestekstje waarvoor de kandidaat 45 seconden leestijd krijgt, een geluidsfragment van 90 seconden, daarna 30 seconden voorbereidingstijd en 60 seconden spreektijd. Daarbij ontwikkelden en valideerden ze kennistoetsen over de betreffende onderwerpen. Ook ontwikkelden ze vier vergelijkbare onafhankelijke taken. De participanten waren 773 studenten Engels als vreemde taal aan Taiwanese universiteiten. 421 studenten participeerden in de validering van de kennistoetsen en 352 deden mee aan het eigenlijke experiment. Uit de resultaten bleek dat onderwerpkennis sterke invloed had op de spreekprestatie in zowel de een geïntegreerde spreektaak als in een onafhankelijke spreektaak.

4.2.8 Kritische reflectie

In dit hoofdstuk stond de vraag naar de voor- en nadelen van geïntegreerd toetsen centraal. De beoogde voordelen die de literatuur noemt zijn: vergroten van de authenticiteit, verbeteren van het terugslageffect en vergroten van de betrouwbaarheid van de toets doordat de input voor de productieve taak wordt aangeboden, en kandidaten dus over een gelijke hoeveelheid voorkennis beschikken.

Het argument dat we in het echte leven vaardigheden combineren is overtuigend, en maakt de integratie van taalvaardigheden in een taaltoets aantrekkelijk. TOEFL iBT is een grootschalig highstakes examen dat gebruik maakt van geïntegreerde taken en er is veel onderzoek gedaan naar het functioneren ervan. De TOEFL lees-luister-spreektaak bestaat uit een leestekstje waarvoor de kandidaat 45 seconden leestijd krijgt, een geluidsfragment van 90 seconden, daarna 30 seconden voorbereidingstijd en 60 seconden spreektijd. Deze conceptualisering van authenticiteit is niet overtuigend. Onderzoek van Rajendram et al. (2019), besproken bij de hoofdvraag, ondersteunt deze opvatting: uit dit onderzoek bleek dat toetsen zoals TOEFL iBT niet goed aansluiten op de taaleisen in een academische context.

Verschillende onderzoeken droegen op verschillende manieren bij aan een antwoord op de vraag naar de betrouwbaarheid van geïntegreerde toetsen. Dit hoofdstuk vatte verschillende onderzoeken samen die vertrokken vanuit verschillende onderzoeksvragen. Of deze onderzoeken er nu expliciet op focusten of niet: veel onderzoekers vonden sterke taakeffecten. Dat wil zeggen dat een specifieke taak voor de ene kandidaat veel moeilijker kan zijn dan de andere, ook als de kandidaten in principe hetzelfde vaardigheidsniveau hebben. Taakeffecten zijn dus een vorm van ruis en doen afbreuk aan de betrouwbaarheid van de toets. Taakeffecten worden in elk geval ten dele veroorzaakt door de voorkennis van kandidaten over het onderwerp van de taak (Huang et al., 2018). Onderzoek liet zien dat deze taakeffecten even sterk spelen in geïntegreerde taken als in onafhankelijke taaltaken. Hoewel de voorstanders van geïntegreerd toetsen claimen dat input de kansen van kandidaten meer gelijk maakt, ondersteunt de literatuur deze claim niet. Om de impact van taakeffecten op scores te verkleinen, is het nodig een voldoende groot aantal taken aan te bieden.

Een andere mogelijke bron van ruis is de beoordelaar. Een geïntegreerde taak resulteert altijd in een spreek- of schrijfproduct dat doorgaans door een menselijke beoordelaar wordt beoordeeld. Onderzoekers keken daarom naar de invloed van respectievelijk taak en beoordelaar op de generaliseerbaarheid van scores (Gebril, 2009; Lee, 2006). Met behulp van generaliseerbaarheidstheorie kan met dit type onderzoek zowel het optimale aantal taken en beoordelaars worden bepaald. Uit de resultaten bleek dat de inzet van meer beoordelaars relatief weinig bijdraagt aan het vergroten van de betrouwbaarheid van de beoordeling, in elk geval veel minder dan de inzet van een

voldoende groot aantal taken. Omdat bij het staatsexamen Nt2 zowel de spreek- en de schrijfexamens een ruime hoeveelheid opgaven én zeer veel beoordelaars kennen, lijkt het de moeite waard de efficiëntie daarvan nader te laten onderzoeken.

Ander onderzoek dat wellicht nadere reflectie verdient, kwam van Wei en Zheng (2017). Dit onderzoek keek naar verschillende itemtypes voor het meten van luistervaardigheid. Uit dit onderzoek bleek dat de meerkeuzevraag met één juist antwoord één van de minst efficiënte itemtypes is. Tegelijk is dit het enige itemtype dat in het staatsexamen Nt2 wordt gebruikt voor het meten van luistervaardigheid. Nader onderzoek naar effectiviteit van taaktypen voor het meten van luistervaardigheid is daarom te overwegen.

4.3 Alternatieven voor het toetsen van vier vaardigheden

De tweede deelvraag luidt: Welke (andere) alternatieven zijn er voor het toetsen van vier vaardigheden (luisteren, spreken, lezen en schrijven) en in welke mate worden deze formats gebruikt bij *highstakes examens*? Deze deelvraag heeft enige overlap met de eerste. De termen '**integratief**' en '**geïntegreerd**' verwijzen naar twee verschillende benaderingen in het toetsen van taalvaardigheid. De vorige paragraaf besprak de integratie van meerdere vaardigheden (luisteren, lezen, spreken, schrijven) in één functionele assessmenttaak, aangeduid als geïntegreerd toetsen. Deze paragraaf bespreekt andere vormen van integratie.

De zoekterm die voor de huidige deelvraag werd gebruikt luidde ("Language testing" OR "language tests") AND ("format" OR "formats" OR "test design"). Met deze zoekopdracht werden veel generieke teksten gevonden die te weinig specifiek waren in de context van deze deelvraag. Enkele van deze teksten zijn ondergebracht bij de hoofdvraag. Ook werden teksten gevonden die al eerder waren gebruikt voor het beantwoorden van de eerste deelvraag. Wat overbleef was een kleine set relevante teksten, over integratief toetsen en toetsen van woordenschat, over meten van meerdere niveaus met één instrument, en over alternatief assessment. Bij deze drie onderwerpen is gericht gezocht naar aanvullende literatuur. De zoekopdrachten die daarbij zijn gebruikt zijn: "cloze test" (hierover is weinig recente literatuur, daarom is ook literatuur van vóór het jaar 2000 bekeken); "WAF-test"; "language assessment" AND "alternative assessment"; "high-takes language assessment" AND "portfolio"; "multilevel language tests" AND "highstakes"; "duolingo language test". De teksten waarop dit hoofdstuk is gebaseerd zijn opgenomen in Tabel 4.

Tabel 4 *Bestudeerde literatuur deelvraag 2*

1. Abbott, M. L., Lee, K. K., & Ricioppo, S. (2021). Does portfolio-based language assessment align with learning-oriented assessment? Evidence from literacy learners and their instructors. *Canadian Journal of Applied Linguistics, 24*(2), 260-285.
2. Abbott, M., & Lee, K. K. (2025). Lessons Learned From a Highstakes Portfolio-Based Language Assessment Initiative for LESLLA Learners. *LESLLA Symposium Proceedings, 20*(1), 35–51.
3. Alderson, J. C. (2009). Test review: Test of English as a foreign language™: Internet-based test (TOEFL iBT®). *Language Testing, 26*(4), 621-631.
4. Bachman, L.F. (1982). The trait structure of cloze test scores. *Tesol Quarterly, 16*(1), 61-70.
5. Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671–704.
6. Banerjee, J., Lestari, S. B., & Rossi, O. (2021). Choosing test format and task types. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing*. Routledge/Taylor & Francis.
7. Green, A. (2006). Watching for washback: Observing the influence of the International English Language Testing System academic writing test in the classroom. *Language Assessment Quarterly, 3*(4), 333-368.
8. Ghafournia, N. (2015). Standard assessments: Merits and demerits and the alternative assessments. *Asian Social Science, 11*(13), 166.
9. Hidri, S. (2018). Discrete point and integrative testing. *The TESOL encyclopedia of English language teaching. John Wiley & Sons*.
10. Jenkins, J., & Leung, C. (2019). From mythical 'standard' to standard reality: The need for alternatives to standardized English language tests. *Language Teaching, 52*(1), 86-110.
11. Kühn, B., & Pérez Cavana, M.L. (2012). Perspectives from the European Language Portfolio. Learner autonomy and self-assessment. Routledge.

12. Lynch, B. (2001). The ethical potential of alternative language assessment. *Experimenting with uncertainty: Essays in honour of Alan Davies*, 228-239.
13. Norris, J. M. (2012) Purposeful Language Assessment: Selecting the right alternative test. *English Teaching Forum 38* (3)
14. Qian, D. D., & Schedl, M. (2004) Evaluation of an indepth vocabulary knowledge measure for assessing reading performance. *Language Testing 21*(1), p.28-52
15. Read, J. (2022). Test review: The international English language testing system (IELTS). *Language testing, 39*(4), 679-694.
16. Sheybani, E., & Zeraatpishe, M. (2018). On the dimensionality of reading comprehension tests composed of text comprehension items and cloze test items. *International Journal of Language Testing, 8*(1), 12-26.
17. Schmitt, N., Ng, J. W. C., & Garras, J. (2011). The word associates format: Validation evidence. *Language Testing, 28*(1), 105-126.
18. Wagner, E. (2020). Duolingo English test, revised version July 2019. *Language Assessment Quarterly, 17*(3), 300-315.
19. Zhang, D., & Koda, K. (2017). Assessing L2 vocabulary depth with word associates format tests: issues, findings, and suggestions. *Asian-Pacific Journal of Second and Foreign Language Education, 2*(1), 1.

4.3.1 Cloze tests

- De cloze test wordt tegenwoordig niet meer geschikt geacht voor het meten van taalvaardigheid.

Onder een integratieve test wordt een gatentekst of cloze test verstaan. Een cloze test vereist het integreren van linguïstische, tekstuele en wereldkennis (Hidri, 2018). Volgens Oller (1979, geciteerd in Hidri, 2018) vraagt een cloze test om taalgebruik, niet alleen om taalkennis. De cloze test was eind jaren '70 enige tijd populair maar verdween onder de invloed van de communicatieve benadering (Bachman, 1991).

Het theoretisch model dat dat ten grondslag lag aan de ontwikkeling van cloze tests was Oller's *unitary trait hypothesis*. Volgens deze theorie is er één onderliggende vaardigheid voor taal (Bachman, 1991). Oller reageerde met deze theorie op de destijds gangbare praktijken van discrete-point testing waarbij de nadruk lag op taalkennis, met name grammatica en woordenschat. In plaats daarvan pleitte hij voor integratieve tests die taalgebruik in langere, betekenisvolle contexten vereisten (Wagner, 2022).

Gatenteksten en C-teksten (een variant waarbij één of meer beginletters van het in te vullen woord worden gegeven) zijn controversieel omdat niet duidelijk is wat ze precies meten. Ze kunnen echter een indicatie geven van het niveau van de taalleerder en worden om die reden nog wel gebruikt in intaketoetsen (Banerjee et al., 2021). Hidri (2018) merkt op dat de face-validiteit van de cloze test laag is.

Een debat tussen voor- en tegenstanders van de cloze test betrof de vraag of deze test in staat is meer te meten dan vaardigheid op zinsniveau. Bachman (1982), onderzocht dit met behulp van een factoranalyse naar de verschillende vaardigheden die kandidaten inzetten bij het maken van een cloze test. Hij identificeerde drie onderliggende competenties: syntactische, verbindende (*cohesive*) en strategische competentie. Het onderzoek bevestigde het bestaan van een onderliggende algemene vaardigheid. Later onderzoek gebruikte echter juist ook factoranalyses om aan te tonen dat taalvaardigheid meerdimensionaal was. Deze onderzoeken weerlegden de unitary trait hypothesis (Bachman, 1991).

Hoewel de cloze test niet in lijn is met hedendaagse modellen van taalvaardigheid en tegenwoordig minder vaak wordt gebruikt, komt deze nog wel voor in bepaalde highstakes contexten, waaronder een universitair toelatingsexamen in Iran. Sheybani en Zeraatpishe (2018) vergeleken met behulp van een factoranalyse de cloze test met een leestaak met meerkeuze tekstbegripsvragen. De data waren afkomstig van 1024 kandidaten. Hoewel de moeilijkheid van beide taken vergelijkbaar was, bleken de twee onderdelen niet dezelfde onderliggende vaardigheid te meten. Op basis van deze bevindingen trekken ook deze auteurs de validiteit van de cloze test in twijfel.

4.3.2 Word Associates Format (WAF)

- Het gebruik van het Word Associates Format in taaltoetsing is controversieel.

Woordenschat is een essentiële voorwaarde van communicatieve competentie en de omvang van de woordenschat wordt in verband gebracht met het niveau van taalvaardigheid (Qian & Schedl, 2004). Woordenschat kan worden gedefinieerd als enerzijds het aantal woorden dat de taalgebruiker beheerst, en anderzijds de kwaliteit van de kennis, dus hoe goed de woorden worden beheerst. Een toetsvorm die beoogt de breedte en de diepte van de woordenschat

te meten, is het Word Associates Format (WAF) (Schmitt et al., 2011). De taak gaat als volgt: kandidaten selecteren uit een lijstje van acht woorden vier woorden gerelateerd aan een doelwoord. De woorden zijn verdeeld in twee sets van vier. Uit de eerste set moeten de kandidaten twee woorden selecteren die synoniem zijn aan het doelwoord, uit de tweede set moeten twee collocaties worden geselecteerd: woorden die vaak samen met het doelwoord worden gebruikt (Schmitt et al., 2011).

Een voorbeeld van een WAF item, afkomstig uit Schmitt et al. (2011) is het volgende:

Doelwoord: fundamental

neutral	core*	perfect	root*	marriage	objective*	agreement*	news

Onderzoek van Schmitt et al. (2011) vond, op basis van hardop-denk-interviews, dat het WAF-format goed meet aan de uiteinden van de schaal: kandidaten met een grote en een kleine woordenschat kunnen goed worden geïdentificeerd. Voor de middenmoters was de relatie tussen WAF-score en ware woordkennis minder duidelijk. De auteurs menen dat de WAF-score de woordenschat mogelijk overschat.

Het WAF-format wordt vaak gebruikt in onderzoek, maar Qian & Schedl (2004) hebben mogelijkheden voor gebruik in een gestandaardiseerde toets onderzocht. Het onderzoek werd uitgevoerd in aanloop naar de implementatie van de destijds nieuwe TOEFL iBT. Aan het onderzoek deden 207 studenten Engels als tweede taal mee, die studeerden aan een Canadese universiteit. De taken waren dertig woordenschatitems die deel uitmaakten van een oude TOEFL-toets voor het meten van leesvaardigheid. Deze items bestonden uit een in een zin aangeboden doelwoord met vier meerkeuze-opties waaruit de participant het best passende synoniem moest selecteren. Daarnaast kregen de participanten 30 tekstbegripsvragen bij vijf teksten, eveneens afkomstig uit TOEFL, en dertig voor het doel geconstrueerde WAF-items (door Qian en Schedl aangeduid als DVK-items, *depth of vocabulary knowledge*) over dezelfde set woorden. Uit de resultaten bleek dat beide woordenschattaken een vergelijkbare correlatie hadden met de leestaak en dus even goede voorspellers waren van leesvaardigheid. De onderzoekers verwachtten een beter terugslageffect van de DVK-taak, omdat die niet alleen de omvang maar ook de kwaliteit van de woordenschat meet. Tegelijk merkten ze op dat de constructielast voor de DVK-items hoog was omdat het niet makkelijk was om goede items te construeren. De taak maakt overigens geen deel uit van het huidige TOEFL-examen en werd dus niet geïmplementeerd.

Zhang en Koda (2017), in een review van WAF onderzoek, vonden weinig onderbouwing voor de validiteit van de WAF-tests. Ze komen tot de conclusie dat design, afname en wijze van scoring zowel de gemeten performance van de kandidaat als de psychometrische eigenschappen van de test beïnvloeden.

4.3.3 Context-gebonden assessment en portfolio

- Alternatief assessment wordt bepleit als ethisch alternatief voor gestandaardiseerde toetsen.
- In Canada wordt portfolio assessment ingezet voor highstakes beslissingen.

Alternatief assessment is een verzamelnaam voor verschillende assessmentprocedures, waaronder zelfbeoordeling, peer-assessment, portfolio, leerdagboeken, en docent-observaties. Ook task-based assessment wordt wel gerekend tot alternatief assessment (Norris, 2012). Waar traditionele taaltoetsen leren en toetsen als gescheiden activiteiten ziet, verbindt alternatief assessment deze (Ghafournia, 2015; Lynch, 2001). Alternatief assessment is erop gericht de ontwikkeling van individuen te onderzoeken en kijkt daarbij naar performance. Het gaat uit van het menselijk vermogen tot leren en ontwikkelen (Lynch, 2001). Norris (2012) stelt dat een alternatief assessment gericht moet zijn op taalgebruik in contexten. Hij adviseert de assessmentprocedure te kiezen op basis van het doel ervan en te focussen op de informatie die het assessment moet opleveren voor alle stakeholders.

Lynch (2001) heeft zich gebogen over ethische aspecten van alternatief assessment. Hij stelt dat traditionele taaltests een verstorend effect hebben op het leerproces en ongelijkheid bewerkstelligen, omdat ze als 'gatekeeper' fungeren. Bijvoorbeeld toegang tot hoger onderwijs kan afhangen van een taaltoets. Daarnaast is voorbereiden op de toets geen voorbereiding op taalgebruik in het echte leven. Lynch stelt dat alternatief assessment aan een aantal criteria moet voldoen: ten eerste authenticiteit, dat hier verwijst naar de mate waarin stakeholders en kandidaat informatie verkrijgen als gevolg van het assessment. Het tweede criterium is rechtvaardigheid (*fairness*): de bedoelde

en onbedoelde gevolgen van het assessment. Toetsontwikkelaars moeten zich bezighouden met de vraag naar het gebruik van de test. Toetsgebruik maakt deel uit van de validiteit en dus ook van de mate waarin het toetsgebruik ethisch is. Rechtvaardigheid is verbonden met macht: wie bepaalt wat er gebeurt met wie? Lynch past zijn criteria toe op zelfbeoordeling. Bij zelfbeoordeling wordt de macht verlegd naar leerders, maar de macht is nog steeds eenzijdig. Beter is het om de verantwoordelijkheid te delen tussen assessor, leerder en medecursisten (Lynch, 2001). Ghafournia (2015) is van mening dat gestandaardiseerde toetsen en alternatief assessment beide voor- en nadelen hebben. Zo is een gestandaardiseerde test meestal goed onderzocht, en een zelfbeoordeling mogelijk subjectief. De beste informatie wordt verkregen door beide vormen van toetsing en assessment aanvullend te gebruiken (Ghafournia, 2015).

Jenkins en Leung (2019) uitten scherpe kritiek op gestandaardiseerde Engels taalexamens voor toegang tot hoger onderwijs. Een belangrijk negatief aspect is dat de toetsen fungeren als 'gatekeeper' en mensen uitsluiten. Verder zijn gestandaardiseerde toetsen gebaseerd op een geïdealiseerd model van standaard-Engels. Dit past niet meer in de huidige context van mobiliteit en superdiversiteit. Recente conceptualisaties van 'English as a Lingua Franca' (ELF) gaan uit van meertalige communicatie. Gestandaardiseerde toetsen zijn niet authentiek en niet valide. Toetstaken in academische toegangstoetsen zijn kunstmatig, en reflecteren niet de academische realiteit, waarin schrijven en spreken veel lezen en veel voorbereidingstijd vereist. Jenkins en Leung (2019) bepleiten daarom een context-gebonden zelf-assessment. Universiteiten zouden aspirant studenten materiaal en taken moeten aanbieden die representatief zijn voor wat studenten in de studie zullen tegenkomen. De student zelf beoordeelt of die de taken voldoende aankan. Dit heeft twee beoogde voordelen: de studenten kunnen het taalleren richten op de doelsituatie en de verantwoordelijkheid komt bij de student te liggen.

Portfolio-assessment is een veelgebruikte vorm van alternatief assessment, dat beoogt de autonomie van taalleerders te ondersteunen en inzicht in het leerproces te vergroten (Kühn & Pérez Cavana, 2012). Een portfolio is een verzameling van verschillende producten die de ontwikkeling van taalleerders vastleggen. Portfolio's worden veel gebruikt in de context van formatief assessment, maar kunnen ook gebruikt worden voor highstakes beslissingen. Dat gebeurt bijvoorbeeld in Canada, waar *portfolio-based language assessment* (**PBLA**) wordt gebruikt in de inburgeringscursussen (Abbott & Lee, 2025). Het Canadese inburgeringsprogramma heet LINC: Language Instruction for Newcomers to Canada, en het raamwerk dat in deze cursussen wordt gehanteerd heet Canadian Benchmarks. Portfolio-assessment wordt op de eerste plaats gebruikt om te beoordelen of leerders kunnen doorstromen naar de volgende benchmark, het volgende niveau. Maar PBLA wordt ook gebruikt om in aanmerking te komen voor naturalisatie. Hiervoor is benchmark 4[5] vereist. Het niveau moet worden aangetoond met behulp van de bewijzen in het portfolio. Het portfolio wordt dus ingezet voor zeer belangrijke beslissingen.

In een onderzoek op basis van interviews met 118 praktisch geschoolde leerders keken Abbott en Lee (2023) naar de ervaringen van leerders met het Canadese portfolio. Het portfolio fungeerde zowel als leermiddel als formeel beoordelingsinstrument. De houding van de leerders ten opzichte van het portfolio was over het algemeen positief: de studenten erkenden dat het portfolio de organisatie van leermateriaal bevorderde en het volgen van hun voortgang faciliteerde. Ze zagen het portfolio echter vooral als een door de docent gereguleerd beoordelings-instrument. De onderzoekers concludeerden dat de beoordelingsfunctie van het portfolio niet faciliterend is voor de bevordering van zelfregulatie: de formatieve en summatieve functie van het portfolio gaan niet goed samen.

4.3.4 Toetsen die meerdere niveaus meten

- Verschillende grootschalige, highstakes toetsen (waaronder TOEFL en IELTS) rapporteren niet een zakslaagbeslissing maar een behaalde score.

Een deel van de teksten die werden gevonden in het kader van de voor deze paragraaf gehanteerde zoekopdracht, bespraken het fenomeen dat sommige toetsen meerdere niveaus meten. In plaats van een zak-slaagbeslissing wordt een behaald niveau aan de kandidaat gerapporteerd. Hoewel wellicht niet helemaal in lijn met de vraag die in deze sectie centraal staat (die immers was gericht op de integratie van vaardigheden), leek deze vorm van integratie voldoende relevant. Deze paragraaf gaat daarom in op de integratie van niveaus.

5 Vergelijkbaar met A2, bron: ChatGPT.

Garcia Laborda & Fernandez Alvarez (2021) beschrijven een aantal grootschalige highstakes toetsen Engels als tweede taal die meerdere niveaus meten. Deze toetsen zijn: IELTS, TOEFL iBT, Aptis, Duolingo English Test, LanguageCert Test International; Linguaskill, Pearson's Test of English, Oxford Test of English, TOELC Exams, en Trinity Integrated Skills in English. De auteurs beschrijven de kenmerken van de toetsen, die ik samenvat in Tabel 5. Het vervolg van de paragraaf beschrijft drie belangrijke toetsen uitgebreider: IELTS, TOEFL, en Duolingo English Test.

Tabel 5 *Multilevel toetsen*

Toets	Tijdsduur (totaal)	Niveaus	Onderdelen (x & y = geïntegreerd)	Uitslag
IELTS	2 uur, 45 minuten	1-9	Luisteren; lezen; spreken; schrijven.	14 dagen
TOEFL iBT	2 uur	Below basic Basic Low intermediate High Intermediate Advanced	Luisteren, lezen, spreken, schrijven (spreken en schrijven op basis van geschreven en gesproken input)	4 tot 8 dagen
Aptis	3 uur	A1-C1	Grammatica & woordenschat; Spreken; Schrijven; Luisteren; Lezen	2 tot 3 dagen
Duolingo	1 uur	A1-C2	Onduidelijk	2 dagen
LanguageCert Test International ESOL	4 uur	A1-C2	Luisteren; Lezen & schrijven; Spreken	3 tot 5 dagen
Cambridge Assessment	2 uur	A1-C2	Lezen; Luisteren; Spreken; Schrijven.	Le + lui onmiddellijk; schrijven (automatisch gescoord) binnen 1 dag; Spreken binnen 2 dagen.
Pearson Test of English Academic	3 uur	A2-C2	Spreken & schrijven; Lezen; Luisteren	Binnen 2 dagen
Oxford Test of English	2 uur	A2, B1 en B2	Lezen; Luisteren; Spreken; Schrijven	Binnen 14 dagen

International English Language Testing System (IELTS)

Read (2022) schreef een uitgebreide review van de IELTS test, de grootste toets Engels als tweede en vreemde taal. Deze paragraaf vat de bevindingen samen.

IELTS is gestart in 1989 als examen voor mensen die een beurs kregen om aan universiteit in de UK te studeren. Met de toets werd vastgesteld hoeveel budget er nodig was voor taalleren voorafgaand aan studie. IELTS kent twee versies, *Academic en General Training*. De *General Training* versie is later ontwikkeld voor mensen die niet-academische opleidingen of beroepsopleiding gaan volgen. De spreek-en luistertaken zijn in beide versies gelijk maar de taken voor de schriftelijke vaardigheden verschillen. In 2019 was het aantal kandidaten dat IELTS aflegden 3,7 miljoen. De toets wordt erkend door 11.000 organisaties in Engelssprekende landen.

De inhoud van IELTS is gebaseerd op twee trends in de jaren '70. De eerste was de communicatieve benadering; deze is zichtbaar in de (gesimuleerde) real life taken van IELTS, waarbij grote waarde wordt gehecht aan de face-validiteit. De tweede trend is het in die jaren groeiend aantal cursussen Engels voor specifieke doelen. Dit werd vertaald in de twee examens voor twee doelen en doelgroepen.

IELTS wordt afgenomen in testcentra op verschillende plaatsen in de wereld. De toets wordt sinds 2017 op de computer afgelegd, met uitzondering van het onderdeel spreken, dat in een face-to-face interview wordt afgenomen. Sinds de COVID pandemie kan IELTS ook thuis worden afgelegd, het interview vindt dan online plaats. IELTS rapporteert een uitslag in termen van 9 niveaus ('bands'). De kandidaat krijgt dus geen zak/slaagbeslissing maar een niveau-aanduiding.

Read (2022) heeft enkele bedenkingen bij de validiteit van IELTS. Ten eerste is het gemeten construct nergens helder beschreven maar organisch ontstaan uit de communicatieve benadering. Het *sociocognitive framework* van Weir, uit 2005, wordt beschouwd als goed passend bij IELTS. Het examen werd aanvankelijk gemaakt voor aspirant-studenten. Later werd het ook ingezet om de taalvaardigheid van kennismigranten te toetsen, met name in Australië waar verstrekken van een visum gekoppeld is aan een score op de toets. Het uitbreiden van toetsdoelen zonder verder onderzoek leidt tot validiteitsvragen. Ook beschouwt Read het viervaardighedenmodel als een validiteitsissue op zich, omdat we in het echte leven vaardigheden vaak combineren.

Voor het examen Luisteren wordt het inputmateriaal ingesproken van een script en is dus niet authentiek. De kandidaat mag één keer luisteren. Onderzoek maakt duidelijk dat slagen veel leunt op toetsstrategieën, niet per se sterk op luistervaardigheid. De voornaamste kritiek op het leesexamen is dat de items eenzijdig zijn en dat globale en interpretatieve taken ontbreken. Voor schrijven krijgt de kandidaat slechts twee taken: een grafiek beschrijven en een essay schrijven. Beide teksten zijn relatief kort, 400 woorden in totaal. De kandidaat krijgt geen inputmateriaal en moet dus gebruik maken van eigen algemene kennis. Ook moet de kandidaat schrijven onder tijdsdruk. Deze kwesties doen afbreuk aan de validiteit.

Het examen spreken wordt afgenomen in een face-to-face interview door één examinator die tevens beoordelaar is. IELTS beschouwt het interview als een belangrijke kracht van het examen. Uit oogpunt van betrouwbaarheid kun je daar vragen bij stellen. Vroeg onderzoek naar IELTS liet een examinator-effect zien, waarbij de score sterk werd bepaald door de interviewstijl van de examinator. Daarna werden er strakkere richtlijnen gegeven, waarbij de examinator meer volgens een vast script werkt. De examinator beoordeelt tijdens de afname op vloeiendheid en coherentie, woordenschat en woordgebruik, grammatica, en uitspraak. IELTS claimt dat de spreektoets betrouwbaar is, als gevolg van de volgende maatregelen:

- Beoordelaars volgen een tweedaagse training.
- De examinator volgt een script.
- Prestaties worden her-beoordeeld als de score op spreken sterk afwijkt van die op de overige vaardigheden.
- Beoordelaars worden gemonitord.
- Beoordelaars volgen eens per twee jaar een opfristraining.

IELTS heeft lange tijd weinig data gepubliceerd over de betrouwbaarheid van de toetsen maar in 2021 zijn er wat data gepubliceerd. Voor de luistertoetsen is een alpha van .88-.91 gerapporteerd; voor lezen .89-.93. Voor de productieve vaardigheden zijn de data van de beoordelaarstraining gerapporteerd: de waardes zijn .83-.86 voor spreken en .81-.89 voor schrijven. Het examen schrijven werd tot voor kort beoordeeld door één beoordelaar. Nu wordt er beoordeeld door twee beoordelaars, die elk één taak beoordelen. Beoordeling vindt plaats in een regionale hub en het beoordelaarsgedrag wordt gemonitord.

Bedenkingen tegen de mate waarin de IELTS schrijftoets representatief is voor schrijven in academische contexten motiveerde (Green, 2006) een onderzoek uit te voeren naar het terugslageffect van de IELTS schrijftoets. Hiervoor observeerde hij ruim 50 uur de lessen in zowel IELTS examentrainingsklassen en algemene cursussen academisch Engels. Participanten in het onderzoek waren 197 leerders en 20 docenten. Aan de hand van een observatieformulier werden verschillende activiteiten geturfd, getimed en geanalyseerd. Uit de resultaten bleek dat beide cursussen overlap vertoonden. Maar de algemene cursussen academisch Engels boden meer vakspecifiek maatwerk en gebruikten materiaal buiten het lesboek om. Ze waren daarmee meer representatief voor de taken in academische contexten. De IELTS-examentrainingsklassen beperkten zich meer tot algemene onderwerpen en de docent besteedde meer aandacht aan de eisen van het examen. Er werd bijvoorbeeld veel geoefend met schrijven onder tijdsdruk. Docenten wezen op de eisen die examinatoren stelden in plaats van op de eisen die worden gesteld tijdens de studie. Er werd gebruik gemaakt van de IELTS-beoordelingsmodellen en in vergelijking met de algemene cursussen academisch Engels was er meer aandacht voor vormaspecten van de tekst. Er werden dus duidelijke aanwijzingen gevonden dat toetsvoorbereiding leidde tot een nauwe opvatting van schrijfvaardigheid.

TOEFL iBT

TOEFL is gericht op taalgebruik in academische contexten en meet de taalvaardigheid in het Engels van mensen voor wie het Engels niet de moedertaal is. TOEFL werd voor het eerst afgenomen in 1961. Sinds 2005 wordt de online versie afgenomen, de TOEFL iBT. Deze paragraaf is gebaseerd op een review van deze toets uit 2009 (Alderson, 2009). De informatie is dus mogelijk enigszins gedateerd.

TOEFL wordt geaccepteerd op 6000 universiteiten en instituties in 136 landen. De toets wordt 30 tot 40 keer per jaar afgenomen en het afleggen van de gehele toets kost vier uur. Scores worden gerapporteerd op een schaal van 0 tot 30 per vaardigheid en vervolgens opgeteld. De maximale score is dus 120. Stakeholders bepalen welke TOEFL-score voor hen toereikend is. Er wordt dus geen zak-slaagbeslissing gerapporteerd.

Voor de receptieve vaardigheden worden items gezaaid, resulterend in een langere toets. Kandidaten krijgen de lees- of de luistertoets met een zaaitekst. Zo blijft de toetstijd relatief laag en krijgt elke kandidaat een even lange toets. Voor lezen krijgt de kandidaat maximaal vijf teksten van 600-700 woorden (drie examenteksten en eventueel twee zaaiteksten). Bij elke tekst horen 12 tot 14 vragen en het totaal aantal vragen waarop de uitslag gebaseerd wordt, is gemiddeld 39. Voor luisteren krijgt de kandidaat vier tot zes lezingen te horen: vier examenteksten en eventueel twee zaaiteksten. De teksten hebben een lengte van drie tot vijf minuten (500-800 woorden), met zes vragen per tekst. Daarnaast zijn er twee tot drie gesprekken van ongeveer drie minuten, met vijf vragen per tekst. Fragmenten mogen één keer worden beluisterd, waarbij de kandidaat aantekeningen mag maken.

Voor spreken krijgt de kandidaat zes taken[6]: twee onafhankelijke spreektaken en vier geïntegreerde taken. De toets meet het spreken over alledaagse onderwerpen en het integreren van informatie uit verschillende bronnen. Bij de geïntegreerde taken hoort een korte leestekst van maximaal 100 woorden. De kandidaat krijgt 30 seconden voorbereidingstijd en 1 minuut spreektijd. De toets wordt beoordeeld met een vierpuntschaal, op de beoordelingsaspecten 'algemeen' (begrijpelijkheid, aansluiting op de taak), presentatie (duidelijkheid, vloeiendheid, intonatie), taalgebruik (grammatica en woordenschat) en uitwerking van het onderwerp (de relatie tussen ideeën). Voor schrijven krijgt de kandidaat twee taken, de eerste is een onafhankelijke schrijftaak waarin de kandidaat een mening moet geven of een keuze moet beargumenteren. De kandidaat krijgt hiervoor 30 minuten. De tweede taak is een geïntegreerde lees-luister-schrijftaak. De te schrijven tekst moet 150 tot 225 woorden zijn en de kandidaat krijgt daarvoor 20 minuten de tijd. De schrijfproducten worden holistisch beoordeeld op een schaal van 1 tot 5. De onafhankelijke taak wordt beoordeeld door een menselijke beoordelaar en een computer-based scoringstool, E-Rater. De geïntegreerde taak wordt door twee mensen beoordeeld. De betrouwbaarheid van lezen, luisteren en spreken is goed, maar de betrouwbaarheid van schrijven was, in 2009, minder goed, waarschijnlijk als gevolg van het beperkt aantal taken (Alderson, 2009).

Duolingo English Test

De Duolingo English Test (DET) wordt steeds meer gebruikt voor toegang tot universiteiten, vooral sinds de COVID-pandemie. In 2019 werd de toets herzien. Wagner (2020) schreef een review over de DET. De DET is een adaptieve toets die de kandidaat thuis kan afleggen. De kandidaat moet beschikken over een rustige ruimte, een goede internetverbinding en een computer met camera en microfoon. De afname duurt ongeveer een uur en de kandidaat krijgt de uitslag binnen twee dagen. De toets rapporteert een uitslag op een schaal van 10-160. De scores 10-55 staan voor een niveau A1/A2, de scores 60-115 staan voor niveau B1/B2 en de scores 120-160 horen bij niveau C1/C2.

De gebruikte taaktypen zijn:

- C-test: in een tekst is de helft van een deel van de woorden weg, de kandidaat moet de woorden compleet maken;
- Woorden beoordelen: de kandidaat moet woorden van pseudo-woorden onderscheiden. Er is een schriftelijke en een mondelinge versie van de taak.
- Dictee
- Hardop voorlezen van een zin
- Open spreektaak (vraag beantwoorden, plaatje beschrijven) (30 seconden)
- Open schrijftaak (plaatje beschrijven, reageren op een stelling) (50 woorden)

Alle taken worden automatisch gescoord. De open taken worden gescoord met behulp van taalmodellen. Deze modellen kijken naar grammaticale complexiteit en accuraatheid, woordgebruik, inhoud, lengte, uitspraak en vloeiendheid.

Wagner (2020) is kritisch op het gebruik van de toets voor toelating tot de universiteit. Zijn voornaamste bezwaren zijn: (a) de toets meet geen begrip noch productie op tekstniveau. De meeste taken beperken zich tot woord- en zinsniveau; (b) de toets meet geen interactionele vaardigheden; (c) de taken zijn niet representatief voor taken in

6 Tegenwoordig worden vier spreektaken aangeboden: https://www.ets.org/toefl/test-takers/ibt/about/content.html

academische contexten, de toets is daarmee niet authentiek; (d) hoewel kandidaat-gedrag met AI wordt gemonitord lijkt de thuisafname fraudegevoelig; (e) geïsoleerde taken (discrete-point) zijn makkelijk voor te bereiden maar de vaardigheid in het maken van deze taken zegt weinig over communicatie in dagelijkse situaties. De toets heeft daarmee een negatief terugslageffect.

4.3.5 *Kritische reflectie*

De vraag die in dit hoofdstuk centraal stond luidt: welke alternatieven zijn er voor het toetsen van vier vaardigheden (luisteren, spreken, lezen en schrijven) en in welke mate worden deze formats gebruikt bij highstakes examens?

Voor de opkomst van de communicatieve benadering in de jaren ’80 was het onder invloed van *Oller’s unitary trait hypothesis* enige tijd gebruikelijk om taalvaardigheid integratief te toetsen via een gatentekst. Ook besprak dit hoofdstuk een woordenschattaak, waarbij woordenschat wordt gezien als een belangrijke voorspeller van taalvaardigheid. Beide formats worden tegenwoordig niet meer geschikt geacht voor een valide en betrouwbare meting vaan taalvaardigheid.

Critici van gestandaardiseerde highstakes toetsing bepleiten een veel grotere rol van de taalleerder in de beoordeling. In elk geval moet de leerder in het proces van evaluatie meer informatie krijgen over diens voortgang. Ook wordt het niet fair geacht belangrijke beslissingen over mensen te baseren op hun taalvaardigheid. In Canada wordt een taalportfolio gebruikt voor belangrijke beslissingen, waaronder naturalisatie. De summatieve, highstakes functie van het portfolio lijkt daarbij echter de formatieve te verstoren (Abbott & Lee, 2025).

Verschillende grotere taalexamens, waaronder TOEFL iBT en IELTS, vertrekken niet vanuit één bepaald te meten niveau waarvoor de kandidaat kan zakken of slagen, maar ze operationaliseren academische taalvaardigheid en rapporteren vervolgens een score op een schaal. Voor kandidaten heeft het model zoals gehanteerd bij IELTS en TOEFL iBT voordelen, omdat de kandidaat in alle gevallen een score krijgt, wellicht beter presteert dan vooraf verwacht, en in elk geval nooit ‘zakt’. Voor afnemers is deze werkwijze wel wat complexer. Scholen, universiteiten en werkgevers moeten zelf bepalen welke score toereikend is in hun context. Voor het staatsexamen Nt2 zou het interessant kunnen zijn om te onderzoeken of de niveaus B1 en B2 geïntegreerd kunnen worden getoetst. Onderzoek van Hulstijn et al. (2012), verder te bespreken bij deelvraag 6, laat alvast zien dat dit voor spreekvaardigheid waarschijnlijk haalbaar is. In dit onderzoek werden acht spreektaken ontwikkeld, die bedoeld waren te verschillen in moeilijkheid. Alle taken bleken goed in staat de taalleerders te verdelen in een B1-groep en een B2-groep. Het lijkt dus niet nodig spreektaken specifiek te ontwikkelen om een prestatie op een bepaald niveau uit te lokken.

4.4 Alternatieven voor afname van het spreekexamen via een talenpracticum-model

Mondelinge vaardigheden worden door velen beschouwd als de meest cruciale taalvaardigheid. Tegelijk is het toetsen ervan de grootste uitdaging (Kenyon & Malone, 2010; Nakatsuhara et al., 2022). Deze paragraaf bespreekt de vraag: Welke alternatieven zijn er voor afname van het spreekexamen via een talenpracticum-model, wat zijn de voor-en nadelen van de alternatieven? De tekst beschrijft achtereenvolgens recente inzichten in het construct spreekvaardigheid, verschillende afnameformats van grootschalige *highstakes* toetsen, de taken die veel worden gebruikt in spreekexamens en de voor-en nadelen van de verschillende modellen. De paragraaf sluit af met een kritische reflectie.

Voor het zoeken naar literatuur over alternatieven voor het spreekexamen in een talenpracticum is de volgende zoekterm gebruikt: (“Testing oral skills” OR “Testing speaking”) AND (“language laboratory” OR “face-to-face” OR “computer-based”). Tabel 6 geeft een overzicht van de gebruikte literatuur.

Tabel 6 *Bestudeerde literatuur deelvraag 3*

1. Alikulova, S. A. (2020). A comparison of computer-assisted and face-to-face speaking assessment: Performance, perceptions, anxiety and computer attitudes. *Экономика и социум*, 11 (78), 61-64.
2. Ardin, P., & Hendryanti, R. (2023). Investigating the Speaking Examiners' Challenges in a Synchronous Online Test. In *Proceedings of the 20th AsiaTEFL-68th TEFLIN-5th iNELTAL Conference (ASIATEFL 2022)* (Vol. 749, p. 22). Springer Nature.
3. Artunç, E. K., & Hart, D. O. (2020). Interactional competence in paired speaking tests: A study on proficiency-based pairings. *System, 89*, 102194.

4. Birjandi, P., & Bagherkazemi, M. (2011). From Face-to-Face to Paired Oral Proficiency Interviews: The Nut Is yet to Be Cracked. *English Language Teaching, 4*(2), 169-175.
5. Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. *Computer-based assessment of foreign language speaking skills, 29*, 51.
6. Ildikó, C. (2002). Is testing speaking in pairs disadvantageous for students? A quantitative study of partner effects on oral test scores. *NOVELTY-BRITISH COUNCIL HUNGARY, 9*(1), 22-45.
7. Isler, C., & Aydın, B. (2021). Developing and validating a computerized oral proficiency test of English as a foreign language (Coptefl). *International Journal of Assessment Tools in Education, 8*(1), 38-66.
8. Iwashita, N., May, L., & Moore, P. J. (2021). Operationalising interactional competence in computer-mediated speaking tests. *Assessing speaking in context: Expanding the construct and its applications*, 283-302.
9. Joo, M. J. (2008). *Korean university students attitudes to and performance on a face-to-face interview (FTFI) and a computer administered oral test (CAOT)* (Doctoral dissertation, Institute of Education, University of London).
10. Karim, S., & Haq, N. (2014). An assessment of IELTS speaking test. *International Journal of Evaluation and Research in Education, 3*(3), 152-157.
11. Kenyon, D., & Malone, M. (2010). Investigating examinee autonomy in a computerized test of oral proficiency. *Computer-based assessment of foreign language speaking skills*, 1-27.
12. Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing, 33*(3), 319-340.
13. Lowe, J., & Yu, X. (2009). Computer assisted testing of spoken English: A study of the SFLEP college English oral test system in China. *Journal of Systemic, Cybernetics and Informatics, 7*(3), 33-38.
14. Onal, A. (2022). Oral Proficiency Interview (OPI) for the Assessment of Speaking. *The Eurasia Proceedings of Educational and Social Sciences, 28*, 11-18.
15. Quaid, E., & Barrett, A. (2021). Interpretations of spoken utterance fluency in simulated and face-to-face oral proficiency interviews. *Language Education & Assessment, 4*(1), 1-18.
16. Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing, 35*(3), 331-355.
17. Roever, C., & Ikeda, N. (2022). What scores from monologic speaking tests can (not) tell us about interactional competence. *Language Testing, 39*(1), 7-29.
18. Sayin, B. A. (2015). Exploring anxiety in speaking exams and how it affects students' performance. *International Journal of Education and Social Science, 2*(12), 112-118.
19. Taylor, L. (2003). The Cambridge approach to speaking assessment. *Research Notes, 13*, 2-4.
20. Timpe-Laughlin, V., Sydorenko, T., & Dombi, J. (2024). Human versus machine: Investigating L2 learner output in face-to-face versus fully automated role-plays. *Computer Assisted Language Learning, 37*(1-2), 149-178.
21. Yujia, Z. H. O. U. (2012). *Test-takers' affective reactions to a computer-delivered speaking test and their test performance*. retrieved 01/08/2012 from http://cblle.tufs.ac.jp/assets/files/publications/working_papers_09/section/295-310.pdf.
22. Xi, X., Norris, J. M., Ockey, G. J., Fulcher, G., & Purpura, J. E. (2021). Assessing academic speaking. *Assessing academic English for higher education admissions*, 152-199. Routledge
23. 손영아. (Youngah Zoon) (2013). The Interlocutor Proficiency Effect on Test-taker *Performance in a Paired Oral Assessment* (Doctoral dissertation, 서울대학교 대학원).

4.4.1 Het construct spreekvaardigheid

- Het construct spreekvaardigheid is in de laatste 25 jaar veranderd, van communicatieve competentie naar taal als een sociale co-constructie.

Voor elke toets is allereerst een heldere definitie nodig van het te meten construct. Dat geldt ook voor spreekvaardigheid. Begin jaren '90, tevens de begintijd van het staatsexamen Nt2, was het communicatieve paradigma dominant. Canale en Swain onderscheidden begin jaren '80 vier componenten die samen communicatieve competentie vormden: grammaticale competentie, discourse competentie, sociolinguïstische competenties en strategische competentie (Salaberry & Burch, 2021; Xi et al., 2021).

Volgens Salaberry en Burch (2021) zijn er in de afgelopen decennia twee belangrijke verschuivingen geweest in de opvatting van taalvaardigheid. De eerste verschuiving wordt aangeduid als de *social turn*, waarbij de aandacht verschuift van taalvaardigheid als een competentie van een individu naar taal als een co-constructie binnen een sociale context. De tweede verschuiving wordt aangeduid als de *multilingual turn*, waarin taalgebruikers in toenemende mate worden gezien als meertalige individuen die tot betekenisuitwisseling komen met behulp van hun gehele

talige repertoire. Beide ontwikkelingen stellen de toetsontwikkelaar voor uitdagingen. In de context van toetsing en assessment van mondelinge vaardigheden heeft met name de *social turn* veel aandacht gekregen. De opvatting dat betekenisuitwisseling plaatsvindt in sociale contexten leidt tot de vraag hoe interactionale competentie zichtbaar kan worden gemaakt. Hierbij is ook aandacht nodig voor het beoordelingsmodel. Waar vroege modellen met name grammaticale, tekstuele, pragmatische en strategische competenties operationaliseerden, is meer aandacht nodig voor pragmatische en interactionele competentie (Xi et al., 2021).

Roever en Kasper (2018) beschrijven deze '*turn*' als de verschuiving van een psycholinguïstisch-individualistisch naar een sociolinguïstisch-interactioneel perspectief. Het eerste perspectief legt nadruk op de vaardigheden van de individuele spreker en kijkt bijvoorbeeld naar diens woordenschat, grammatica, uitspraak en vloeiendheid. Het tweede perspectief kijkt naar het communicatieve doel in een bepaalde situaties en beoordeelt de linguïstische keuzes en hun impact op de gesprekspartners binnen die context.

In een authentieke communicatieve situatie is interactionele competentie (IC) een gezamenlijke inspanning van gesprekspartners. IC is een essentieel element in de constructie van een dialoog, en omvat zowel de competentie voor effectieve interactie als de competentie aanwezig onder de deelnemers aan de interactie (Iwashita et al., 2021). Onder interactionele competentie vallen bijvoorbeeld beurtwisseling, reparatie, gebruik en interpretatie van lichaamstaal en gebruik van hulpmiddelen. IC wordt bepaald door de lokale context. Opvattingen over IC kunnen verschillen. In de meer extreme opvatting is IC een co-constructie die daarom niet kan worden toegeschreven aan de competenties van één individu (Xi et al., 2021). Dat maakt, volgens sommigen, dat IC niet goed te testen is met gestandaardiseerde taaltoetsen; in plaats daarvan is context-gebonden assessment in de klas beter passend (Salaberry & Burch, 2021). Iwashita et al. (2021) zien echter wel mogelijkheden voor het operationaliseren van IC in grootschalige en gestandaardiseerde taaltoetsen. Hoewel via de computer afgenomen spreektoetsen vragen om enige mate van interactie, wordt de interactie doorgaans niet beoordeeld; de beoordeling richt zich op vloeiendheid, accuraatheid, en inhoud, in lijn met een smalle psycholinguïstische definitie van spreekvaardigheid. Iwashita en collega's wijzen erop dat gesproken interactie via computer in toenemende mate normaal is in het dagelijks leven, vooral sinds de pandemie. Ze verwachten daarom dat zich in de toekomst nieuwe mogelijkheden voor valide toetsen van interactie zullen voordoen, met gebruik van technologieën als virtual reality en spraaktechnologie.

Roever en Ikeda (2022) wijzen erop dat een gestandaardiseerde spreektoets, waarin interactie niet echt een rol speelt, veronderstelt dat de kandidaat met het slagen voor de toets niettemin kan functioneren in mondelinge interactie. Hun onderzoek toetst deze veronderstelling. Een groep van 150 participanten deed mee aan het onderzoek. De participanten waren aspirant-studenten, eerstejaarsstudenten en studenten uit een hoger jaar, in subgroepen van ongeveer gelijke grootte. De studenten legden TOEFL iBT voor spreekvaardigheid af en een spreektoets die interactie toetste in rollenspellen. Correlaties tussen beide scores leken op het eerste gezicht vrij hoog: r=.76. Nadere analyse leerde echter dat de TOEFL-scores van de subgroepen aspirant-studenten en eerstejaarsstudenten veel lager correleerden met de scores op de interactietoets (r=.57) en dat de hogere gemiddelde correlatie dus werd veroorzaakt door de vaardigheid van de hogerejaars studenten. De onderzoekers concluderen daaruit dat TOEFL iBT niet voldoende meet of de kandidaat de spreekvaardigheden kunnen toepassen in interactie.

4.4.2 *Afnameformats*

- Sommige grootschalige spreekexamens worden via de computer afgenomen en andere in een interview, met één of twee kandidaten tegelijk.
- In interview-examens vervult de examinator vaak een dubbele rol, van afnameleider en beoordelaar.
- Sommige computertoetsen worden automatisch gescoord, andere door getrainde beoordelaars.

Er zijn verschillende doelen om gespreksvaardigheid te toetsen en ook verschillende manieren om dat te doen. Allereerst wordt er onderscheid gemaakt tussen directe en semidirecte toetsing van taalvaardigheid (Galaczi, 2010). Directe toetsen zijn gebaseerd op een interview tussen een examinator en de kandidaat. Er zijn een aantal variaties op dit format. Zo worden er soms twee examinatoren ingezet, waarvan er één het gesprek leidt terwijl de ander beoordeelt. In sommige examens worden er ook twee kandidaten tegelijk getoetst, waarbij het feit dat de kandidaten onderling met elkaar in gesprek kunnen als voordeel wordt genoemd. Dit maakt de toets meer authentiek en de afname uiteraard ook efficiënter. Het is ook mogelijk dat het gesprek tussen de examinator(en) en de kandidaat of kandidaten plaatsvindt via een videoverbinding. Met name tijdens de COVID 19-pandemie is dit model in opkomst geraakt (Ardin & Hendryanti, 2023).

Semidirecte taalexamens maakt gebruik van de computer (en voorheen van cassettebandjes) om de opdrachten aan de kandidaat te presenteren. Veel onderzoekers (bijvoorbeeld Alikulova, 2020; Isler & Aydın, 2021) bespreken de noodzaak om van een interviewformat over te stappen op een computertoets (**CMT: *Computer Mediated Test***) vanwege capaciteitsproblemen: door hoge kandidaat-aantallen zijn interviews niet langer een optie. Zo bespreekt Alikulova (2020) de implementatie van een examen Engels in China, de College English Test. Deze toets is verplicht voor universitaire studenten in China en moet worden afgelegd door 10 miljoen kandidaten per jaar. Aangezien de capaciteit voor het afleggen van een interviewtoets volkomen ontoereikend was, kon slechts een klein deel van de studenten een spreekexamen afleggen. Het gevolg was dat in de lessen ook weinig aandacht aan spreekvaardigheid werd besteed. De invoering van de CMT moest hier dus niet alleen de afnamecapaciteit vergroten maar via het terugslageffect van de toets ook zorgen voor meer aandacht voor spreken in het taalonderwijs (Alikulova, 2020; zie ook Lowe & Yu, 2009).

CMT's kunnen in meer of mindere mate geautomatiseerd zijn. In sommige gevallen is alleen de afname geautomatiseerd en zijn er getrainde beoordelaars nodig om de prestaties van de kandidaat te beoordelen. In andere gevallen is ook de scoring geautomatiseerd. Galaczi (2010) geeft enkele voorbeelden van toetsen die gebruik maken van verschillende afnamecondities en scorecondities. Deze voorbeelden zijn samengevat in Tabel 7.

Tabel 7 *Spreekexamens besproken in Galaczi (2010)*

F2F interview **Menselijke beoordelaar**	**CMT** **Menselijke beoordelaar**	**CMT** **Automatische score**
Cambridge ESOL's suite of General	BULATS online speaking test	Pearson's PTE Academic speaking test
Speaking test	TOEFL iBT speaking test[7]	Versant Test
ACTFL OPI tes t	COPI	TOEFL iBT speaking practice tests
	SOPI	

Kenyon en Malone (2010) beschreven de ontwikkeling van een CMT voor spreekvaardigheid, genaamd Computerized Oral Proficiency Instrument (COPI). COPI is de opvolger van het Simulated Oral Proficiency Interview (SOPI), dat werd afgenomen met de cassetterecorder. De ontwikkeling van SOPI volgde op de ontwikkeling van de niveaubeschrijving van het American Council on the Teaching of Foreign Languages (ACTFL), vergelijkbaar met het CEFR in Europa, en het bijbehorende Oral Proficiency Interview (OPI). Deze toetsen worden ingezet voor het beoordelen van spreek-vaardigheid in meerdere vreemde talen. In OPI kreeg de kandidaat een serie vragen en taken. Op basis daarvan bepaalde de examinator het niveau van de spreekvaardigheid van de kandidaat, uitgedrukt in termen van de vier ACTFL-niveaus zijn: Novice, Intermediate, Advanced en Superior. OPI vereiste grote capaciteit van getrainde examinatoren en dat was met name voor kleinere talen problematisch. Dat was de aanleiding voor de ontwikkeling van SOPI, later gevolgd door COPI.

COPI is in twee opzichten een interessante CMT voor spreekvaardigheid: ten eerste kan met de toets, in navolging van OPI, een range van niveaus bepaald worden. Ten tweede heeft de kandidaat behoorlijk wat zeggenschap over hoe die de toets wil afleggen. De afname verloopt als volgt: kandidaten kiezen of ze een eerste taak willen op novice, intermediate, advanced of superior niveau. Ook is er keuzevrijheid met betrekking tot het onderwerp van de taak. Na het uitvoeren van de taak kunnen kandidaten een vervolgtaak kiezen die makkelijker, even moeilijk of moeilijker is dan de taak die zij zojuist hebben gedaan. De computer heeft een onderliggend algoritme dat ervoor zorgt dat de kandidaat genoeg taken krijgt om een niveau te kunnen bepalen. Ter voorbereiding van de taak kan de kandidaat, binnen grenzen, zelf bepalen hoeveel denktijd die nodig heeft en hetzelfde geldt voor de spreektijd. Kandidaten op advanced en superior niveau kunnen kiezen of ze de instructies in het Engels of in de doeltaal willen horen. Kandidaten op de lagere niveaus krijgen de instructies altijd in het Engels, de moedertaal van de beoogde kandidaten.

7 Deze toets combineert menselijke beoordeling met automatische score.

Kenyon en Malone (2010) rapporteerden over onderzoek dat de ontwikkeling van COPI begeleidde. Het onderzoek vergeleek de computertoets COPI met de talenpracticumtoets SOPI. In een vergelijkende studie deden alle kandidaten (N=55) de toets onder beide afnamecondities. Een subset van kandidaten (N=46) legt ook een face-to-face interviewtoets (OPI) af. De onderzoekers vonden een correlatie van .95, waarbij de COPI score meestal wat hoger was dan de OPI-score. De kandidaten vonden dat COPI hen beter in staat stelde hun vaardigheden te demonstreren. Ook maakte COPI, met autonomie voor de kandidaat, hen minder nerveus dan de SOPI-conditie, met een band die doorloopt en het tempo bepaalt.

4.4.3 Toetstaken

- Grootschalige *highstakes* examens verschillen met betrekking tot de spreektaken die worden ingezet. Imitatietaken, onafhankelijke spreektaken en geïntegreerde spreektaken komen voor.

Verschillende taken worden gebruikt om spreekvaardigheid uit te lokken. Nakatsuhara et al. (2022) onderscheidden vier veel voorkomende taaktypes:

1. Imitatie: bijvoorbeeld het herhalen van een zin, of voorlezen van tekst. Dit taaktype verdween onder invloed van de communicatieve benadering maar kwam terug in toetsen die gebruik maken van automatische spraakherkenning. Het wordt beschouwd als in beperkte mate authentiek.
2. Onafhankelijke taak: hierbij doet de kandidaat beroep op eigen kennis. Voorbeelden zijn vraag en antwoord, en monoloog. De kritiek op dit taaktype is dat er maar beperkt input wordt gegeven, terwijl kandidaten verschillen met betrekking tot achtergrondkennis. Dit kan een bron van bias zijn.
3. Geïntegreerde taak: kandidaten spreken over een onderwerp waarover input is gegeven, bijvoorbeeld in een leestekst of een luistertekst. Deze taken worden beschouwd als betere reflectie van real-life communicatie. Het nadeel is dat de vaardigheden voor het verwerken van de input (lezen of luisteren) en de output (spreken) niet nauwkeurig los van elkaar te meten zijn. De kwaliteit van spreken wordt dus mogelijk beïnvloed door hoe goed kandidaat kan lezen of luisteren.
4. Interactieve taak: kandidaten werken in paren of groepen, er is aandacht voor interactie. Een dergelijke taak stelt in staat om zowel linguïstische als interactionele competentie te toetsen. Het nadeel is dat er impact kan zijn van de examinator op de prestatie. In het geval van examineren in paren kan er invloed zijn van de medekandidaat op het proces. Onderzoek hiernaar vindt wisselende maten van impact.

Nakatsuhara et al. (2022) verwachten dat in de toekomst nieuwe mogelijkheden ontstaan voor toetsing van spreekvaardigheid, zoals het gebruik van *virtual reality* en interactieve formats op de computer met gebruik van spraakherkenning. Die ontwikkelingen staan echter nog in de kinderschoenen.

De taaktypes 1 (imitatie) en 2 (onafhankelijke taak) worden veel gebruikt in highstakes toetsen. Zo beschrijft Yujia (2012) de taken van de CMT *Global Test of English Communication* (GTEC). De kandidaten beginnen met een voorleestaak, gevolgd door een zinsrepetitietaak. Vervolgens moet de kandidaat op basis van plaatjes een eenvoudig verhaal vertellen en tot slot op basis van een grafiek een mening geven over een onderwerp. Taylor (2003) beschrijft de taken in het *Cambridge Speaking Assessment*. Dit examen kent meerdere onderdelen om verschillende vormen van gesproken interactie te kunnen toetsen, zoals vraag en antwoord, lange spreekbeurten en discussie. Dit examen wordt face-to-face afgenomen. Ook de IELTS spreektest wordt face-to-face afgenomen. De onderdelen zijn: (1) eenvoudige vragen en antwoorden, (2) een monoloog over een wat complexer onderwerp en (3) interactie over het onderwerp van de monoloog (Karim & Haq, 2014). Taaktype 3, de geïntegreerde taak, wordt onder andere gebruikt in TOEFL iBT. In dit examen moet de kandidaat spreken op basis van lees- en luisterteksten (Kyle et al., 2016).

4.4.4 Voor- en nadelen van de modellen

- Scores op interviewexamens en computerexamens correleren hoog.
- Kandidaten hebben vaak de voorkeur voor een interviewexamen.
- De betrouwbaarheid is hoger op een computerexamen.

Voor alle taaltaken in gestandaardiseerde spreekexamens geldt dat ze bekritiseerd worden omdat ze maar in beperkte mate authentiek en communicatief worden geacht. Het format is altijd gebaseerd op stimuli en respons. De examinator, hetzij life, hetzij via de computer, leidt het gesprek en heeft daarmee de regie. Hoewel het interviewformat over het algemeen als authentieker wordt beschouwd dan het spreken via een computer, ontkomt toch ook het interview-examen niet aan deze kritiek (Karim & Haq, 2014; Onal, 2022).

Onderzoek naar de vergelijkbaarheid van spreekexamens op basis van interviews en CMT's vindt zonder uitzondering hoge correlaties (Galaczi, 2010; Joo, 2008; Kenyon & Malone, 2010; Quaid & Barrett, 2021; Sayin, 2015; Timpe-Laughlin et al., 2024; Yujia, 2012). Zo onderzocht Joo (2008) de houding van Koreaanse universiteitsstudenten (N=42) ten aanzien van een face-to-face interview en een CMT spreekvaardigheid en vergeleek vervolgens hun prestaties op de tests. De studenten hadden in vergelijking met eerdere studies over directe en semidirecte tests een meer positieve houding ten opzichte van de CMT; toch gaven ze de voorkeur aan het interview boven de CMT. Uit de resultaten bleek echter dat de scores op de CMT wat hoger waren dan die op het interview en dat ook de beoordelaarsovereenstemming hoger was op de CMT.

Ook Yujia (2012) onderzocht de relatie tussen de affectieve reactie van kandidaten (N=76) en hun prestaties op een CMT voor spreekvaardigheid in het Engels in Japan, de Global Test of English Communication. Over het algemeen waren de kandidaten nerveus voor de test maar vonden ze de computertest valide en fair. De onderzoekers vonden een relatie tussen de inschatting van de student van de moeilijkheid en de testscore, waarbij degenen die de test makkelijk vonden hoger scoorden.

Quaid en Barret (2021) vergeleken de prestaties van kandidaten (N=4) op de Aptis General Speaking Test, een CMT, met voor dit doel geconstrueerde vergelijkbare taken die werden afgenomen in een interviewformat. Kwantitatieve analyse wees uit dat de prestaties onder beide condities overeenkwamen. Sayin (2015) keek naar de spreekangst van kandidaten van een toets Engels in Turkije onder twee condities: een face-to-face interview en een CMT. De verwachting was dat de computerconditie een positief effect zou hebben op de spreekangst. Het onderzoek liet echter zien dat de afnameconditie weinig invloed had op spreekangst en evenmin op de resultaten. Ook meer recent onderzoek naar de spreekvaardigheid van middelbare scholieren onder de twee afnamecondities, face-to-face interview en geautomatiseerde versie op de computer, vond dat prestaties voldoende vergelijkbaar zijn (Timpe-Laughlin et al., 2024); in de CMT waren de spreekbeurten wat langer en onder de interviewconditie waren er wat meer beurtwisselingen.

Uit onderzoek naar voorkeuren van kandidaten blijkt meestal dat kandidaten een (lichte) voorkeur hebben voor een interview-format maar de CMT doorgaans zien als een goede manier om hun spreekvaardigheid te demonstreren (Galaczi, 2010; Isler & Aydın, 2021; Joo, 2008; Lowe & Yu, 2009; Timpe-Laughlin et al., 2024). Participanten in een onderzoek van Lowe en Yu (2009), uitgevoerd onder kandidaten van een examen Engels in China (N=660), gaven aan in een computerexamen meer te letten op accuraatheid en grammaticale correctheid en in het interview meer op inhoud en boodschap. Kandidaten verwachtten een meer analytische beoordeling onder de computerconditie en meer holistisch onder F2F conditie. Ze verwachtten wel dat de beoordeling op basis van een CMT meer betrouwbaar is en minder beïnvloed door examinatorbias.

Er zijn nog steeds grootschalige highstakes examens die gebruik maken van een face-to-face interview, met één of soms twee kandidaten. Sommige toetsontwikkelaars zien in het paarsgewijs toetsen van kandidaten een oplossing voor sommige validiteitsproblemen van de interviewtoets: de sociale dimensie van taalvaardigheid zou zo beter zichtbaar worden. Birjandi en Bagherkazemi (2011) betogen echter dat het paarsgewijs toetsen ook niet zonder problemen is. Onderzoek naar interviewexamens met twee kandidaten tegelijk heeft gekeken naar het effect dat de prestatie van de ene kandidaat heeft op dat van de andere. Uitkomsten van zulk onderzoek variëren. Verschillende studies vonden geen significant effect van de gesprekspartners op elkaars prestaties (Ildikó, 2002; (손영아 , Youngah Zoon, 2013), terwijl Artunç en Hart (2020) in een studie met 100 participanten vonden dat de gesprekspartner wel invloed had op de prestaties van de kandidaat: als een spreekvaardige kandidaat gekoppeld was aan een andere vaardige kandidaat werkten ze onderwerpen beter uit, boden ze elkaar ondersteuning en was hun beurtwisseling gebalanceerd. Ook bij laagvaardige paren was er sprake van gebalanceerde beurtwisseling maar die konden de onderwerpen minder goed uitdiepen. Koppels met één laagvaardige en één hoog-vaardige kandidaat vertoonden asymmetrische patronen met betrekking tot beurtwisseling en inhoudelijke bespreking van het onderwerp. Dit benadeelde beide kandidaten.

In termen van betrouwbaarheid en beoordelaarsovereenstemming hebben de interviewexamens niet de beste reputatie (Onal, 2022). Verschillende examens, waaronder IELTS, hebben maar één toetsleider die tevens de score bepaalt. In dat opzicht scoren CMT's beter, zeker als ook scoring geautomatiseerd is. Indien er gebruik wordt gemaakt van menselijke beoordelaars is training noodzakelijk. Ook zijn CMT's kostenefficiënt op gebied van ontwikkeling en afname. Wel zijn CMT's gevoeliger voor hacken (Galaczi, 2010).

4.4.5 *Kritische reflectie*

De vraag of er alternatieven zijn voor het via de computer afgenomen spreekexamen, zoals dat bij het staatsexamen Nt2 gebruikt wordt, kan met ja beantwoord worden. Hoewel heel wat grootschalige spreekexamens gebruik maken van de computer, komt ook het interview met één of twee kandidaten en één of twee examinatoren nog voor. Op het gebied van capaciteit (het interview vraagt meer inzet van mensen en middelen), betrouwbaarheid van de beoordeling, en beoordelaarsgemak biedt een computertoets voordelen boven een interview.

Het literatuuronderzoek maakte ook duidelijk dat het interviewexamen op het gebied van validiteit nadelen heeft die overeenkomen met die van het computerexamen: het taalgebruik in het interview is kunstmatig en beperkt, en het gesprek wordt gedomineerd door de examinator die de macht heeft in het gesprek. Er zijn in dat opzicht weinig redenen om de afname van de spreekexamens via de computer te heroverwegen.

De literatuur biedt wel interessante ideeën voor verbetering van de huidige condities van het spreekexamen. Zo beschrijven Kenyon en Malone (2010) hoe kandidaten bij het afleggen van de taaltoets COPI een vrij hoge mate van autonomie krijgen: ze hebben keuzemogelijkheden met betrekking tot het niveau van de taak, het onderwerp van de taak, en de denk- en spreektijd. Dit reduceert stress bij kandidaten.

Problemen van de kandidaten van het onderdeel spreken bij het staatsexamen Nt2 zouden eerder opgelost moeten worden door te kijken naar de afnamecondities. In de gelezen literatuur werd ‘last van de buren’ nergens genoemd als een nadeel van computerexamens. Het lijkt een specifiek probleem van het staatsexamen Nt2 te zijn. Het aantal kandidaten per examenruimte en geluidsisolatie door bijvoorbeeld gebruik te maken van betere koptelefoons zijn knoppen waaraan kan worden gedraaid.

4.5 Lezen van een beeldscherm en van papier

De vraag die deze paragraaf bespreekt luidt: Is lezen van een beeldscherm anders dan lezen van papier? Is een leesexamen waarbij van scherm moet worden gelezen valide en betrouwbaar? Deze vraag leeft niet alleen bij het staatsexamen Nt2. Lezen van beeldscherm is relatief nieuw en is tegelijk ook in toenemende mate het nieuwe normaal. Er hebben zich dan ook heel wat onderzoekers gebogen over de vraag of de vaardigheid lezen onder de condities ‘papier’ en ‘scherm’ vergelijkbaar is. Onderzoek heeft zich gericht op lezen als veranderd construct, op tekstbegrip van verschillende soorten teksten, op leestempo en op de impact van verschillende technische specificaties, zoals lettertype en schermgrootte. Daarnaast hebben onderzoekers gekeken naar de voorkeuren van lezers.

Omdat lezen van scherm relatief nieuw is, is het mogelijk dat onderzoeksbevindingen van twintig tot 25 jaar terug achterhaald zijn: lezers zijn immers steeds meer digitaal gaan lezen en daardoor wellicht ook meer aan lezen van scherm gewend geraakt. In dit deel van het literatuuronderzoek is daarom een aangepaste zoekperiode aangehouden, van 2015-2025. De gebruikte zoekterm is “Reading from paper” OR “reading from screen”. De gebruikte literatuur is opgenomen in Tabel 8.

Tabel 8 *Bestudeerde literatuur deelvraag 4*

1. Aziz, M., Hassan, A., & Aljamili, O. (2024). The effect of paper reading versus screen reading on the inferential reading performance among university ESL learners. *Jurnal Arbitrer, 11*(1), 85-95.
2. Bouazza, K. (2023). *The Effect of Reading Medium on EFL Students' Reading Comprehension Reading on Paper versus Reading on Screen: A Case of Master Two Students at Mascara University* (Doctoral dissertation).
3. Çınar, M., Doğan, D., & Seferoğlu, S. S. (2021). The effects of reading on pixel vs. paper: a comparative study. *Behaviour & Information Technology, 40*(3), 251-259.
4. Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of research in reading, 42*(2), 288-325.
5. Elliott, L. J., Ljubijanac, M., & Wieczorek, D. (2019, June). The effect of screen size on reading speed: A comparison of three screens to print. In *International Conference on Applied Human Factors and Ergonomics* (pp. 103-109). Cham: Springer International Publishing.
6. Furenes, M. I., Kucirkova, N., & Bus, A. G. (2021). A comparison of children's reading on paper versus screen: A meta-analysis. *Review of educational research, 91*(4), 483-517.

7. Jeong, Y. J., & Gweon, G. (2021). Advantages of print reading over screen reading: A comparison of visual patterns, reading performance, and reading attitudes across paper, computers, and tablets. *International Journal of Human–Computer Interaction, 37*(17), 1674-1684.
8. Khadawardi, H. (2021). Transfer of print-based academic reading strategies to on-screen reading. *International Journal of English Language Education, 9*(1), 60-105.
9. Kong, Y., Seo, Y. S., & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education, 123*, 138-149.
10. Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing, 33*(3), 319-340.
11. Lizunova, I. V., van der Weel, A., Garcia-Febo, L., & Pshenichnaya, E. V. (2022). Reading on paper and screens: advantages, disadvantages, and digital divide. *Annals of Library and Information Studies (ALIS), 69*(1), 34-43.
12. Lu, H., Paul, S., & Clarke, P. (2022). Reading on screen versus paper: A comparison of L1-English and L1-Chinese students in an English-medium university. (preprint).
13. Mendhakar, A., Sierak, K., Chana, K., & Leder, H. (2025). Does the reading platform matter? An eye-tracking study. *AILA Review, 38*(1), 105-128.
14. Ocal, T., Durgunoglu, A., & Twite, L. (2022). Reading from screen vs reading from paper: does it really matter? *Journal of College Reading and Learning, 52*(2), 130-148.
15. Peras, I., Klemenčič Mirazchiyski, E., Japelj Pavešić, B., & Mekiš Recek, Ž. (2023). Digital versus paper reading: A systematic literature review on contemporary gaps according to gender, socioeconomic status, and rurality. *European journal of investigation in health, psychology and education, 13*(10), 1986-2005.
16. Richardson, J. T. (2022). *The legibility of serif and sans serif typefaces: Reading from paper and reading from screens* (p. 161). Springer Nature.
17. Rifiyani, A. J., Widowati, D. R., Ni'mah, D., & Ubaidillah, M. F. (2022). To read on-screen or printed texts? A quantitative study into EFL students' reading achievement. *International Journal of Education and Learning, 4*(3), 191-201.
18. Tüchler, A. F. (2024). Reading Print vs. on Screen: Exploring Reading Comprehension, Information Location Recall, and Reading Habits Across Media/Author Aisha Futura Tüchler, MSc.
19. Walsh, G. (2016). Screen and paper reading research–A literature review. *Australian Academic & Research Libraries, 47*(3), 160-173.
20. Yu, J., Zhou, X., Yang, X., & Hu, J. (2022). Mobile-assisted or paper-based? The influence of the reading medium on the reading comprehension of English as a foreign language. *Computer Assisted Language Learning, 35*(1-2), 217-245.

4.5.1 Het construct leesvaardigheid

- Lezen van papier vergt lineaire verwerking van de tekst.
- Lezen van scherm is niet-lineair en vaak multimediaal.

De eerste vraag die we moeten beantwoorden, is wat we verstaan onder lezen. Walsh (2016) wijst erop dat er verschillende vaardigheden gecombineerd moeten worden in het leesproces:

> *On a practical level, reading is a combination of recognising words and deciphering their meaning, as well as comprehending the wider context within which the words are read. (...) comprehension of text, in comparison to basic reading, requires higher levels of cognitive and linguistic skills for the reader to be able to understand both the literal and inferential meanings of text. (Walsh, 2016, p. 161)*

De vraag die zich vervolgens opdringt, is of lezen van papier en van beeldscherm vergelijkbaar is, en dezelfde vaardigheid vereist van de lezer. Dat is niet (helemaal) het geval. Lezen van papier en van scherm worden gezien als verschillende constructen (Lizunova et al., 2022). Lezen van papier vergt het lineair verwerken van een tekst, gemotiveerd door het leesdoel, zoals bijvoorbeeld het vinden van specifieke informatie. Digitaal lezen is doorgaans niet-lineair en multimodaal. Het vergt naast leesvaardigheid ook digitale geletterdheid: de vaardigheid in het navigeren door multimediale informatie (Peras et al., 2023). Lizunova en collega's (2022) menen dat digitale teksten, omdat ze vluchtig en niet-tastbaar zijn, een andere emotionele respons oproepen bij de lezer. Digitale teksten nodigen ertoe uit kleine tekstfragmenten te lezen en stimuleren zo gefragmenteerde aandacht, ze nodigen eerder uit tot scannen en skimmen en minder tot aandachtig lezen.

Leestoetsen worden verondersteld te meten hoe goed een tekst begrepen wordt. Tekstbegrip is gebaseerd op een mentale representatie van de tekst en omvat meerdere gerelateerde componenten, zoals woordherkenning, het ontwikkelen van ideeën vanuit deze woorden en deze verbinden tot een samenhangende mentale representatie (Clinton, 2019). Tekstbegripstaken kunnen verdeeld worden in taken die letterlijk begrip vragen en taken die een beroep doen op het maken van **inferenties**: het proces waarbij de kandidaat met behulp van aanwijzingen uit de tekst en de eigen voorkennis conclusies trekt over informatie die niet letterlijk in de tekst staat (Clinton, 2019).

De Progress in International Reading Literacy Study (PIRLS) is een internationaal onderzoek naar de leesvaardigheid van kinderen in de 'fourth grade', vergelijkbaar met groep 6 van de basisschool in Nederland. De editie 2016, afgenomen bij 340.000 leerlingen, was één van de eerste grootschalige onderzoeken die de leesvaardigheid van leerlingen zowel via beeldscherm als via papier toetsten. PIRLS 2016 vond grote individuele verschillen tussen de twee condities in de leesvaardigheid van studenten, ook gerelateerd aan verschillen tussen de genders, waarbij meisjes de betere lezers waren (Mullis et al., 2017; Peras et al., 2023).

4.5.2 De lezer

- Vaardigheid in lezen op papier voorspelt de vaardigheid in lezen van scherm.
- Lezers hebben een voorkeur voor lezen van papier.
- Lezers gebruiken meer leesstrategieën als ze lezen van papier.
- Meisjes zijn doorgaans betere lezers dan jongens.
- Mensen die opgroeien met een rijke leescultuur zijn eveneens de betere lezers.

Leestheorieën zien lezen als een proces van interactie tussen de lezer en de tekst. Veel van het vergelijkend onderzoek naar lezen van scherm en van papier gaat in op de uitkomst van die interactie: het tekstbegrip. Maar er zijn ook onderzoeken die daarnaast, of uitsluitend, focussen op de lezer. Dat onderzoek kan gaan over voorkeuren van lezers, over leesgedrag of over achtergrondkenmerken van lezers, zoals gender, of de leescultuur waarin zij opgroeien. Onderzoek naar voorkeuren van lezers laat consistent een voorkeur voor papier zien (Bouazza, 2023; Jeong & Gweon, 2021; Ocal et al., 2022). Clinton (2019) vond dat lezers bij lezen van papier beter in staat hun leesperformance in te schatten dan bij lezen van scherm. In een onderzoek onder 74 volwassen lezers vergeleken Jeong en Gweon (2021) lezen van papier en van beeldscherm. Ze keken onder andere naar de attitudes van de participanten. Lezers rapporteerden een beter tekstbegrip, meer vertrouwen en betere concentratie, en minder vermoeidheid bij het lezen van papier.

Peras et al. (2023) voerden een systematische review uit op 23 onderzoeken die lezen van scherm en van papier vergeleken in de groep 6-18 jarigen. Ze keken daarbij naar de invloed van verschillende lezer-gerelateerde factoren die leesvaardigheid konden beïnvloeden, zoals leeftijd, leesniveau, sociaaleconomische status en gender. Ze vonden dat de beste predictor voor lezen van scherm de vaardigheid in lezen op papier is. Met andere woorden: betere lezers lezen beter in beide condities. Ook familieomstandigheden, zoals het aantal boeken dat de lezers thuis hebben, leesplezier bij de ouders en opleiding van de ouders voorspelden de vaardigheid in lezen van scherm. Onderzoek naar genderverschillen liet zien dat meisjes vaak betere lezers zijn dan jongens. Op het gebied van de invloed van sociaaleconomische status (SES) is weinig onderzoek, maar een Duitse analyse van PISA 2012-data liet zien dat SES een deel van de variantie in leesvaardigheid voorspelt (10%); dit percentage was echter hoger bij lezen van papier dan bij lezen van scherm (Peras et al., 2023).

Khadawardi (2021) voerde een kwalitatief onderzoek naar lezen van scherm van academische teksten en keek naar de leesstrategieën die mensen daarbij gebruiken. Ze vergeleek deze met de strategieën die gebruikt worden bij lezen van papier. Leesstrategieën zijn bewuste acties gericht op het begrijpen van de tekst en het oplossen van problemen bij het komen tot begrip. De participanten waren twintig Arabisch sprekende studenten (Master's en PhD), die studeerden aan een Engelse universiteit. Uit de resultaten bleek dat lezers meer strategieën gebruiken bij het lezen van papier. Khadawardi concludeert dat lezen van papier daarom effectiever is en leidt tot beter tekstbegrip.

In een onderzoek dat lezen van papier en van de mobiele telefoon vergeleek, keken ook Yu en collega's naar leesstrategieën. De participanten in dit onderzoek waren eerstejaarsstudenten in China, die Engels leerden als verplicht vak. Deelnemers vulden een vragenlijst in en rapporteerden minder effectief gebruik van leesstrategieën bij het lezen van de telefoon. Ook rapporteerden ze beter tekstbegrip wanneer zij lazen van papier (J. Yu et al., 2022).

4.5.3 Tekstbegrip

- Onderzoek naar leesbegrip van informatieve teksten laat zien dat lezen van papier leidt tot beter begrip dan lezen van scherm.
- Het voordeel van papier ten opzichte van scherm is kleiner geworden.

Een van de centrale vragen in onderzoek naar lezen van papier versus scherm is of het medium invloed heeft op leesbegrip. Clinton (2019) voerde een systematische review en meta-analyse uit van onderzoek dat tussen 2008 en 2018 was gepubliceerd. Deze analyse omvatte 33 studies waarin prestaties op leesbegripstaken onder de condities van papier en scherm werden vergeleken. Uit de resultaten bleek dat lezen van scherm een klein maar significant negatief effect had op leesperformance. Dit effect trad alleen op bij informatieve teksten, niet bij verhalende teksten. Het nadeel van schermlezen gold bovendien zowel voor vragen die letterlijk begrip vereisten als voor vragen die het maken van inferenties uitlokten. Ook Kong et al. (2018) voerden een meta-analyse uit, gebaseerd op 17 studies (waarvan negen uitgevoerd vóór 2013 en acht daarna). Hun bevindingen kwamen grotendeels overeen met die van Clinton (2019): tekstbegrip was beter bij lezen van papier. Deze onderzoekers merkten op dat het verschil in tekstbegrip tussen de twee condities over de tijd kleiner leek te worden.

Onderzoek van Lu et al. (2022) nuanceerde deze bevinding. In een studie onder 120 studenten in het Verenigd Koninkrijk, verdeeld naar moedertaal (Engels of Chinees), werd het effect van medium onderzocht bij informatieve teksten. Studenten die van papier lazen, scoorden gemiddeld hoger dan studenten die van scherm lazen, maar dit verschil betrof uitsluitend de inferentievragen. Voor de letterlijke begripsvragen werd geen verschil gevonden tussen de twee condities. Vanuit een ander perspectief keek Walsh (2016) in een literatuuronderzoek naar de voor- en nadelen van scherm versus papier in relatie tot tekstbegrip. Zij concludeerde dat studies naar het lezen van langere, complexe academische teksten doorgaans een voordeel voor papier laten zien. Samenvattend laten deze onderzoeken zien dat, tot ongeveer 2018, lezen van papier over het algemeen betere resultaten opleverde dan lezen van scherm, vooral bij informatieve teksten en inferentievragen.

Recente reviews hebben onderzocht of het verschil tussen lezen van papier en van scherm in de loop van de tijd afneemt. Peras et al. (2023) voerden een systematische review uit van 23 studies bij kinderen en jongeren tussen zes en achttien jaar oud, gepubliceerd tussen 2015 en 2022. Ongeveer de helft van de studies vond geen verschil in tekstbegrip tussen papier en scherm, terwijl de andere helft een voordeel voor papier rapporteerde. In één geval werd een voordeel van het scherm gevonden op woordenschat. In dit geval bood de digitale applicatie een functie om woordbetekenissen toe te lichten. Leerlingen die dit programma gebruikten, scoorden hoger op aan woordenschat gerelateerde taken. Ook Furenes et al. (2021) voerden een review uit, gericht op jongere kinderen (1–8 jaar) en op het begrip van verhalen. Zij analyseerden 39 studies en concludeerden dat de kinderen verhalen beter begrepen wanneer ze van papier werden gelezen. Een voordeel van het digitale formaat werd ook hier alleen gevonden voor woordenschat, en dan uitsluitend wanneer de applicatie een ingebouwde woordenlijst bevatte. Zelfs in die gevallen werd echter een negatief effect op tekstbegrip vastgesteld. De onderzoeken onder kinderen en jongeren wijzen dus overwegend op een voordeel voor papier, hoewel digitale hulpmiddelen sommige nadelen van schermlezen tot op zekere hoogte kunnen compenseren.

Een onderzoek onder basisschoolkinderen in de groepen 6 tot 8, uitgevoerd in Oostenrijk, keek zowel naar tekstbegrip als naar het vermogen van lezers om informatie te lokaliseren in de tekst (Tüchler, 2024). Het onderzoek bestond uit twee delen. Het eerste deel, met 120 participanten werd leesbegrip getoetst onder de condities papier en computerscherm. De kinderen die van papier lazen presteerden significant beter. Het tweede experiment, met 128 participanten, keek naar het vermogen om informatie te lokaliseren bij lezen van papier en van scherm. De kinderen presteerden goed in beide condities.

Recente studies (na 2020) onder volwassenen en studenten laten een meer divers beeld zien. Sommige onderzoeken rapporteren geen verschil tussen papier en scherm, terwijl andere een voordeel vinden voor het ene of het andere medium. Çınar et al. (2021) onderzochten het lezen van verhalende teksten bij 126 middelbare scholieren, waarbij teksten werden gelezen van papier, computerscherm, tablet of telefoon. Er werd geen effect van leesmedium op tekstbegrip gevonden. Ocal et al. (2022) vergeleken papier en beeldscherm bij het lezen van informatieve teksten onder 69 studenten in de Verenigde Staten en kwamen tot eenzelfde conclusie: er werd geen verschil in performance gevonden tussen beide condities.

Een studie uit Indonesië von juist een voordeel voor het lezen van scherm. Rifiyani et al. (2022) onderzochten studenten van een lerarenopleiding Engels die willekeurig over de condities papier en scherm werden verdeeld. De studenten lazen een nieuwsbericht en beantwoordden twintig tekstbegripsvragen. De groep die van scherm las, presteerde significant beter.

Andere recente studies blijven een voordeel van papier vinden. Bouazza (2023) onderzocht dertig Algerijnse studenten Engels als vreemde taal die drie academische teksten lazen, waarvan twee informatief en één narratief. Studenten die van papier lazen, presteerden beter op de informatieve teksten, terwijl er geen verschil werd gevonden voor de narratieve tekst. De onderzoeker concludeerde dat lezen van papier nog steeds leidt tot betere concentratie en tekstbegrip. Aziz et al. (2024) onderzochten bij 426 Pakistaanse studenten de vaardigheid om inferenties te maken—onderverdeeld in analyseren, evalueren en creëren. Voor alle drie vaardigheden presteerden studenten significant beter bij lezen van papier. De auteurs verklaarden dit door het verlies aan overzicht dat optreedt bij scrollen in digitale teksten. Omdat kandidaten in de twee condities niet exact dezelfde tekst kregen, is echter onzeker of de gevonden verschillen uitsluitend aan het medium toe te schrijven zijn.

Over de hele linie lijkt het verschil in tekstbegrip tussen lezen van papier en van scherm mettertijd kleiner te worden, zoals ook Kong et al. (2018) al suggereerden. Terwijl oudere studies consequent een voordeel voor papier aantoonden, laten recente onderzoeken een meer genuanceerd beeld zien, met zowel neutrale (Ocal et al., 2022) als tegengestelde (Rifiyani et al., 2022) resultaten. Dit wijst mogelijk op toenemende vertrouwdheid met digitale media, al worden complexe taken nog altijd beter uitgevoerd als er van papier wordt gelezen (Aziz et al., 2024; Bouazza, 2023).

Leestempo en leesgedrag

- Het medium (scherm of papier) heeft geen invloed op het leestempo.
- Eye-trackingsonderzoek suggereert dat medium invloed heeft op leesgedrag.

Onderzoek naar vergelijking van de condities scherm en papier richt zich op het proces van lezen. Een veelgebruikt criterium bij de operationalisering van dit proces is de leestijd. Leestijd wordt gezien als een indicator voor het leesgemak (Clinton, 2019). Reviews van onderzoek naar verschillen tussen lezen van scherm en van papier laten, met betrekking tot leestempo, over het algemeen geen verschillen zien tussen beide condities (Clinton, 2019; Kong et al., 2018; Walsh, 2016).

Recent is er ook onderzoek uitgevoerd dat gebruik maakt van eye-tracking. Dit onderzoek volgt de oogbeweging van lezers om inzicht te verkrijgen in het leesproces. Zulk onderzoek werd onder andere uitgevoerd door Jeong et al. (2021). Zij onderzochten de effecten van leesmedium (op papier, op computerscherm en op tablet) op de oogbewegingen, waarbij fixaties worden geïnterpreteerd als momenten waarop de cognitieve verwerking van de tekst plaatsvinden. Dit experiment werd uitgevoerd in Zuid-Korea, met 74 volwassen studenten. De studenten lazen negen informatieve teksten over verschillende onderwerpen, afkomstig uit een taalexamen. Het onderzoek toonde aan dat lezers tijdens het lezen van papier een kortere fixatieduur en een hoger aantal fixaties vertoonden dan tijdens het lezen van het beeldscherm.

Vergelijkbaar onderzoek werd uitgevoerd door Mendhakar en collega's (2025). Dit onderzoek maakte eveneens met gebruik van eye-tracking en vergeleek lezen uit een papieren boek met lezen via e-reader, iPad, en computerscherm. Participanten waren 80 universitaire studenten van verschillende studierichtingen. De taak betrof lezen van een verhalende tekst (twee hoofdstukken uit *Pride and prejudice*). De leestijd werd gemeten, evenals de oogfixaties. Studenten lazen sneller op de PC en de iPad. Er werden verschillen in leestempo gevonden tussen de e-reader en het boek. De onderzoekers vonden minder en langere oogfixatie op bij het lezen op de PC, de andere condities vertoonden vergelijkbare patronen. De auteurs concluderen dat in de hand gehouden apparaten (iPad, e-reader) vergelijkbaar zijn met een papieren boek.

Technische aspecten

- Er zijn geen eenduidige onderzoeksresultaten over het best leesbare lettertype.
- Het formaat van het scherm heeft geen invloed op leestempo of tekstbegrip.

Richardson (2022) wijdde een heel boek aan de vraag of een schreefletter beter leesbaar is dan een schreefloze letter. Het boek geeft, onder andere, een overzicht van onderzoek naar leesbaarheid van lettertypes met en zonder schreef, onder de condities lezen van papier en lezen van scherm. Vijf gerapporteerde studies vergeleken het leestempo bij het lezen van teksten met schreefletter of een schreefloos lettertype. Eén vond dat het leestempo hoger lag bij de schreefletter, twee vonden dat het tempo hoger lag bij de schreefloze letter en twee vonden geen verschil. Drie onderzoeken keken naar tekstbegrip bij teksten met fonts met en zonder schreef. Eén vond een voordeel voor de schreefletter, één vond een voordeel voor de schreefloze letter en één vond geen verschil. Zeer kort samengevat is er geen wetenschappelijk bewijs voor de algemene aanname dat bij lezen van papier een schreefletter beter leesbaar is. Bij lezen van scherm zijn er aanwijzingen in het voordeel van de schreefloze letter: enkele onderzoeken vonden dat bepaalde schreefloze letters (Consolas, Letter Gothic, Verdana) beter leesbaar zijn dan bepaalde schreefletters (Times New Roman, Courier). Echter, de afstand tussen de letters is relevanter dan de aan- of afwezigheid van de schreef.

Een ander technisch aspect van leesbaarheid waarnaar onderzoek is gedaan, is het schermformaat: dit zou invloed kunnen hebben op het leesgemak. Elliott et al. (2019) rapporteren over een experiment dat deze vraag centraal stelde: beïnvloedt het formaat van het scherm het leestempo van studenten, in vergelijking met lezen van papier? De participanten lazen vier korte verhalen, elk gedurende tien minuten. Ze lazen uit een papieren boek, een 17 inch computerscherm, een 10 inch tablet en een mobiele telefoon met een 3 inch scherm. Het aantal woorden dat binnen tien minuten werd gelezen, werd geteld en daarna werd het aantal woorden per minuut berekend. De onderzoekers vonden geen significante verschillen tussen de verschillende leescondities. Het leestempo van individuen was opvallend consistent. Schermgrootte heeft dus geen invloed op het leestempo van lezers, bij het lezen van een verhalende tekst. Vergelijk ook het hiervoor gerapporteerde onderzoek van Çınar en collega's (2021), die focusten op tekstbegrip en eveneens schermgrootte meenamen in hun onderzoeksopzet. In het experiment dat zij uitvoerden werden verhalende teksten aangeboden via papier, computerscherm, tablet of telefoon. Het onderzoek vond geen effect van leesmedium op het tekstbegrip.

Kritische reflectie

Lezen van scherm is een construct dat afwijkt van lezen op papier vanwege multimediale, non-lineaire karakter van schermlezen. Dat heeft gevolgen voor de operationalisering in een leesexamen. Als we in een leesexamen lineaire teksten op scherm zouden aanbieden, dan zou dat vragen kunnen oproepen over de validiteit van het leesexamen.

Hoewel het verschil tussen lezen van scherm en van papier mettertijd kleiner is geworden, lijkt het nog steeds niet helemaal verdwenen, met name als het gaat om informatieve, langere, en academische teksten. Hoe groter de cognitieve last voor de lezer, hoe meer het lezen van papier nog altijd een voordeel lijkt te bieden. Het staatsexamen Nt2 heeft daarom goede redenen de teksten aan te bieden op papier, al zijn er uiteraard praktische redenen denkbaar om een overgang naar lezen van scherm te overwegen.

4.6 Beoordeling van Schrijven en Spreken

Het tweede deel van dit literatuuronderzoek, dat de deelvragen 5 tot en met 9 bespreekt, gaat over aspecten van de beoordeling van de productieve vaardigheden spreken en schrijven. Het toetsen en beoordelen van productieve vaardigheden is een uitdaging, omdat er altijd interactie is van de kandidaat en diens performance, de toets, de taak, de beoordelaar, het beoordelingsmodel, het construct dat gemeten wordt en de theorie waarop de schaal gebaseerd is (Kuiken & Vedder, 2021). Bij spreekexamens is er bovendien invloed van de menselijke of virtuele gesprekspartner(s) (Fan & Yan, 2020). Er is dan ook heel wat onderzoek gedaan naar de verschillende aspecten van toetsing van de productieve vaardigheden.

Dit deel start met een meer algemene onderzoeksvraag: Wat zegt wetenschappelijke onderzoek op het gebied van beoordeling van tweede- en of vreemde-taalexamens, met name op het gebied van de onderdelen Schrijven en Spreken? De deelvragen 6 tot en met 9 gaan meer specifiek in op verschillende typen beoordelingsmodellen. De literatuur voor deelvraag 5 is gevonden met de volgende zoekterm: ("assessing" OR "testing") AND ("speaking" OR "writing" OR "productive skills"). Een overzicht van de gebruikte literatuur voor deze deelvraag is te vinden in Tabel 9.

Tabel 9 *Bestudeerde literatuur deelvraag 5*

1. Banerjee et al (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing writing* 26 (p. 5-19)
2. Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing, 29*(1), 91-108.
3. Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing writing, 8*(2), 73-83.
4. Cumming, A., Cho, Y., Burstein, J., Everson, P. & Kantor, R. (2021). *Assessing academic writing. In: Xi, X. & Norris, J.M. (eds.) Assessing Academic English for Higher Education Admissions.* Routledge/Taylor & Francis.
5. Fan, J., & Yan, X. (2020). Assessing speaking proficiency: a narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in psychology, 11*, 330.
6. Harsch, C. & Malone, M.E. (2021). Language proficiency Frameworks and scales. In: Winke, P. & Brunfaut, T. (eds.) The Routledge Handbook of Second Language Acquisition and Language Testing.
7. Hoang, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY Access. *Language Assessment Quarterly, 13*(4), 359-376.
8. Isaacs, T. (2016). Assessing speaking. *Handbook of second language assessment, 12*, 131-146. De Gruyter Mouton.
9. Joo, S. H. (2016). Self-and peer-assessment of speaking. *Studies in Applied Linguistics and TESOL*, 16(2).
10. Kane, M. T. (2022). Articulating a validity argument. In: Fulcher, G. & Harding, L. *The Routledge Handbook of Language Testing.* (pp. 32-47). Routledge/ Taylor & Francis.
11. Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing, 38*(4), 602-626.
12. Knoch, U. (2022). Assessing writing. (236-253). In Fulcher, G. & Harding, L. *The Routledge Handbook of Language Testing.* (pp. 236-253). Routledge/ Taylor & Francis.
13. Kuiken, F., & Vedder, I. (2016). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing, 34*(3), 321- 336.
14. Kuiken, F. en Vedder, I. (2021). Scoring approaches. (pp. 125-134). In: Winke, P. & Brunfaut, T. (eds.) *The Routledge Handbook of Second Language Acquisition and Language Testing.* (pp. 125-135). Routledge/ Taylor & Francis.
15. Nakatsuhara, F., Khabbazbashi, N., & Inoue, C. (2022). Assessing speaking. In Fulcher, G. & Harding, L. *The Routledge Handbook of Language Testing*. pp 209-222. Routledge/ Taylor & Francis.
16. Nakatsuhara, F., Inoue, C. & Khabbazbashi, N. (2021). Measuring L2 speaking. In: Winke, P. & Brunfaut, T. (eds.) *The Routledge Handbook of Second Language Acquisition and Language Testing.* (pp. 285-294). Routledge/ Taylor & Francis.
17. Pill, J. & Smart, C. (2021). Raters. Behavior and training. In Winke, P. & Brunfaut, T. (eds.). The Routledge handbook of second language acquisition and language testing. pp 135-144. Routledge / Taylor & Francis.
18. Poe, M., & Elliot, N. (2019). Evidence of fairness: Twenty-five years of research in assessing writing. *Assessing Writing, 42*, 100418.
19. Romova, Z., & Andrew, M. (2011). Teaching and assessing academic writing via the portfolio: Benefits for learners of English as an additional language. *Assessing Writing, 16*(2), 111-122.
20. Runge, A., Goodwin, S., Attali, Y., Poe, M., Mulcaire, P., Lo, K. L., & LaFlair, G. T. (2025). A multistage interactive writing task for the assessment of English language writing proficiency. *Language Testing, 42*(4), 423-446.
21. Salaberry, M. R., & Burch, A. R. (Eds.). (2021). Assessing speaking in context: *Expanding the construct and its applications* (Vol. 149). Multilingual Matters.
22. Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing, 35*(2), 161-192.
23. Xi, X., Norris, J.M., Ockey, G.J., Fulcher, G. & Purpura J.E. (2021). Assessing academic speaking. In: Xi, X. & Norris, J.M. (eds.) *Assessing Academic English for Higher Education Admissions.* Routledge/Taylor & Francis.
24. Yoon, H.-J. & Burton, D. (2021). Measuring L2 writing. (pp. 295-304). In: Winke, P. & Brunfaut, T. (eds.) The Routledge Handbook of Second Language Acquisition and Language Testing.

Deze deelvraag is breed en de gevonden literatuur vertoont gedeeltelijke overlap met de literatuur voor de deelvragen 6 t/m 9. Om overlap te voorkomen is voor deze eerste, brede deelvraag literatuur geselecteerd die de constructen spreken en schrijven bespreekt en een algemene introductie geeft op het toetsen van deze vaardigheden. Verder is literatuur geselecteerd die een algemene introductie geeft op beoordelingsmodellen voor spreken en schrijven en op de rol van de beoordelaar. Tot slot is een korte beschrijving opgenomen van de twee meest bekende taaltoetsen voor het Engels, TOEFL en IELTS, en de wijze van beoordeling van de productieve vaardigheden in deze toetsen.

4.6.1 Het construct spreekvaardigheid

- Spreken wordt op verschillende manieren geoperationaliseerd, waarbij de nadruk in toenemende mate is komen te liggen op taalgebruik in contexten.

Paragraaf 4.4 besprak de veranderde inzichten in het construct spreekvaardigheid. Kort samengevat was in de jaren '90 het communicatieve paradigma leidend, gebaseerd op vier componenten van communicatieve competentie (grammaticaal, discourse, sociolinguïstisch en strategisch). Sindsdien hebben zich twee belangrijke verschuivingen voorgedaan (Salaberry & Burch, 2021). De eerste verschuiving wordt aangeduid als de *social turn*: taal wordt gezien als het product van sociale interactie in plaats van als individuele vaardigheid. Dit maakt de vraag belangrijk hoe *interactionele competentie (IC)* zichtbaar en beoordeelbaar kan worden gemaakt. De tweede verschuiving is de *multilingual turn*: taalgebruikers worden gezien als meertalige individuen die hun volledige talige repertoire kunnen inzetten.

Nakatsuhara et al. (2022) zijn van mening dat naast het sociale model, dat spreken beschouwt als een co-constructie tussen gesprekspartners, ook het psycholinguïstische model nog gangbaar is. Vanuit dit perspectief worden onderliggende componenten van taalvaardigheid gemeten als maat voor de efficiëntie waarmee taalleerders taal verwerken. Deze visie onderbouwt bijvoorbeeld geautomatiseerd spreekassessment. Isaacs (2016) betoogt dat het Britse perspectief meer gericht is op interactie (en dus validiteit), terwijl Amerikaanse toetsontwikkelaars traditioneel een groot accent leggen op de psychometrische betrouwbaarheid. Als gevolg daarvan wordt het onderdeel spreken van zowel IELTS als Cambridge, beide Britse toetsen, afgenomen als een interviewexamen, terwijl het Amerikaanse TOEFL onderdeel spreken wordt afgenomen met de computer, zodat de toets-omstandigheden beter kunnen worden gecontroleerd. Aangezien toetstaken en beoordelingsmodellen gebaseerd zouden moeten zijn op de theoretische onderbouwing van het construct, kunnen veranderde inzichten gevolgen hebben voor de toetsing van spreekvaardigheid.

4.6.2 Het construct schrijfvaardigheid

- Schrijven is een meerdimensionaal construct dat wordt getoetst door beoordeling van schrijfproducten.

De verschillende perspectieven op het construct spreekvaardigheid gelden ook voor schrijven. Schrijfvaardigheid is een meerdimensionaal construct. Het omvat microprocessen van woordkeus, spelling, en interpunctie; compositie-processen, zoals plannen, schrijven van kladversie, revisie, en editing; en macroprocessen zoals respecteren van genre-conventies en coherentie in argumentatie (Cumming, 2002; Cumming et al., 2021). Ondanks uiteenlopende perspectieven en accenten zijn toetsontwikkelaars het er over eens dat een schrijftoets in de T2 het schrijven van tekst moet omvatten, en niet gebaseerd moet zijn op het separaat meten van deelcomponenten. Dat heeft geleid tot drie uitgangspunten:

1. Schrijven als vorm van interactie: taken specificeren zaken als onderwerp, doel, en beoogd publiek. Schrijven is dus gebonden aan communicatieve context.
2. Gebruik van beoordelingsmodellen voor het scoren: deze modellen kijken onder andere naar taakuitvoering. Ze zijn vaak gerelateerd aan een schaal, bijvoorbeeld een niveaubeschrijving.
3. Toetstaken op basis van een beschrijving van het domein. Wat moeten mensen kunnen schrijven in een bepaalde setting en welke eisen worden gesteld aan de kwaliteit van de producten? (Cumming et al., 2021).

In de geschiedenis van het beoordelen van schrijfvaardigheid speelt het spanningsveld tussen betrouwbaarheid, validiteit, en praktische toepasbaarheid een grote rol. Tot het begin van de twintigste eeuw waren schrijftoetsen gebaseerd op het schrijven van essays. Vanwege zorgen om de betrouwbaarheid werden deze in het midden van de twintigste eeuw, met name in de Verenigde Staten, vervangen door meerkeuzetoetsen. Eind jaren '70 werden de meerkeuzetoetsen vanwege zorgen om de validiteit van meerkeuzetoetsen vervangen door directe schrijftaken. Vanwege efficiëntie wordt meer recent technologie ingezet voor de beoordeling, maar dat leidt weer tot nieuwe vragen over de validiteit, omdat alleen specifieke elementen van schrijfvaardigheid automatisch gescoord kunnen worden (Knoch, 2022).

4.6.3 Het ERK

- Het ERK bevat uitgebreide beschrijvingen van het construct taalvaardigheid.

Toetsontwikkeling begint bij het bepalen van het te meten construct. In Europa is het Europees Referentiekader voor talen het meest gebruikte referentiedocument dat taalvaardigheid beschrijft op verschillende niveaus (Harsch & Malone, 2021). Volgens Yoon en Burton (2021) bevat het ERK één van de meest gedetailleerde beschrijvingen

van het construct schrijfvaardigheid. Deze beschrijving omvat zowel linguïstische als sociaal-cognitieve elementen van taalgebruik.

Een valide beoordeling zou zowel moeten focussen op adequaatheid van de taakuitvoering als op linguïstische aspecten. Het ERK beschrijft het meerdimensionale en sociaal-cognitief georiënteerde proces van schrijfvaardigheid, bijvoorbeeld in de descriptoren voor monitoring en reparatie, die schrijfvaardigheid als proces van revisie benaderen. Vanuit dit perspectief is ontwikkeling van schrijfvaardigheid niet alleen een cognitief proces, maar ook een proces van kennisconstructie door sociale interactie. Actieve interactie met anderen is dan ook voorwaardelijk voor de ontwikkeling van schrijfvaardigheid. Maar Runge et al. (2025) wijzen erop dat deze aspecten zelden zijn geïmplementeerd in gestandaardiseerde toetsen.

4.6.4 Validiteit en betrouwbaarheid

- Het argument-based validation framework biedt een model voor evaluatie van de kwaliteit van toetsen.

Toetsontwikkelaars moeten duidelijke relaties kunnen leggen tussen het construct dat ze meten, de operationalisering van het construct en de beslissingen die op basis van het construct kunnen worden genomen. Dit wordt wel aangeduid als *argument-based validation*. Argument-based validation is een proces waarin meerdere bewijzen worden verzameld om de validiteitsclaim te onderbouwen. In de woorden van Kane (2022,): 'An argument-based approach to validation imposes two basic requirements: first, state the claims that are being made, and second, evaluate the credibility of these claims' (p. 32).

Het argument-based validation framework kent zes verschillende typen inferenties die gevalideerd moeten worden (Fan & Yan, 2020):

1. Domeindefinitie: onderbouwing van de relatie tussen de gemeten vaardigheid en de beoogde vaardigheid in de doelcontext;
2. Evaluatie: onderbouwing van de kwaliteit van beoordeling en beoordelaars;
3. Generalisatie: onderbouwing van de generaliseerbaarheid van scores en de afwezigheid van construct-irrelevante ruis.
4. Uitleg: onderbouwing van de wijze waarop kandidaten worden ingedeeld naar vaardigheid.
5. Extrapolatie: Onderbouwing van de claim dat de uitslag voorspellend is voor het functioneren in doelsituaties.
6. Gebruik: Onderbouwing van de interpreteerbaarheid en bruikbaarheid van scores voor de te nemen beslissingen.

Een review van ruim 100 onderzoeken naar de validiteit van spreektoetsen maakte duidelijk dat validiteitsonderzoek zich voornamelijk heeft gericht op Evaluatie, Generalisatie en Uitleg. Veelbesproken onderwerpen zijn: het construct spreekvaardigheid, beoordelaarseffecten, betrouwbaarheid, en beoordelingsmodellen (Fan & Yan, 2020). De kwaliteit van de beoordeling is een centraal aandachtspunt in toetsing gebaseerd op menselijke beoordelaars (Isaacs, 2016).

Het is een uitdaging om de beoordeling van productieve vaardigheden voldoende betrouwbaar te maken. Middelen daartoe zijn beoordelaarstraining, opfristraining en monitoring. In een systematische review van 259 artikelen, gepubliceerd tussen 1980 en 2013, keken Wind en Peterson (2018) naar de manier waarop betrouwbaarheid wordt gemeten en gerapporteerd. Het onderzoek liet zien dat er een breed scala aan technieken wordt gebruikt om betrouwbaarheid te meten, waarvan beoordelaarsovereenstemming (het percentage overeenstemmende scores), interbeoordelaarsbetrouwbaarheid (beoordelaarsconsistentie), generaliseerbaarheidstheorie en Rasch-modellen het meest worden gebruikt.

4.6.5 Nieuwe taaktypen

- De inzet van grote taalmodellen (AI) biedt mogelijkheden voor nieuwe taaktypen.
- Vanuit een kritisch perspectief wordt ook democratisering van taaltoetsing bepleit, waarbij taken context-gebonden zijn en er aandacht is voor het proces.

Verschillende typen taken worden gebruikt om productieve vaardigheden uit te lokken. Taken waarbij kandidaten moeten schrijven of spreken op basis van een korte prompt, worden aangeduid als onafhankelijke taken. In toenemende mate zijn geïntegreerde taken populair. In een geïntegreerde taak moet de kandidaat schrijven of spreken op basis van geschreven of gesproken input. Echter, zowel op het gebied van construct als scoring zijn er ook uitdagingen bij dit taaktype, eerder besproken in paragraaf 4.2. Daarnaast laat onderzoek zien dat niet alleen de

taalvaardigheid van kandidaten hun toetsresultaat beïnvloedt, maar ook de kenmerken van de taken zelf. Kandidaten kunnen namelijk verschillend presteren op verschillende taken, omdat niet alle taken voor elke kandidaat even moeilijk of even toegankelijk zijn (Yoon & Burton, 2021). Zo kan een kandidaat op de ene taak goed presteren en op een andere taak minder goed, zonder dat dit noodzakelijkerwijs een verschil in taalvaardigheid weerspiegelt. Deze systematische verschillen in prestaties die worden veroorzaakt door de taak, en niet door de onderliggende vaardigheid, worden aangeduid als **taakeffecten**. Zie bijvoorbeeld ook het onderzoek van Lee (2006) besproken in paragraaf 4.2.

Meer recent worden de mogelijkheden van grote taalmodellen (AI) gebruikt bij de operationalisering van schrijfvaardigheid en de interactieve dimensie ervan. Runge et al. (2025) bijvoorbeeld beschreven een interactieve schrijftaak voor het toetsen van schrijfvaardigheid, opgenomen in de Duolingo English Test. Deze taak wordt afgenomen met de computer en kent twee fasen. In de eerste fase reageert de kandidaat op een prompt. Een taalmodel detecteert een onderwerp dat de kandidaten niet besproken heeft in de eerste respons en vraagt de kandidaat daar verder op in te gaan. Op deze manier wordt in een computertoets (een aspect van) de procesmatige kant van schrijfvaardigheid nagebootst. Om te onderzoeken of het tweede deel van de taak op de beoogde wijze werd uitgewerkt, werden 300 producten beoordeeld op relevantie door getrainde beoordelaars. Deze producten en de beoordelingen werden ingevoerd in ChatGPT, dat op deze manier getraind werd om dezelfde beoordeling uit te kunnen voeren. Daarna werden ruim 25.000 responsen automatisch beoordeeld. De onderzoekers vonden dat de aansluiting van de vervolgtaak op de initiële taak goed verliep en dat daarmee een zekere mate van interactiviteit in een gestandaardiseerde, grootschalige highstakes schrijftoets[8] kon worden geoperationaliseerd met behulp van grote taalmodellen.

Veel onderzoek benadrukt de procesmatige een meerdimensionale aspecten van schrijven en de rol van feedback (bijvoorbeeld Romova & Andrew, 2011). Volgens Poe en Elliot (2019) is er sprake van verschillende overlappende veranderingen in de opvatting van schrijfvaardigheid: de opvatting van schrijven als een sociale activiteit, de proces-georiënteerde opvatting van schrijven, sociale leertheorieën, en de rol van sociale ongelijkheid. Onderzoekers die vanuit deze perspectieven kijken vinden doorgaans dat assessment het best in de klas, of anderszins 'in de gemeenschap' kan plaatsvinden. Onderzoekers hebben bijvoorbeeld bepleit om schrijfproducten in overleg tussen assessoren te beoordelen en ander onderzoek heeft de processen die assessoren daarbij gebruiken onderzocht. Ook hebben onderzoekers gekeken naar de rol van de sociale context op de interactie van kandidaten met taken, bijvoorbeeld in toetsen Engels als vreemde taal waarvan taken niet goed aansloten bij de belevingswereld van Aziatische kandidaten, vanwege culturele bias (Poe & Elliot, 2019). Deze onderzoekers bepleiten een context-gebonden procesgeoriënteerde aanpak, bijvoorbeeld door gebruik te maken van portfoliomethodiek, peerassessment en zelfbeoordeling (Hoo, 2016). Alternatieve, meer holistische beoordelingsvormen worden ook overwogen als antwoord op ethische kwesties met betrekking tot taaltoetsing. Alternatieve vormen van toetsing zouden meer recht doen aan de complexiteit van schrijfvaardigheid en aan verschillen tussen mensen (Cumming, 2002; Hoo, 2016).

In een onderzoek onder 41 gevorderde leerders van het Engels in Nieuw-Zeeland maakten Romova en Andrew (2011) gebruik van portfolio-assessment bij studenten die een cursus academisch schrijven volgden. De studenten legden een portfolio aan met schrijfproducten, waarbij elke tekst na het eerste concept nog twee keer werd herschreven. Dit proces van revisie reflecteert het gangbare proces van academisch schrijven. Het bood de leerders feedback op de teksten, en dus de mogelijkheid te leren van fouten en zwakke punten in de tekst te verbeteren. Het maakte de leerders ook vertrouwd met schrijven als proces.

4.6.6 Beoordelingsmodellen en beoordelaars

- Er zijn verschillende typen beoordelingsmodellen, waarbij holistische en analytische modellen de bekendste zijn.
- Beoordelingsmodellen worden ontwikkeld op basis van niveaubeschrijvingen of op basis van prestaties van leerders.
- Naast beoordeling door menselijke beoordelaars komt ook automatische scoring van schrijf- en spreekproducten steeds meer voor.

8 Duolingo English Test wordt op veel universiteiten geaccepteerd als bewijs van voldoende vaardigheid in het Engels (Garcia Laborda & Fernandez Alvarez, 2021).

Scores op een performancetoets vormen de link tussen de prestatie van de kandidaat en de vaardigheidsclaim van de toets (Knoch et al., 2021). Beoordelingsmodellen zijn het instrument dat meestal wordt gebruikt om tot die score te komen. Een beoordelingsmodel is een beschrijving van taalvaardigheid, bestaand uit een aantal geconstrueerde niveaus waartegen de leerder wordt beoordeeld (Kuiken & Vedder, 2021). Kuiken en Vedder (2019, 2021) onderscheiden verschillende typen beoordelingsmodellen, waarbij categorieën deels kunnen overlappen. Zo kunnen analytische en holistische modellen worden onderscheiden, naast primary trait en multiple trait modellen. Modellen kunnen intuïtief of empirisch tot stand komen. En ze kunnen taak-afhankelijk of taakonafhankelijk zijn (Kuiken & Vedder, 2019, 2021). Over alle types is debat geweest; er is niet één duidelijk beste model.

Bij holistische beoordelen wordt er per taak een totaalscore gegeven voor de performance. Een holistische score kan impressionistisch of op basis van een beoordelingsvoorschrift worden gegeven. Bijvoorbeeld TOEFL iBT hanteert een holistisch model op basis van een beschrijving van elk van de niveaus. Holistisch scoren is snel en praktisch, en daarmee ook relatief goedkoop. Een belangrijk nadeel is dat de beoordeling geen nuttige diagnostische informatie oplevert en dus geen basis biedt voor feedback aan de kandidaat. Met een holistisch model is het ook moeilijker om tot goede beoordelaarsovereenstemming te komen (Cumming et al., 2021; Kuiken & Vedder, 2021).

Bij een analytische beoordeling worden verschillende aspecten afzonderlijk beoordeeld. Kuiken en Vedder (2016) wijzen erop dat een goed beoordelingsmodel in staat moet zijn om niet alleen talige aspecten op te pikken, zoals grammatica en woordenschat, maar ook functionele adequaatheid moet beoordelen. Ze zien functionele adequaatheid als een aspect met meerdere dimensies: inhoud (is de uiting inhoudelijk adequaat en relevant?), taakvereisten (voldoet de uiting aan de opdracht, met betrekking tot bijvoorbeeld register en genre), begrijpelijkheid (is de prestatie begrijpelijk voor de ontvanger?) en coherentie en cohesie (is de prestatie samenhangend?)

Bij primary trait scoring krijgen kandidaten net als bij holistische beoordeling één score, maar het scoringsvoorschrift is gemaakt voor de specifieke taak. Er wordt beschreven welk gedrag de taak verlangt. Het voordeel is dat er een goede aansluiting is van model op taak. Het nadeel is dat het arbeidsintensief is om voor elke taak een model te maken, en dat de generaliseerbaarheid van de score naar andere taken onduidelijk is. Multiple trait scoring lijkt enigszins op analytisch scoren. In beide gevallen worden meerdere elementen beoordeeld. Bij multiple trait scoring zijn de aspecten gekozen omdat ze relevant zijn voor de specifieke taak; deze aspecten hoeven niet deel uit te maken van één onderliggend model. Een analytische score is gebaseerd op een model van taalvaardigheid en kent doorgaans taakonafhankelijke aspecten. Voordeel van beide modellen is dat ze meer informatie voor leerders opleveren. Ook zijn deze modellen meestal makkelijker te hanteren voor beoordelaars. Een nadeel is dat beoordeling meer tijd vergt (Kuiken & Vedder, 2021).

Beoordelingsmodellen kunnen ook worden ingedeeld op basis van hoe ze zijn ontwikkeld (Kuiken & Vedder, 2021). We kunnen onderscheid maken tussen modellen gebaseerd op expertoordelen enerzijds (het ERK is op die manier tot stand gekomen) en empirisch onderbouwde modellen gebaseerd op performance van taalleerders anderzijds. Tegenwoordig worden modellen vaak goed onderbouwd door gebruik van meerdere methodes, zoals een combinatie van statische procedures, expertoordelen en analyse van producten van kandidaten. In een review van 36 studies die de ontwikkeling van een beoordelingsmodel beschrijven, kwamen Knoch et al. (2021) tot een meer fijnmazige beschrijving van informatiebronnen die gebruikt worden om beoordelingsmodellen te ontwikkelen en onderbouwen. Dat zijn toets-externe bronnen, zoals theorie, een raamwerk, een syllabus, expertoordelen en bestaande schalen, en toets-interne bronnen zoals performance samples, beoordelaarsfeedback en -performance en assessmenttaken. Hun onderzoek liet zien dat doorgaans meerdere bronnen worden gebruikt.

Beoordelingsmodellen moeten geregeld worden herzien om na te gaan of ze nog aansluiten bij het beoogde gebruik. Banerjee et al. (2015) beschreven dit proces van revisie voor het beoordelingsmodel van een grootschalige taaltoets Engels, genaamd Examination for the Certificate of Proficiency in English (ECPE). Het examen toetst vier vaardigheden en het beoogd niveau is C2 van het ERK. Kandidaten schrijven een essay van minimaal 300 woorden. Het niveau C2 veronderstelt goede en brede beheersing van de woordenschat, gevarieerd en accuraat gebruik van grammaticale structuren, de vaardigheid in het ontwikkelen van een argumentaties, publieksgerichtheid en goede organisatie van de tekst. De revisie van het bestaande model was gebaseerd op drie bronnen: theorie over de beste voorspellers van schrijfvaardigheid, een analyse van schrijfproducten van kandidaten en input van beoordelaars. Vanuit elk van de drie invalshoeken werd de bestaande schaal geanalyseerd en werden verbeteringen voorgesteld. Het herziene model bevatte de aspecten inhoud en ontwikkeling van het onderwerp, organisatie en samenhang van

ideeën, woordenschat, grammatica en aansluiting op publiek. De auteurs bepleiten triangulatie bij de revisie van beoordelingsmodellen.

Hoewel beoordelingsmodellen impact kunnen hebben op de beoordeling betogen Pill en Smart (2021) dat het de beoordelaar is die in het proces van beoordelen de centrale positie inneemt. De beoordeling moet consistent en reproduceerbaar zijn. Ervaring en kennis van het te meten domein helpen beoordelaars bij het uitvoeren van consistente beoordelingen. Beoordelaarstraining helpt bij het bevorderen van de kwaliteit van de beoordeling, maar beoordelaarseffecten kunnen nooit helemaal worden geëlimineerd. Beoordelaars kunnen (te) streng, mild of inconsistent beoordelen. Het inzetten van meerdere beoordelaars per kandidaat is daarom belangrijk.

In toenemende mate wordt ook geautomatiseerde evaluatie van schrijfvaardigheid ingezet (Pill & Smart, 2021). Enerzijds als beoordelingstool, maar ook als feedbacktool in het leerproces. Zo maakt TOEFL iBT bijvoorbeeld gebruik van een menselijke beoordeling én een automatische scoring van spreek- en schrijfprestaties. Voordelen van automatische scoring zijn efficiëntie, consistentie en kosteneffectiviteit. Als nadeel geldt dat het geautomatiseerde systeem meestal een beperkt aantal linguïstische kenmerken van de tekst in kaart brengt en geen aandacht besteedt aan communicatieve adequaatheid (Knoch, 2022, zie ook Pallotti, 2022). Zo bleek bijvoorbeeld uit een onderzoek van Hoang en Kunnan (2016) dat een geautomatiseerd beoordelings- en feedbacksysteem het ontbreken van inhoudelijk adequaatheid niet oppikte, en daarnaast matig overeenstemde met de beoordeling door menselijke beoordelaars. Runge et al. (2025) betogen echter dat meer recente modellen steeds geavanceerder worden en gebruik maken van natuurlijke taalmodellen (Runge et al., 2025).

4.6.7 Beoordeling van IELTS en TOEFL

- TOEFL gebruikt een holistisch beoordelingsmodel en IELTS gebruikt een analytisch model.
- Bij IELTS Spreken beoordeelt de examinator; bij IELTS Schrijven beoordelen twee beoordelaars elk één taak.
- Bij TOEFL wordt de beoordeling door een menselijke beoordelaar gecombineerd met een automatische score.

De waarschijnlijk meest bekende taaltoetsen zijn IELTS (VK) en TOEFL (VS). Beide toetsen zijn ontwikkeld voor het meten van Engels als tweede en vreemde taal. Verwijzingen naar zowel TOEFL als IELTS zijn in de huidige tekst op verschillende plaatsen te vinden, maar deze paragraaf zet de beoordeling van beide toetsen nog eens naast elkaar, te beginnen met spreken, gevolgd door schrijven.

TOEFL iBT Spreken bestaat uit vier taken: één onafhankelijke taak en drie geïntegreerde taken waarbij de kandidaat moet spreken op basis van gelezen en beluisterde input. Het holistische beoordelingsmodel kijkt naar de dimensies presentatie (‘delivery’), taalgebruik en ontwikkeling van het onderwerp. Presentatie omvat de aspecten uitspraak, intonatie en vloeiendheid, taalgebruik omvat grammatica en woordgebruik, en ontwikkeling van het onderwerp combineert inhoudelijke aspecten en coherentie. Voor de geïntegreerde taken omvat dit aspect ook de mate van volledigheid van verwerking van het bronmateriaal (Bridgeman et al., 2011; Xi et al., 2021). Het TOEFL onderdeel spreken wordt afgenomen via de computer. Zie ook de review van TOEFL door Alderson (2009), in paragraaf 4.3.4.

Het IELTS onderdeel Spreken bestaat uit drie onderdelen: in het eerste onderdeel beantwoordt de kandidaat algemene vragen over bijvoorbeeld familie of werk. In het tweede deel moet de kandidaat een korte presentatie geven over een vertrouwd onderwerp, waarbij bepaalde punten moeten worden behandeld. In het derde deel stelt de examinator vragen over aspecten van de presentatie en de kandidaat moet deze vragen beantwoorden en zijn of haar standpunten onderbouwen. Typisch academische spreekvaardigheid wordt niet getoetst. Het analytische beoordelingsmodel kijkt naar de aspecten vloeiendheid & coherentie, woordgebruik, grammaticale beheersing, en uitspraak. Hoewel de toets wordt afgenomen in een face-to-face interview en daarmee in principe ruimte zou bieden voor interactie, controleert de examenleider het gesprek en wordt interactie niet beoordeeld (Xi et al., 2021). Zie ook de IELTS review van Read (2022) in paragraaf 4.3.4.

TOEFL iBT Schrijven bestaat uit twee taken. De eerste taak is onafhankelijk en de tweede taak is geïntegreerd. In de onafhankelijke taak moet de kandidaten in ongeveer 300 woorden een mening geven over een onderwerp. Ze hebben daarvoor 30 minuten de tijd. Voor de geïntegreerde taak krijgen kandidaten input in de vorm van een zeer korte lees- en luistertekst. De teksten vertegenwoordigen tegengestelde meningen. De kandidaat heeft 20 minuten de tijd en de respons moet 150-225 woorden lang zijn. De toets wordt afgenomen via de computer. De schrijfprestaties worden holistisch beoordeeld op een taak-specifieke schaal, en de geïntegreerde taken worden

beoordeeld op een zespuntschaal. Elke taak word beoordeeld door een menselijke beoordelaar en gescoord door een automatisch (Xi et al., 2021). Het automatische scoringssysteem voor het spreekexamen heet SpeechRater. Dit AI-gebaseerde systeem analyseert verschillende aspecten van de gesproken respons, te weten vloeiendheid, uitspraak, prosodie, woordenschat en grammatica.

Het IELTS onderdeel Schrijven bevat eveneens twee schrijftaken. In de eerste taak moet de kandidaat informatie uit een tabel of grafiek bespreken en samenvatten. De kandidaat krijgt daarvoor 20 minuten en de beoogde tekstlengte is minimaal 150 woorden. Voor de tweede taak moet de kandidaat een korte argumenterende tekst schrijven van minimaal 250 woorden. De beschikbare tijd is 40 minuten. IELTS Schrijven kan zowel op papier als via de computer worden afgelegd. Het beoordelingsmodel voor de schrijftaak kijkt naar vier aspecten: taakuitvoering, coherentie & cohesie, woordgebruik, en grammatica, alle gescoord op een negenpuntschaal. De prestaties worden beoordeeld door één beoordelaar per taak (Cumming et al., 2021; Read, 2022).

4.6.8 Kritische reflectie

Ontwikkelaars van taaltoetsen zijn zich in de loop van de afgelopen decennia in toenemende mate bewust geworden van het interactieve en meerdimensionale karakter van spreek- en schrijfvaardigheid. Beide vaardigheden worden tegenwoordig getoetst door middel van communicatieve en zoveel mogelijk authentieke taken. De beoordeling vindt meestal plaats door menselijke beoordelaars, die daarbij uiteenlopende modellen kunnen gebruiken, en soms door automatische scoringssystemen.

De paragraaf besprak de beoordeling van de bekende taaltoetsen TOEFL iBT en IELTS. Het valt daarbij op dat het staatexamens Nt2 in vergelijking met deze bekende toetsen beduidend meer taken aanbieden. Bij het staatexamen-programma Spreken I bijvoorbeeld worden aan kandidaten 16 taken voorgelegd[9], bij programma II zijn dit er 13. Ter vergelijking: TOEFL geeft vier spreektaken en IELTS drie. Voor schrijven zijn dit (ongeveer) 12 taken voor programma I en 10 voor programma II. Ter vergelijking: zowel IELTS als TOEFL geven twee taken.

Ook het aantal beoordelaars is bij het staatsexamen Nt2 opvallend veel groter. Dit komt doordat er in eerste instantie twee beoordelaars per kandidaatproduct beoordeelden en er vervolgens is overgestapt op het segmenteren van de beoordeling van spreken en schrijven (om beoordelaarseffecten te voorkomen), waarbij beoordelaars niet een heel examen beoordelen maar losse opgaven. Hierdoor kunnen bij de beoordeling van het examen van één kandidaat 10 tot 15 beoordelaars betrokken zijn. Een dergelijke grondige aanpak ben ik in de literatuur nergens anders tegengekomen. Ter vergelijking: TOEFL zet een menselijke beoordelaar en een automatische score in, IELTS zet voor spreken één beoordelaar in en voor schrijven twee beoordelaars die elk één taak beoordelen.

Zowel het inzetten van meerdere taken als het inzetten van meerdere beoordelaars zijn maatregelen die toetsontwikkelaars kunnen nemen om tot een goede betrouwbaarheid te komen. Vergelijking met deze gerenommeerde toetsen roept echter wel de vraag op of dergelijke grote aantallen strikt noodzakelijk zijn. Vanuit het oogpunt van efficiëntie lijkt een evaluatie van de huidige inzet van taken en beoordelaars de moeite waard.

4.7 Positieve beoordelingsmodellen

De vraag die dit hoofdstuk bespreekt, luidt: Wat is er bekend over positieve beoordelingsmodellen? Worden deze modellen ingezet in highstakes examens? Deze vraag komt voort uit een ondervonden nadelig effect van bepaalde beoordelingsaspecten van de beoordelingsmodellen voor spreken en schrijven bij het staatsexamen Nt2. Omdat bij de beoordeling van bijvoorbeeld woord- en zinsvorming het aantal fouten wordt geteld, bevoordeelt het beoordelingsmodel de kandidaten die weinig risico nemen en degenen die zeer compacte responsen produceren. Tegelijk bestaat de intuïtie dat betere taalbeheersing juist gepaard gaat met langere en grammaticaal meer complexe uitingen. De vraag is dus hoe beoordelingsmodellen dit type complexe productie kunnen belonen.

De zoekterm “positive rating scales” leverde echter geen resultaten op, deze term is niet gangbaar. De literatuur voor deze deelvraag is gevonden met zoekterm: “complexity” AND (“language testing” OR “writing” OR “speaking” OR “productive skills”). Naar complexiteit, in combinatie met accuraatheid en vloeiendheid (**CAF**: complexity, accuracy,

9 Op basis van openbare bronnen.

fluency) is vrij veel onderzoek gedaan, ook in de context van toetsing en assessment. Pallotti (2021) wijst er echter op dat de CAF-maten niet uitsluitend positief zijn:

> *Complexity and fluency have to do with the interlanguage systems as such, of which they describe several characteristics in positive terms, such as the presence of some structures, or its delivery rate. Accuracy, on the other hand, implies the comparison with another system, namely the target language, which is conducted in negative terms by counting errors (cases in which certain structures are not produced). (Pallotti, 2021, p. 201)*

Naast beoordeling met CAF-criteria, of althans met criteria voor complexiteit en vloeiendheid, zouden wellicht ook holistische beoordelingsmodellen als 'positief' kunnen kwalificeren. Immers, bij holistisch beoordelen geeft de beoordelaar één score voor de prestatie, en daarbij kunnen verschillende zaken, waaronder complexiteit, een rol spelen. Omdat holistisch beoordelen uitgebreid aan de orde komt bij deelvraag 8, bespreekt de huidige deelvraag uitsluitend CAF-onderzoek. De literatuur die gebruikt is voor het beantwoorden van deze onderzoeksvraag is te vinden in Tabel 10.

Tabel 10 *Bestudeerde literatuur deelvraag 6*

1. Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics, 28*(1), 147-164.
2. Bulté, B., Housen, A., & Pallotti, G. (2025). Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning, 75*(2), 533-574.
3. Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. Journal of Second Language Writing, 26, 66-79.
4. Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2011). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR)1. *Language Testing, 29*(2), 203-221.
5. Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly: An International Journal, 3*(2), 151-169.
6. Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing.
7. Kuiken, F., & Vedder, I. (2011). Task complexity and linguistic performance in L2 writing and speaking: The effect of mode. In *Second language task complexity* (pp. 91-104). John Benjamins Publishing Company.
8. Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31*(3), 329-48.
9. Kuiken, F., & Vedder, I. (2019). Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. *International journal of applied linguistics, 29*(2), 192-210.
10. Lu, X. (2017). Automated measurement of syntactic complexity in corpusbased L2 writing research and implications for writing assessment. *Language Testing, 34*(4), 493-511.
11. Mora, J. C., Mora-Plaza, I., & Bermejo Miranda, G. (2024). Speaking anxiety and task complexity effects on second language speech. *International Journal of Applied Linguistics, 34*(1), 292-315.
12. Pallotti, G. (2021). Measuring Complexity, Accuracy, and Fluency. In: Winke, P. & Brunfaut, T. (eds.) The Routledge Handbook of Second Language Acquisition and Language Testing. P. 201-211
13. Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing, 36*(2), 161-179.
14. Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied linguistics, 37*(6), 828-848.
15. Révész, A., Kourtali, N. E., & Mazgutova, D. (2017). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning, 67*(1), 208-241.
16. Rezazadeh, M., Tavakoli, M., & Rasekh, A. E. (2011). The role of task type in foreign language written production: Focusing on Fluency, Complexity, and Accuracy. *International Education Studies, 4*(2), 169-176.
17. Riemenschneider, A., Weiss, Z., Schröter, P., & Meurers, D. (2024). The Interplay of Task Characteristics, Linguistic Complexity, and Language Proficiency in High-Stakes English as a Foreign Language Writing. *Tesol Quarterly, 58*(2), 775-801.
18. Tabari, M. A., & Wang, Y. (2022). Assessing linguistic complexity features in L2 writing: Understanding effects of topic familiarity and strategic planning within the realm of task readiness. *Assessing Writing, 52*, 100605.
19. Vasylets, O., Gilabert, R., & Manchón, R. M. (2017). The effects of mode and task complexity on second language production. *Language Learning, 67*(2), 394-430.
20. Yan, X., Kim, H. R., & Kim, J. Y. (2020). Dimensionality of speech fluency: Examining the relationships among complexity, accuracy, and fluency (CAF) features of speaking performances on the Aptis test. *Language Testing, 38*(4), 485-510.

4.7.1 CAF onderzoek

- CAF staat voor complexiteit, accuraatheid en vloeiendheid.

Deze paragraaf, gebaseerd op Pallotti (2021), licht toe hoe complexiteit, accuraatheid en vloeiendheid worden geoperationaliseerd en onderzocht.

Bij het leren van een nieuwe taal krijgt de leerder in toenemende mate beschikking over taalelementen in het nieuwe taalsysteem. Complexiteit omvat lexicale *complexiteit*. Hierbij wordt gekeken naar de variatie van lexemen (minimale betekenisvolle eenheden) in de tekst. Lexicale complexiteit kan berekend worden op verschillende manieren, rekening houdend met tekstlengte. Een ander aspect van complexiteit is morfologische complexiteit: ook hier wordt gekeken naar het aantal verschillende morfologische processen in de tekst, zoals bijvoorbeeld verbuigingen, vervoegingen en samenstellingen. Als maat voor syntactische complexiteit wordt vaak gekeken naar de lengte van eenheden, bijvoorbeeld **T-units** (vergelijkbaar met zinnen), en het aantal en type bijzinnen.

Vloeiendheid zegt iets over met de mate van controle en beheersing over de nieuwe taal. CAF onderzoek kijkt vaak naar tempo, haperingen en herstel. Tempo kan worden geoperationaliseerd door het aantal syllaben per tijdseenheid te meten. Haperingen worden gemeten door het aantal en de lengte van pauzes. Er wordt ook gekeken naar het moment waarop de pauzes vallen. Herstel wordt in kaart gebracht door het meten van het aantal herhalingen, reparaties en valse starts. Dit geeft een beeld van de onzekerheid van de taalleerder en van het proces van monitoring van de taalproductie. Zie bijvoorbeeld ook Yan et al. (2020).

Accuraatheid vergelijkt het taalgebruik van de taalleerder met de standaardtaal en brengt afwijkingen in kaart. Zoals gezegd kan dit aspect eigenlijk niet als 'positieve beoordeling' worden beschouwd. Het kan conceptueel ook ter discussie worden gesteld, als we aannemen dat 'tussentaal' een normaal aspect is van taalontwikkeling. Daarnaast hangt accuraatheid sterk samen met vloeiendheid: accurate taalproductie hangt af van de mate waarin de vaardigheid in de nieuwe taal is geautomatiseerd. Accuraatheid wordt meestal geoperationaliseerd als het aantal fouten per eenheid (bijvoorbeeld per zin) of door te kijken naar de proportie foutloze eenheden.

CAF-onderzoek baseert zich overwegend op de volgende theoretische modellen, die de relatie tussen cognitieve complexiteit en kwaliteit van de taaluiting leggen (Kuiken & Vedder, 2007, 2011):

1. *The Limited Attentional Capacity Model*, ontwikkeld door Skehan en Foster. Dit model stelt dat taalgebruikers maar een beperkte hoeveelheid aandacht beschikbaar hebben tijdens het spreken of schrijven. Omdat deze aandacht niet voldoende is om alle aspecten van taal tegelijk optimaal te verwerken, moeten leerders voortdurend prioriteiten stellen tussen vloeiendheid, nauwkeurigheid en complexiteit. Wanneer de taak cognitief complexer wordt zullen leerders focussen op betekenis ten koste van de vorm.
2. *The Cognition Hypothesis*, ontwikkeld door Robinson. Deze stelt dat cognitief complexe taken leiden tot meer complex taalgebruik, omdat leerders daarbij meer aandacht moeten richten op vorm en betekenis. In tegenstelling tot Skehan verwacht Robinson dat een hogere cognitieve belasting niet ten koste gaat van taalprestatie, maar juist taalontwikkeling stimuleert door meer diepgaande verwerking.

De overeenkomst tussen beide theorieën is, dat ze ervan uitgaan dat leerders een beperkte hoeveelheid aandacht in kunnen zetten bij het uitvoeren van een taak. Ze verschillen wat betreft de voorspelling hoe deze aandacht zal worden gericht. Resultaten van CAF-onderzoek zijn tot op heden niet consistent, vanwege grote variatie in taakkenmerken en operationalisering van CAF in de verschillende onderzoeken (Pallotti, 2021). Bulté, Housen en Pallotti (2025) bepleiten daarom een beperkte interpretatie van complexiteit, gebaseerd op structurele, linguïstische eigenschappen. Daarnaast gebruiken ze de term 'moeilijkheid' om te verwijzen naar de cognitieve last die gepaard gaat met het leren en gebruiken van verschillende linguïstische elementen.

Als het gaat om toepasbaarheid van CAF voor toetsing wijst Pallotti (2021) allereerst op het feit dat CAF-maten geen rekening houden met de communicatieve context. Hij adviseert daarom om naast CAF-maten ook adequaatheid als criterium toe te voegen. Daarnaast werpt hij de vraag op hoe de CAF-maten moeten worden vastgesteld. Handmatig berekenen van CAF-maten is tijdrovend en vergt veel training. Automatisch scoren is mogelijk en gaat sneller, maar neemt communicatieve aspecten, zoals adequaatheid, niet mee.

Onderzoek van Lu (2017) keek onder andere naar het automatisch scoren van CAF. Hij onderzocht het concept grammaticale complexiteit en ging na hoe dit geoperationaliseerd wordt in schrijftheorie, in beoordelingsmodellen van highstakes toetsen en in drie verschillende geautomatiseerde analysemethodes (Biber-Tagger, Coh-Matrix en Syntactic Complexity Analyzer) die gebruikt worden in onderzoek naar syntactische complexiteit. De onderzoeksvraag luidde: hoe wordt syntactische complexiteit gedefinieerd in theoretische definities van schrijfvaardigheid, in beoordelingsmodellen die worden gebruikt in highstakes toetsen en in software die wordt gebruikt in onderzoek naar syntactische complexiteit? In theoretische zin is syntactische complexiteit de variatie en mate van complexiteit van syntactische structuren gebruikt in schrijfproducten. In toetsing zijn de beoordelingsmodellen de facto de operationalisering van het construct schrijfvaardigheid. Viervaardigheden-modellen gebruiken vaak de criteria spelling, woordenschat en grammatica. Er wordt dus belang gehecht aan syntaxis maar meer gedetailleerde analyse vindt doorgaans niet plaats. Grote highstakes toetsen voor het Engels (zoals TOEFL iBT, Cambridge en IELTS) noemen syntactische complexiteit en variatie in hun beoordelingsmodel. Ook softwarepakketten voor geautomatiseerde analyses van syntactische complexiteit operationaliseren complexiteit op verschillende manieren. Ondanks verschillen in aanpak laat onderzoek zien dat er een positieve relatie is tussen syntactische complexiteit en schrijfscores, onafhankelijk van welke software er wordt gebruikt.

4.7.2 Hoe hangt CAF samen met tweede-taalontwikkeling?

- Gedurende tweede en vreemde taalontwikkeling nemen complexiteit en vloeiendheid toe en neemt het aantal fouten af.

Een lijn van CAF-onderzoek kijkt naar de vraag hoe de complexiteit, accuraatheid en vloeiendheid van taalproductie zich ontwikkelt met de ontwikkeling van de vaardigheid in de tweede taal. Deze paragraaf beschrijft deze groep onderzoeken.

Onderzoek van Kuiken en Vedder (2019) keek naar syntactische complexiteit in relatie tot taalniveau, eerste en tweede taal, en de verschillen tussen talen. De participanten (N=91) in dit onderzoek waren taalleerders van het Nederlands, Italiaans en Spaans, met een taalniveau van A2 tot B1. Hun prestaties werden vergeleken met moedertaalsprekers van dezelfde talen. De participanten schreven twee betogende teksten en legden daarnaast een C-toets af. De schrijfproducten werden geanalyseerd op vier aspecten van syntactische complexiteit: algemene complexiteit, complexiteit door coördinatie (nevenschikking), complexiteit door subordinatie (onderschikking) en het gebruik van complexe naamwoordgroepen. Correlaties van de scores voor complexiteit en de **C-toets** werden berekend. Uit de resultaten bleek dat voor het Nederlands en het Spaans geen toegenomen complexiteit bij meer gevorderde leerders kon worden aangetoond. Voor Italiaans was dit wel het geval. Ook kon voor het Nederlands niet worden aangetoond dat moedertaalsprekers meer complexe uitingen produceerden. Voor Italiaans en Spaans was dat wel het geval (Kuiken & Vedder, 2019).

Onderzoek dat al eerder aan de orde kwam bij deelvraag 1, werd uitgevoerd door Plakans et al. (2019). De studie onderzocht de relatie tussen scores op de TOEFL iBT geïntegreerde lees-luister-schrijftaak (RLW) en de linguïstische kenmerken van het schrijfproduct. De onderzoekers wilden weten of CAF de score op een geïntegreerde taak voorspelt, en welke CAF-maten het meest voorspellend zijn. CAF werd als volgt geoperationaliseerd: voor complexiteit werden het aantal T-units (ongeveer vergelijkbaar met zinnen) geteld, voor accuraatheid werden grammaticale fouten geteld, verdeeld in syntactische en morfologische fouten. Als maat voor vloeiendheid werd de tekstlengte genomen. Uit de resultaten bleek dat vloeiendheid en complexiteit toenamen met de score, terwijl het aantal grammaticale fouten afnam met de score. Essays van hoger scorende kandidaten waren dus langer en complexer en met minder fouten. Het onderzoek liet dan ook zien dat CAF-maten de score op de taak kunnen voorspellen, met tekstlengte als beste voorspeller.

Onderzoek met een vergelijkbare vraag keek naar spreekvaardigheid (Révész et al., 2016). Hier was het doel na te gaan of CAF-maten voorspellend waren voor communicatieve adequaatheid. Er werden 500 spreekproducten, afkomstig van 80 leerders van het Engels en twintig moedertaalsprekers bij het onderzoek betrokken. De samples waren afkomstig van een universitaire toelatingstoets in de VS. Alle producten werden beoordeeld door getrainde beoordelaars op inhoud, begrijpelijkheid en register. Elk product werd door twee beoordelaars beoordeeld. Daarnaast werden alle producten met behulp van software getranscribeerd en daarna geanalyseerd op verschillende maten van complexiteit, accuraatheid en vloeiendheid. Alle CAF-maten bleken voorspellend voor communicatieve adequaatheid, met vloeiendheid als beste voorspeller.

Een ander onderzoek gericht op spreekvaardigheid onderzocht de relatie tussen vloeiendheid en taalniveau (Yan et al., 2020). De data voor dit onderzoek waren afkomstig van de Aptis spreektest. In deze toets krijgen kandidaten twee minuten om hun persoonlijke ervaringen te beschrijven en hun mening te geven over een abstract onderwerp. De uitslag wordt gerapporteerd op een schaal van 0 tot 50, en de scores worden gerelateerd aan de ERK-niveaus A1 tot en met C1. Voor elk van de ERK-niveaus werden 25 spreekproducten geselecteerd. Deze werden geanalyseerd op vloeiendheid. Er werd gekeken naar holistische maten van vloeiendheid (tempo, articulatie, pauzes); deze werden gemeten met software (Praat). Daarnaast werden alle spreekuitingen getranscribeerd en werd de plaats van haperingen verder geanalyseerd. Uit de resultaten bleek dat alle maten van vloeiendheid een af- of toenemende trend vertoonden gerelateerd aan het vaardigheidsniveau. Prestaties op hogere niveaus werden gekenmerkt door snellere spraak en langere uitingen. Op alle niveaus kwamen pauzes voor. Bij de hogere niveaus leken deze gebruikt te worden voor syntactische planning. Ook waren reparaties meer succesvol naarmate het niveau steeg.

Veel CAF-onderzoek kijkt naar leerders van het Engels. Iwashita (2006) onderzocht de vraag of wat opgaat voor het Engels ook geldt voor andere talen, hier het Japans. Het onderzoek keek naar syntactische complexiteit van prestaties op spreektaken. De participanten in het onderzoek waren 33 leerders van het Japans. Zij voerden spreekopdrachten uit en legden daarnaast een gestandaardiseerde taaltoets af. Uit de analyses bleek dat syntactische complexiteit van de spreekproducten toenam met vaardigheid. De lengte van T-units en het aantal bijzinnen per T-unit bleken de beste voorspellers van spreekvaardigheid in het Japans. Deze bevinding komt overeen met bevindingen uit ander onderzoek, gericht op het Engels.

Terwijl toenemende taalvaardigheid verondersteld wordt gepaard te gaan met toenemende syntactische complexiteit van de taalproductie, is er maar weinig onderzoek gedaan naar hoe deze toegenomen complexiteit tot uiting komt in de beoordeling van schrijfvaardigheid door menselijke beoordelaars. Crossley en McNamara (2014) keken daarnaar in een longitudinaal onderzoek op basis van schrijfproducten van 57 leerders van het Engels als tweede taal. Ze verzamelden over de loop van een semester van elke participant drie schrijfproducten. De producten werden geanalyseerd op elf syntactische kenmerken met behulp van software (Coh-metrix). Daarnaast werden de producten geanalyseerd door getrainde beoordelaars met behulp van een model, dat onder andere keek naar syntaxis. Met behulp van de software kon worden vastgesteld dat gedurende het semester de syntactische complexiteit van de producten sterk toenam en meer in lijn kwam met de typerende kenmerken van academische schrijfvaardigheid. Dit bevestigt de relatie tussen taalvaardigheid en syntactische complexiteit. Slechts één aspect van de automatische score was voorspellend voor de menselijke beoordeling, de afname van het aantal bijzinnen in de tekst. Een ander aspect, het voorkomen van bijzinnen met ‘dat’ en ‘om te’, werd door de beoordelaars in verband gebracht met toegenomen kwaliteit van de schrijfproducten. Echter, uit onderzoek blijkt dat deze constructies typisch zijn voor gesproken taal en niet voor academische schrijfvaardigheid. Met andere woorden: de beoordelaars signaleerden veel aspecten van de toegenomen complexiteit niet en daarnaast beloonden ze aspecten die niet construct-relevant zijn. Dat betekent dat de toegenomen schrijfvaardigheid van de leerders niet adequaat beloond werd in het assessment van hun producten.

Ook Kuiken en Vedder (2014) keken naar de rol van beoordelaars in relatie tot de linguïstische complexiteit van schrijfproducten. Het onderzoek vergeleek de oordelen van menselijke beoordelaars met een analyse van syntactische complexiteit, lexicale diversiteit en adequaatheid, en keek bovendien naar de afwegingen die beoordelaars maken. De schrijfproducten in deze studie waren afkomstig van 39 leerders van het Italiaans, 18 moedertaalsprekers van het Italiaans, 32 leerders van het Nederlands en 17 moedertaalsprekers van het Nederlands. Elke participant leverde twee argumenterende teksten. In totaal deden zeven ervaren, getrainde beoordelaars mee aan het onderzoek, drie voor Italiaans en vier voor Nederlands. De beoordelaars gebruikten voor dit doel ontwikkelde globale schalen en beoordeelden communicatieve adequaatheid en linguïstische complexiteit. Na de beoordeling van alle producten vonden panelgesprekken plaats waarin de beoordelaars hun beoordelingen beargumenteerden. Uit de resultaten bleek ten eerste dat de oordelen van beoordelaars over communicatieve adequaatheid significant samenhingen met hun oordelen over linguïstische complexiteit. Ook correleerde de scores van beoordelaars met betrekking tot linguïstische complexiteit met de analyses van lexicale diversiteit en accuraatheid, maar niet met de syntactische complexiteit. Uit het panelgesprek met beoordelaars bleek dat beoordelaars veel waarde hechtten aan heldere formulering, overtuigende voorbeelden, afwezigheid van onnodige herhalingen, gebruik van signaal- en verbindingswoorden en een heldere conclusie. Ook creativiteit en het gebruik van laagfrequente woorden werd

gewaardeerd. Kenmerken van linguïstische complexiteit zoals zinsstructuren, bijzinnen, en voorzetsels werden minder belangrijk geacht. Ook accuraatheid (grammaticale correctheid) werd niet erg belangrijk gevonden zolang de begrijpelijkheid niet werd aangetast.

Syntactische complexiteit hangt samen met ontwikkeling van vaardigheid in de tweede taal. Recent onderzoek laat echter zien dat die toename van complexiteit niet altijd lineair verloopt en dat de ontwikkeling van lexicon, morfologie en syntaxis niet altijd gelijktijdig verlopen. Bulté en Housen (2018) waren daarom geïnteresseerd in individuele verschillen en in groepspatronen in de ontwikkeling van leerders. Ze volgden de ontwikkeling van tien jonge, beginnende leerders van het Engels gedurende twee jaar. Van elke participant waren elf schrijfproducten beschikbaar. De gehanteerde maten voor complexiteit waren: gemiddelde lengte van de T-units, lengte van hoofd- en bijzinnen en van naamwoordgroepen. Uit de analyse bleek dat er geen duidelijk patroon zichtbaar was met betrekking tot de lengte van hoofdzinnen en nevenschikkende bijzinnen. Deze werden niet systematisch langer over de loop van twee jaar. Voor de aspecten onderschikkende bijzinnen, lengte van T-units en naamwoordgroepen was een opwaartse trend zichtbaar: de producten werden meer complex over tijd. Inzoomend op de individuele leerders was een oplopende trend zichtbaar, de complexiteit van de schrijfproducten nam gemiddeld toe. De individuele verschillen waren echter groot en leerlingen ontwikkelden hun syntactische vaardigheden op verschillende manieren.

Een onderzoek dat relevant is voor de staatsexamencontext vergeleek de prestaties van Nt2-leerders (N=181) op acht spreektaken met hun linguïstische competenties (Hulstijn et al., 2012). Het onderzoek had tot doel om de niveaus B1 en B2 van het Europees Referentiekader (ERK) nader te onderzoeken, omdat de validering van het ERK vrij beperkt is: er zijn geen leerders-samples gebruikt bij de ontwikkeling en er is met betrekking tot spreekvaardigheid ook niet een stevig theoretisch model waarop het ERK is gestoeld. De vragen die in dit onderzoek centraal stonden waren: (1) In welke mate zijn kennis van woordenschat en grammatica in staat leerders in te delen naar vaardigheid; (2) Hoe verschillen leerders op B1 en B2 wat betreft hun woordenschat en (3) Hoe verschillen B1ers en B2ers met betrekking tot grammaticale beheersing. Participanten legden acht voor het doel geconstrueerde spreektaken af, die in moeilijkheid verschilden op de aspecten complexiteit, formaliteit en taalhandeling (beschrijven versus argumenteren). Daarnaast legden ze tests af om hun productieve woordenschat, verwerkingssnelheid en uitspraak in kaart te brengen. De spreektaken werden beoordeeld op communicatieve adequaatheid door in totaal twaalf beoordelaars, waarbij elke prestatie werd beoordeeld door vier beoordelaars. Het onderzoek vond dat linguïstische kennis, met name woordenschat, grammatica, verwerkingssnelheid en uitspraak goede voorspellers zijn voor het taalniveau; woordenschat was daarbij de beste voorspeller. Participanten op niveau B1 bleken (naar schatting) gemiddeld 4000 en de B2ers gemiddeld 7000 woorden te beheersen. Ook bleken de B2ers significant beter te presteren op de grammaticatoets dan de B1ers. Het onderzoek biedt enige onderbouwing voor de validiteit van de taalniveaus B1 en B2 van het ERK. Verder is een interessante bevinding dat de acht taken, die ontwikkeld waren als verschillend in moeilijkheid, allemaal goed in staat waren de leerders te verdelen in een B1groep en een B2groep. Het lijkt dus niet nodig spreektaken specifiek te ontwikkelen om een prestatie op een bepaald niveau uit te lokken.

4.7.3 Hoe hangt CAF samen met taakkenmerken?

- Onderzoek naar taakkenmerken in relatie tot CAF vindt geen eenduidige resultaten.
- Wel wordt duidelijk dat taken invloed kunnen hebben op de prestaties van kandidaten.

De tweede hoofdlijn in het CAF-onderzoek kijkt naar de relatie tussen taak en taalproductie (Pallotti, 2021). Taakkenmerken die invloed kunnen hebben op de productie zijn ander andere register, genre, taaktype, en onderwerp (Riemenschneider et al., 2024).

Kuiken en Vedder (2007) onderzochten de invloed van taakcomplexiteit in een onderzoek onder studenten Italiaans (n=84) en Frans (n=75). De studenten voerden elk twee schrijftaken uit, die gemanipuleerd waren in termen van cognitieve complexiteit. Voor beide taken schreven de participanten een brief over een vakantiebestemming. In de eenvoudige taak moesten ze drie en in de complexe taak zes elementen verwerken. Participanten maakten daarnaast een **cloze-test** waarmee hun taalniveau werd gemeten. De producten werden geanalyseerd op accuraatheid, syntactische complexiteit, en lexicale variatie. Uit de resultaten bleek ten eerste dat er invloed is van het niveau van de participant op de complexiteit van de taakuitvoering. Er bleek ook invloed te zijn van de taak: de studenten maakten minder lexicale fouten in de complexe taak. Er was maar in zeer beperkte mate interactie tussen niveau van de studenten en taakcomplexiteit: alleen voor de studenten Frans en alleen voor spelling werden er meer fouten gevonden in de complexe taak.

Ook Rezazadeh et al. (2011) onderzochten de invloed van de schrijftaak op de complexiteit, adequaatheid en vloeiendheid van schrijfproducten. De participanten in dit onderzoek waren 168 half-gevorderde leerders van het Engels die studeerden aan een universiteit in Iran. De studenten werden verdeeld in twee groepen, en elke groep maakte een schrijftaak. De ene subgroep schreef een instructieve tekst waarin ze de weg van het station naar de universiteit moesten uitleggen aan een vriend. De andere groep schreef een argumenterende tekst over de negatieve invloed van toetsen op onderwijs. De instructie werd verondersteld minder cognitief complex te zijn, omdat het een alledaagse situatie betrof. De schrijfproducten werden geanalyseerd. Eerst werd het aantal T-units bepaald (hoofdzin en bijbehorende bijzinnen). Als maat voor vloeiendheid werd het aantal woorden, aantal T-units en bijzinnen per tekst genomen. Als maat voor accuraatheid werd gekeken naar de proportie foutloze t-units en de proportie foutloze bijzinnen. Als maat voor complexiteit werd gekeken naar aantal en type bijzinnen. Uit de resultaten bleek dat de instructieve tekst leidde tot betere schrijfproducten in termen van vloeiendheid en accuraatheid. De argumenterende teksten scoorden hoger op complexiteit.

Waar de focus van Rezazadeh et al. (2011) lag op schrijfproducten, trachtten Révész et al. (2017) inzicht te krijgen in de cognitieve processen van tweede-taalleerders. Het onderzoek keek tijdens het schrijven naar vloeiendheid, pauzes, en revisie-gedrag in relatie tot taakkenmerken. De participanten waren 73 gevorderde leerders van het Engels (niveau C1, gemeten met IELTS, band 7-7.5) op een universiteit in het VK. De participanten voerden een eenvoudige of een complexe schrijftaak uit. De eenvoudige taak bood input in de vorm van ideeën die konden worden verwerkt in de tekst, bij de complexe taak ontbrak deze input. Het gedrag van de participanten werd in kaart gebracht met behulp van software die het toetsenbordgebruik volgde. Daarnaast vulden de participanten een vragenlijst in. De geproduceerde teksten werden geanalyseerd op lexicale diversiteit (woordfrequentie en -variatie) en syntactische complexiteit (zinslengte, bijzinnen, en zinscomplexiteit). De resultaten lieten zien dat de taakcomplexiteit het gedrag van leerders beïnvloedt: met name verschillen in aantal pauzes en revisies was significant hoger in de complexe versie van de taak. De eenvoudigere taak leidde tot meer geavanceerd woordgebruik en hogere syntactische complexiteit. Ook gaven de kandidaten aan de taak met input makkelijker te vinden. De onderzoekers vermoeden dat de geschreven input meer verwerkingsruimte vrijlaat voor aandacht voor linguïstische processen, leidend tot een beter uitgewerkt schrijfproduct.

Mora en collega's focusten op de affectieve aspecten van het spreken. Ze wilden weten hoe taakcomplexiteit de vloeiendheid, accuraatheid, begrijpelijkheid en uitspraak van taalleerders beïnvloeden. Daarnaast wilden ze onderzoeken wat de invloed van taakcomplexiteit is op spreekangst (Mora et al., 2024). Spaanse leerders van het Engels (N=42) voerden een eenvoudige en een complexe spreektaak uit. In de taak moesten de participanten de rol van brandweercommandant vervullen en aangeven hoe ze een brandend pand zouden ontruimen. Taakcomplexiteit werd gemanipuleerd door de hoeveelheid elementen die ze in hun oplossing moesten betrekken. De hartslag van de participant werd gemeten als maat voor spreekangst. Daarnaast rapporteerden de deelnemers direct na de taakuitvoering over de spreekangst die ze hadden ervaren. Uit het onderzoek bleek dat de uitingen meer vloeiend en begrijpelijk waren in de eenvoudige taak, en meer lexicaal en grammaticaal accuraat in de complexe taak, in overeenstemming met Robinson's Cognition Hypothesis. Er was geen relatie tussen de gemeten en gerapporteerde hartslag en in tegenstelling tot de verwachtingen werd er ook geen relatie gevonden tussen taakcomplexiteit en spreekangst.

Vasylets et al. (2017) maakten gebruik van dezelfde 'brandweercommandant-taak', in een onderzoek dat tot doel had taakcomplexiteit te vergelijken in schrijf- en spreekproducten. De participanten in dit onderzoek waren 78 leerders van het Engels van twee Spaanse universiteiten. De helft deed de eenvoudige en de complexe versie van de taak mondeling, de andere helft schriftelijk. Uit de resultaten bleek een effect van respectievelijk spreken of schrijven op de complexiteit van de producten. De sprekers werkten meer ideeën uit. De producten van de schrijvers vertoonden langere zinnen, meer bijzinnen, grotere lexicale diversiteit en ideeën werden meer uitgediept.

Complexiteit van een taak kan worden gemanipuleerd door het aantal elementen dat in de taak moet worden verwerkt, maar ook de mate van bekendheid met het onderwerp waarover moet worden geschreven kan een rol spelen. Daarnaar keken Tabari en Wang (2022). Zij onderzochten de invloed van vertrouwdheid met het onderwerp op de complexiteit van schrijfproducten en keken daarnaast ook naar de rol van voorbereidingstijd. Participanten waren 64 leerders van het Engels aan een universiteit in de VS. Ze voerden elk een schrijftaak uit over een vertrouwd en een niet-vertrouwd onderwerp. De helft van de participanten kreeg wel en de andere helft kreeg geen voorbereidingstijd. De producten werden geanalyseerd op 11 aspecten van lexicale en syntactische complexiteit.

Uit de resultaten bleek dat vertrouwdheid met het onderwerp een groot effect had op de syntactische kwaliteit van de schrijfproducten, en bovendien enig effect op de lexicale complexiteit. Planningstijd droeg bij aan syntactische complexiteit maar had geen effect op woordgebruik en -variatie.

Riemenschneider et al., (2024) tot slot, keken naar de interactie tussen vaardigheid, taak, en complexiteit. Ze combineerden dus de twee belangrijke lijnen van CAF-onderzoek. De participanten waren 362 Duitse kandidaten die een schrijftoets Engels aflegden voor toegang tot de universiteit. De schrijftaak bestond uit drie opdrachten, gerelateerd aan een leestekst. De kandidaten schreven een samenvatting, daarna volgde een analyse van bepaalde aspecten van de leestekst en tot slot moest de kandidaat een argumentatie schrijven. De vaardigheid van de kandidaat werd berekend door hun schoolcijfers voor Engels van de laatste twee jaar te middelen. Linguïstische complexiteit werd gemeten met software (CTAP: Common Tekst Analysis Platform) die 319 aspecten van linguïstische complexiteit in kaart bracht. Deze maten hebben betrekking op lexicale, syntactische, morfologische complexiteit en op cohesie, taalverwerking en tekstlengte. Uit de analyse bleek dat de taken verschillen in de complexiteit van de productie die ze uitlokken. De derde (argumenterende) schrijftaak, die een groot beroep deed op onafhankelijke taalproductie, was het best in staat de verschillen in vaardigheid tussen leerders te meten.

De teksten met betrekking tot taakkenmerken verschillen in focus, in manipulatie van taakcomplexiteit en in bevindingen. De bevindingen worden samengevat in Tabel 11. De samenvatting maakt duidelijk dat onderzoek met betrekking tot het effect van de taak op complexiteit van de producten zowel in methodologisch opzicht als wat betreft de resultaten niet eenduidig is.

Tabel 11 *Taakkenmerken in relatie tot complexiteit*

Onderzoekers	Vaardigheid	Manipulatie van complexiteit	Eenvoudige taak	Complexe taak
Kuiken en Vedder (2007)	schrijven	Aantal te beschrijven elementen		Minder lexicale fouten
Rezazadeh et al. (2011)	schrijven	Tekstsoort (instructie vs argumentatie)	Betere vloeiendheid en accuraatheid	Grotere complexiteit
Révész et al. (2017)	spreken	Mate van ondersteuning, ideeën wel/niet aangereikt	Meer geavanceerd woordgebruik en grotere syntactische complexiteit	Meer haperingen
Mora et al. (2024)	spreken	Aantal te bespreken elementen	Betere vloeiendheid en begrijpelijkheid	Betere grammaticale en lexicale accuraatheid
Tabari en Wang (2022)	spreken	Vertrouwdheid van het onderwerp en voorbereidingstijd	Meer geavanceerd woordgebruik en grotere syntactische complexiteit a.g.v. vertrouwdheid van onderwerp. Grotere syntactische complexiteit a.g.v. voorbereidingstijd.	
Riemenschneider et al. (2024)	schrijven	Tekstsoort		Grotere complexiteit.

4.7.4 Kritische reflectie

De vraag die in dit hoofdstuk centraal stond was, wat er bekend is over positieve beoordelingsmodellen en of deze worden ingezet in highstakes examens. De literatuurstudie is toegespitst op complexiteit, accuraatheid en vloeiendheid (CAF), waarvan met name de aspecten complexiteit en vloeiendheid als 'positief' te kwalificeren zijn (Pallotti, 2021). Onderzoek naar CAF laat een positieve relatie zien tussen CAF-maten en vaardigheid van leerders. Met andere woorden: taalleerders op hoger niveau gebruiken complexere structuren, een grotere variatie van woorden, en hun taalgebruik is vloeiender. Ook maken ze minder fouten. CAF-onderzoekers waarschuwen

echter voor het uitsluitend beoordelen met CAF-maten en menen dat er ook altijd taak-specifieke criteria voor communicatieve adequaatheid moeten worden gebruikt.

CAF-onderzoekers verschillen in het aantal CAF-maten dat wordt geanalyseerd. De analyses kunnen zeer fijnmazig zijn. Riemenschneider et al. (2024) bijvoorbeeld analyseerden 319 verschillende CAF-maten in hun onderzoek. Handmatige beoordeling in de context van toetsing en examinering is dan ook niet goed denkbaar. Wel zijn er verschillende geautomatiseerde modellen voor de analyse van CAF, die mogelijk inzetbaar zijn.

Een deel van het CAF-onderzoek is gericht op de effecten van de taak op de complexiteit van de taakuitvoering. Taken worden gemanipuleerd in termen van cognitieve last, bijvoorbeeld het aantal elementen dat verwerkt moet worden, de vertrouwdheid van onderwerpen, voorbereidingstijd, en tekstsoort. Hoewel dit onderzoek geen eenduidige resultaten laat zien, wordt het wel (wederom) duidelijk dat het effect van taken op de prestatie van kandidaten groot kan zijn. Deze bevindingen geven misschien wat richting aan het beter begrip van taakeffecten.

Hoewel een deel van de besproken CAF-onderzoeken gebruik maakte van toetsdata, kwam er in deze selectie van literatuur geen tekst voor die CAF beschreef als deel van de beoordeling van kandidaatprestaties, dus in de context van uitslagverstrekking. In paragraaf 4.7 is echter besproken dat bij de beoordeling van TOEFL iBT gebruik wordt gemaakt van automatische scoring naast menselijke beoordeling en ook bijvoorbeeld Duolingo English Test wordt volledig automatisch gescoord, zie paragraaf 4.3.4. Ook paragraaf 4.9, die de vraag naar holistisch beoordelen bespreekt, zal mogelijk nuttige aanvullende informatie opleveren. In een holistische beoordeling wordt een (meer of minder expliciete) impressie van zaken als syntactische complexiteit en woordgebruik meegenomen.

4.8 Dichotoom beoordelen

Bij de beoordeling van taalproductie zijn meestal menselijke beoordelaars betrokken. Scores op spreek- en schrijftoetsen worden daarom niet alleen beïnvloed door het niveau van de kandidaat, maar ook door de beoordelaars en de gehanteerde beoordelingsmodellen. Beoordelaars zijn namelijk een bron van variatie of 'ruis', bijvoorbeeld als gevolg van hun streng- of mildheid, ervaring, achtergrond, of verschillen in toepassing van criteria. Training van beoordelaars helpt, maar elimineert de variatie niet (Lukácsi, 2021; O'Grady, 2019, 2025; H. Park & Yan, 2019). Deze paragraaf bespreekt onderzoek naar dichotome beoordelingsmodellen voor het beoordelen van de productieve vaardigheden spreken en schrijven. Volgens Khamboonruang (2022) is er weinig onderzoek gedaan naar binaire beoordelingsmodellen maar suggereert het onderzoek dat er is, dat binaire schalen makkelijker toepasbaar zijn door beoordelaars en kunnen leiden tot meer consistente beoordeling. De onderzoeksvraag luidt: Wat is er bekend over dichotoom beoordelen? Zijn er examens die werken met dichotome beoordelingsmodellen voor het beoordelen van open opgaven Spreken en Schrijven? En hoe zitten die beoordelingsmodellen in elkaar?

De literatuur die is gebruikt voor deze deelvraag is -deels- gevonden met de zoekterm "rating scales" AND "language testing" AND "dichotomous". Omdat de zoekopdracht weinig relevante teksten opleverde, is de lijst aangevuld door gebruik te maken van referenties in de gelezen teksten. Vanwege de beperkte beschikbaarheid van relevante literatuur zijn bij deze vraag minder dan twintig teksten verwerkt. De gebruikte teksten zijn te vinden in Tabel 12. De gevonden modellen zijn te verdelen in checklists en beslisbomen. Daarnaast is een korte paragraaf opgenomen over een psychometrische analyse van schalen met verschillende lengte.

Tabel 12 *Bestudeerde literatuur deelvraag 7*

1. Abbasi Sardari, S., Ahmadi, A., & Ravand, H. (2025). Creating an empirically derived checklist to diagnose second language learners' integrated writing ability. *Language Testing in Asia, 15*(1), 52.
2. Chou, M. H. (2013). Teacher Interpretation of Test Scores and Feedback to Students in EFL Classrooms: A Comparison of Two Rating Methods. *Higher Education Studies, 3*(2), 86-95.
3. Deygers, B., Van Gorp, K., Joos, S., & Luyten, L. (2011, July). Rating scale design: a comparative study of two analytic rating scales in a task-based test. In *Exploring Language Frameworks. Proceedings of the ALTE Kraków Conference* (pp. 271-287).
4. East, M., & Young, D. (2007). Scoring L2 writing samples: Exploring the relative effectiveness of two different diagnostic methods. *New Zealand Studies in Applied Linguistics, 13*(1), 1-21.

5. Effatpanah, F., Baghaei, P., & Boori, A. A. (2019). Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia, 9*(1), 12.
6. Fulcher, G., Davidson, F., & Kemp, J. (2011) Effective rating scale development for speaking tests: Performance Decision Trees. Language Testing 28(1), 5-29.
7. Hirai, A., & Koizumi, R. (2008). Validation of the EBB scale: A case of the story retelling speaking test. *JLTA Journal Kiyo, 11*, 1-20.
8. Khamboonruang, A. (2022). Building an Initial Validity Argument for Binary and Analytic Rating Scales for an EFL Classroom Writing Assessment: Evidence from Many-Facets Rasch Measurement. *rEFLections, 29*(3), 675-699.
9. Kim, Y.-H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified Model. *Language Testing, 28*(4), 509-541.
10. Lukácsi, Z. (2021). Developing a level-specific checklist for assessing EFL writing. *Language Testing, 38*(1), 86-105.
11. O'Grady, S. (2019). The impact of pre-task planning on speaking test performance for English-medium university admission. *Language Testing, 36*(4), 505-526.
12. O'Grady, S. (2025). Task design and rater effects in task-based language assessment. *TESOL Journal, 16*(1), e904.
13. Park, H., & Yan, X. (2019). An investigation into rater performance with a holistic scale and a binary, analytic scale on an ESL writing placement test. *Papers in Language Testing and Assessment, 8*(2), 34-64.
14. Safari, F., & Ahmadi, A. (2023). Developing and evaluating an empirically-based diagnostic checklist for assessing second language integrated writing. *Journal of Second Language Writing, 60*, 101007.
15. Shi, X., Ma, X., Du, W., & Gao, X. (2024). Diagnosing Chinese EFL learners' writing ability using polytomous cognitive diagnostic models. *Language Testing, 41*(1), 109-134.
16. Shirazi, M. A. (2013). Using an analytic dichotomous evaluation checklist to increase interand intrarater reliability of EFL writing evaluation. *Iranian Journal of Applied Linguistics (IJAL), 16*(1), 25-57.
17. Turner, C. E., & Upshur, J. A. (1995). Some Effects of Task Type on the Relation between Communicative Effectiveness and Grammatical Accuracy in Intensive ESL Classes. *TESL Canada Journal, 12*(2), 18-31.
18. Weaver, C. (2006). Evaluating the use of rating scales in a highstakes Japanese university entrance examination. *SPAAN FELLOW, 1001*, 57.
19. Yokouchi, Y. (2025). Revisiting the effectiveness of a performance decision tree-style rubric compared to a grid-style rubric. *Language Testing in Asia, 15*(1), 8.

4.8.1 Checklists

- Checklists die een groot aantal aspecten onderscheiden kunnen de basis bieden voor genuanceerde feedback aan taalleerders.
- Maar dichotome oordelen worden vaak als weinig genuanceerd ervaren.

Sommige auteurs duiden hun binaire beoordelingsmodel aan als 'checklist'. Een checklist is een gedetailleerde lijst met omschrijvingen die gebruikt kunnen worden om te beoordelen in hoeverre kandidaten verschillende vaardigheden tonen (Sardari et al., 2025). Het ERK beveelt het gebruik van checklists aan als prestaties op één bepaald niveau moeten worden beoordeeld (Lukácsi, 2021).

Volgens Khamboonruang (2022) is er weinig onderzoek gedaan naar binaire beoordelingsmodellen maar suggereert het onderzoek dat er is, dat binaire schalen makkelijker toepasbaar zijn door beoordelaars en kunnen leiden tot meer consistente beoordeling. Het onderzoek beschrijft het functioneren van twee nieuw ontwikkelde beoordelingsmodellen, een analytisch en een binair beoordelingsmodel voor de beoordeling van schrijfvaardigheid in het Engels. De binaire schaal bestond uit een checklist met 22 criteria die betrekking hadden op zeven domeinen van schrijfvaardigheid en die elk met 0 of 1 beoordeeld moesten worden. De categorieën zijn: hoofdidee, ondersteunende ideeën, ondersteunende details, conclusie, cohesie, zinsgebruik, woordgebruik. De analytische schaal beoordeelde elk van deze categorieën op een driepuntschaal. Er waren 51 schrijfproducten betrokken bij het onderzoek en deze werden met beide modellen beoordeeld door drie getrainde beoordelaars. **Many-facet Rasch-analyses** werden uitgevoerd om te kijken naar de beoordelaar, de vaardigheid van de leerder en de beoordelingscriteria. Uit de analyses bleek dat beide schalen het construct schrijfvaardigheid accuraat konden meten. De beoordelaars beoordeelden consistent en de beoordelaarsovereenstemming was hoog (.86 voor de binaire schaal en .81 voor het analytisch model). Hoewel beide schalen goed functioneerden verdient volgens de auteur de binaire schaal de voorkeur vanwege iets beter psychometrisch functioneren en lagere cognitieve last voor beoordelaars.

In een poging te komen tot een betrouwbaar instrument voor de beoordeling van schrijfvaardigheid in het Engels, ontwikkelde Shirazi (2013) een uitgebreide checklist, genaamd Analytic Dichotomous Evaluation Checklist (ADEC). De checklist bevatte 68 item in vijf sub-schalen: inhoud, organisatie, grammatica, woordenschat en spelling & interpunctie, en kwam tot stand door literatuuronderzoek en door hardop-denksessies met beoordelaars. Beoordelaars beoordeelden de aan- of afwezigheid van een bepaald kenmerk in het schrijfproduct. De prestaties van de checklist werden vergeleken met een bestaand holistisch beoordelingsmodel dat vijf niveaus onderscheidde. Bij het onderzoek waren acht beoordelaars betrokken, die twintig schrijfproducten in totaal vier keer beoordeelden, twee keer met ADEC en twee keer met het holistische model, met tijd tussen elk van de beoordelingssessies. Met behulp van Rasch-analyses werd gekeken naar **beoordelaarsovereenstemming (interbeoordelaarsbetrouwbaarheid)** en **intra-beoordelaarsbetrouwbaarheid**. De beoordelingen met behulp van de checklist lieten een betere beoordelaarsovereenstemming zien en ook de intra-beoordelaarsbetrouwbaarheid (de interne consistentie van beoordelaars) was overwegend beter.

Een onderzoek van H. Park en Yan (2019) vergeleek de betrouwbaarheid van een holistisch beoordelingsmodel met een dichotoom analytisch model. De dataset bestond uit 452 essays afkomstig van de English Placement Test, afgenomen aan een Amerikaanse universiteit. Er waren negen getrainde beoordelaars betrokken bij het onderzoek. Het onderzoek verliep in twee fases. In de eerste fase beoordeelden de beoordelaars met behulp van het holistische beoordelingsmodel. De globale criteria betroffen kwaliteit van de argumentatie en lexicale en grammaticale aspecten (lexico-grammar). Het model deelde kandidaten in op vier niveaus, A, B, C en D, met A als het hoogste niveau. Daarnaast namen de beoordelaars deel aan een interview en een hardop-denksessie. Uit het onderzoek bleek dat de beoordelaarsovereenstemming zeer laag was. De beoordelaars bleken ook na training hun eigen uiteenlopende criteria toe te voegen aan het model. Ook focusten ze eenzijdig op de criteria voor argumentatie, omdat ze de criteria voor lexicale en grammaticale aspecten moeilijk toepasbaar vonden. Op basis van deze informatie werd in fase 2 een analytisch model ontwikkeld. Het analytische model splitste de elementen argumentatie en lexicale en grammaticale elementen uit in verschillende deelvragen die met ja of nee beantwoord konden worden, bijvoorbeeld: ‘Is het essay makkelijk te begrijpen?’; ‘Zijn er grote grammaticale fouten?’ en ‘Is het woordgebruik vaak passend?’. De beoordelaars gebruikten deze schaal met een nieuwe set schrijfproducten en de beoordelaarsovereenstemming bleek sterk verbeterd. De beoordelaars gaven ook aan dat ze de analytische, dichotome schaal makkelijker vonden om te gebruiken.

Kim (2011) onderzocht het psychometrisch functioneren van een checklist met 35 descriptoren. De checklist, genaamd **Empirically-derived Descriptor-based Diagnostic checklist (EDD)**, werd voor het onderzoek ontwikkeld. De 35 items waren gerelateerd aan vijf deelaspecten van schrijfvaardigheid: inhoud, organisatie, grammatica, woordgebruik en spelling & interpunctie. De vraag was of deze checklist goed onderscheid kon maken tussen vaardige en minder vaardige leerders, of de beoordeling betrouwbaar was, en of de lijst de basis kon zijn voor feedback aan leerders. Er werden in totaal 480 schrijfproducten, afkomstig van TOEFL, beoordeeld door 10 docenten met behulp van de checklist, waarbij elke beoordelaar een subset van de schrijfproducten beoordeelde. Analyses lieten zien dat de checklist overwegend goed functioneerde en dat leerders met dezelfde somscores sterk kunnen verschillen in hun sterktes en zwaktes over de verschillende aspecten. De checklist kan daarom een goede basis zijn voor feedback aan leerders.

Effatpanah et al. (2019) repliceerden het onderzoek van Kim (2011). Zij combineerden gebruik van de EDD-checklist met een specifieke statistische techniek om te komen tot gedetailleerde leerdersprofielen die als basis kunnen dienen voor feedback op schrijfvaardigheid. De schrijfprestaties voor dit onderzoek waren afkomstig van 500 studenten Engels aan universiteiten in Iran. Vier getrainde beoordelaars beoordeelden elk 125 prestaties. Een set van 147 teksten werd door twee beoordelaars beoordeeld, met een overeenstemming van .82. Het gebruikte statistische model was een **Cognitive Diagnostic Modeling (CDM)** model. Dit type modellen breekt een taak op in meerdere strategieën en processen (hier gebaseerd op de checklist) en schat de vaardigheid van de kandidaat op elk van de subvaardigheden. Uit de analyses bleek dat Iraanse studenten vooral moeite hebben met de aspecten inhoud en woordenschat, en minder moeite met grammatica, organisatie en spelling. De auteurs concludeerden daaruit dat inhoud en woordgebruik de meest cruciale aspecten van academisch schrijven zijn.

Luckácsi (2021) beschreef de ontwikkeling van een checklist voor Euroexam International, een toetsontwikkelaar die toetsen op verschillende niveaus ontwikkelt voor het beoordelen van Engels en Duits. Het onderzoek richtte zich op de ontwikkeling van een checklist voor het beoordelen van Schrijven op niveau B2. Het schrijfexamen

B2 bestaat uit twee opdrachten die moeten worden uitgevoerd in 60 minuten. Het oorspronkelijk gehanteerde beoordelingsmodel onderscheidde de aspecten taakuitvoering, register & stijl, coherentie, cohesie, grammaticale range & accuraatheid, en lexicale range & accuraatheid, en hanteerde een vijfpuntschaal. Het probleem met dit beoordelingsmodel was een sterke centrale tendens, waarbij de uiteinden van de schaal nauwelijks werden toegekend. De centrale onderzoeksvraag was daarom of een niveau-specifieke checklist kon worden ontwikkeld die de verschillende gradaties van schrijfvaardigheid van kandidaten op B2 reflecteert. Op basis van zowel het ERK als een set schrijfproducten werd een concept checklist met 36 stellingen ontwikkeld, getest en bijgesteld. Voorbeelden van stellingen zijn: de tekst is leesbaar, de schrijver geeft een duidelijke beschrijving van het probleem, de tekst is helder en beknopt. Vervolgens beoordeelden veertien beoordelaars in totaal 184 schrijfproducten, waarbij elke beoordelaar 60 teksten beoordeelde. Na data-analyse werd één item verwijderd dat niet bleek bij te dragen aan de beslissing. De checklist leidde tot een meer betrouwbare beoordeling die inhoudelijk nauw aansluit bij het ERK en daarmee valide wordt geacht. Beoordelaars gaven aan dat de checklist met 35 items relatief snel en makkelijk toepasbaar was.

Analytische beoordelingsmodellen worden geacht – in vergelijking met holistische modellen- een betere basis te zijn voor feedback aan kandidaten. Maar de brede categorieën in de meeste analytische modellen maken het geven van diagnostische feedback ook maar beperkt mogelijk. Dit motiveerde Sardari et al. (2025) een checklist te ontwikkelen voor de beoordeling van één van de schrijftaken van IELTS, de taak waarin de kandidaat een grafiek moet beschrijven. Het project kende twee fases. In de eerste fase werd op basis van hardop-denksessies met zes ervaren docenten een checklist ontwikkeld. Omdat in een dichotome beoordeling mogelijk informatie verloren gaat of te veel wordt vereenvoudigd, werden twee versies ontwikkeld, met een dichotome schaal per descriptor (verder: checklist 1) en met een vierpuntschaal per descriptor (checklist 2). In de tweede fase werden checklist 1 en 2 ingezet voor de beoordeling van prestaties van kandidaten bij twee verschillende prompts. Bij dit deel van het onderzoek waren vijf beoordelaars betrokken. Daarna werd de beoordelaarsovereenstemming (Cohen's kappa) voor elk van de instrumenten berekend. De kappa voor checklist 1 was .75 en voor checklist 2 was deze .73. Ook werd gekeken naar de correlaties van somscores op de beide checklists met de IELTS-score op basis van het IELTS-beoordelingsmodel. De correlatie tussen checklist 1 en het IELTS-model was .73, die tussen checklist 2 en IELTS was .69 en de correlatie tussen de beoordeling met beide checklists was .80. De auteurs concluderen dat checklists geschikt zijn voor de beoordeling van dit taaktype en voordelen bieden op het punt van diagnostische feedback aan kandidaten.

Een vergelijkbaar onderzoek met een vergelijkbaar doel, het verbeteren van feedback aan kandidaten, werd uitgevoerd door Safari en Ahmadi (2023). Zij ontwikkelden een binaire checklist met 30 items bij de TOEFL geïntegreerde schrijftaak. Zij vertrokken bij de ontwikkeling van de checklist vanuit de knelpunten die kandidaten benoemden bij het uitvoeren van hun schrijftaak. Vier ervaren beoordelaars waren betrokken bij het onderzoek. Zij beoordeelden 100 producten bij twee opdrachten van de TOEFL geïntegreerd lees-luister-schrijftaak. Ook hier werd beoordelaarsovereenstemming berekend; deze was .91 voor de eerste opdracht en .89 voor de tweede opdracht. Beoordelaars vonden de schaal makkelijk te hanteren. De relatie van de checklist met de holistische TOEFL scores varieerde van .70 tot .82, waarbij beoordelaars verschilden in de mate van overeenstemming. Er werden kandidaten betrokken bij het onderzoek om na te gaan in hoeverre de beoordeling op de checklist nuttige feedback bood. De kandidaten vonden de feedback waardevol.

Tegengestelde resultaten werden gevonden in een onderzoek van Chou (2013). Het onderzoek werd uitgevoerd in Taiwan, onder studenten die Engels als vreemde taal leerden. Driehonderd studenten voerden een spreekopdracht uit, een voorbereide dialoog in tweetallen. Er waren 15 beoordelaars bij het onderzoek betrokken, die de spreekproducten beoordeelden met een analytisch beoordelingsmodel en met een checklist. Het analytische model beoordeelde de aspecten relevantie, uitspraak & vloeiendheid, grammatica, communicatiestrategieën, en pragmatische competentie, op een schaal van nul tot vier. Er werd een checklist met precies dezelfde aspecten ontwikkeld, met in totaal tien vragen. Hoewel de beoordeling met de checklist sneller ging, vonden zowel de docenten als de kandidaten het analytische beoordelingsmodel het meest informatief. Merk op dat deze checklist zeer beknopt was.

Een onderzoek uitgevoerd door East en Young (2007) vergeleek het functioneren van twee nieuwe beoordelings-modellen. Beide modellen werden onderzocht om een eenvoudig binair model te vervangen. Het binaire model was een korte checklist met ja/nee vragen. Het voldeed niet omdat het niet in staat was om de sterke en zwakke punten van leerders aan het licht te brengen en bood zo weinig basis voor feedback. De nieuwe ontwikkelde modellen in dit onderzoek worden besproken bij deelvraag 9, in paragraaf 4.10. Onderzoek van Shi et al. (2023) kwam eveneens

voort uit onvrede met en dichotome checklist. In deze studie werden de categorieën van een checklist voor een schrijftoets Engels als vreemde taal herzien en uitgebreid. Beoordelaars vonden de dichotome beslissing te weinig genuanceerd en deze werd daarom uitgebreid naar beoordeling op een driepuntschaal. De schrijfproducten van 1166 leerders van het Engels werden beoordeeld met de polytome checklist en de onderzoekers concludeerden dat de polytome checklist goed functioneerde en een betere basis was voor feedback aan kandidaten.

Ook Deygers et al. (2011) beschreven de ontwikkeling van een nieuwe, polytome beoordelingsschaal voor het Certificaat Nederlands als Vreemde Taal (CNaVT) academisch Nederlands, een toets met beoogd niveau B2. De nieuwe polytome schaal moest een eerdere dichotome schaal vervangen. Met behulp van expertpanels uit de academische wereld en met behulp van het ERK werden criteria bepaald die op vier niveaus, van onvoldoende tot uitstekend, beoordeeld moesten worden. Vier onervaren, maar getrainde beoordelaars beoordeelden elk een selectie van 125 schrijf- en spreekproducten, met zowel de oude dichotome als de nieuwe polytome schalen. Ook namen ze deel aan een panelgesprek. De data werden kwalitatief en kwantitatief geanalyseerd. Uit de resultaten bleek dat de dichotome schaal tot een betere beoordelaarsovereenstemming leidde. Beoordelaars hadden ook een voorkeur voor de dichotome schaal omdat de beoordeling hen minder tijd kostte en ze minder vaak twijfelden tussen schaalpunten. Echter, na aanpassing en nadere precisering van de polytome schaal werd het onderzoek herhaald met een nieuwe groep onervaren beoordelaars. Deze beoordelaars spraken een voorkeur uit voor de polytome schaal.

4.8.2 Beslisbomen

- Beslisbomen zijn dichotome beoordelingsmodellen met een specifieke structuur.
- Ze zijn gebaseerd op data van leerders (en dus niet op een niveaubeschrijving).
- Beslisbomen zijn taak-specifiek.

Zoals eerder al aangegeven in paragraaf 4.6 kunnen beoordelingsmodellen ingedeeld worden naar wijze van ontwikkeling en op grond daarvan getypeerd worden als *measurement driven* of *data driven* (Fulcher et al., 2011; Kuiken & Vedder, 2021). Een *measurement driven* benadering ordent descriptoren op een schaal, gebaseerd op expertoordelen. Een *data-driven* benadering vertrekt vanuit het bestuderen van de taaluitingen van leerders. Een specifieke *data-driven* manier van dichotoom beoordelen maakt gebruik van **beslisbomen**, ook aangeduid als **EBB-schalen**. EBB staat voor *Empirically derived, Binary-choice, Boundary-definition scale*. Dit type beoordelingsmodel heeft tot doel een aantal problemen met analytische schalen te vermijden. Het gaat om problemen zoals centrale tendens (de neiging van beoordelaars om de middelste schaalpunten te kiezen), beperkt gebruik van de range van schaalpunten en halo-effecten. De EBB-schaal zou moeten leiden tot hoge intra- en interbeoordelaarsbetrouwbaarheid.

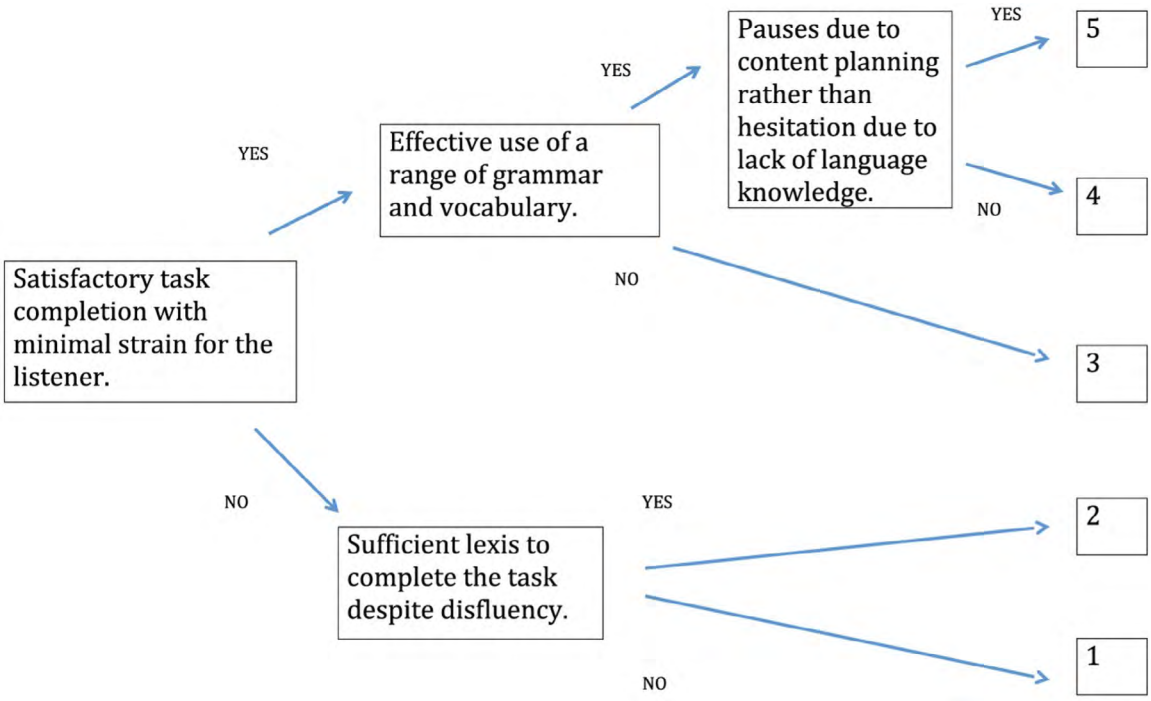
Het gebruik van beslisbomen werd geïntroduceerd door Turner en Upshur (1995) in een onderzoek in Canada onder jonge leerders van het Engels, met Frans als moedertaal. De leerlingen legden twee spreektaken af en deze werden elk beoordeeld met twee beslisbomen, één voor communicatieve effectiviteit en één voor grammaticale correctheid. De beslisbomen waren ontwikkeld op basis van de spreekproducten van de leerlingen en waren taak-specifiek. In totaal werden er dus vier beslisbomen ontwikkeld. Het doel van het onderzoek was om de relatie te onderzoeken tussen communicatieve effectiviteit en grammaticale correctheid. De beslisbomen waren effectief om aan te tonen dat deze relatie taakafhankelijk was: bij het ene taaktype leken communicatieve effectiviteit en grammaticale correctheid zich gelijktijdig te ontwikkelen, terwijl bij het andere taaktype bleek dat grammaticale correctheid zich niet gelijktijdig met communicatieve effectiviteit ontwikkelde, maar dat de scores juist uiteen konden lopen.

Fulcher et al. (2011) uitten kritiek op *measurement-driven* beoordeling. Ze zijn van mening dat daarin de meetbaarheid leidend wordt, in plaats van het taalconstruct, het doel van de beoordeling en de communicatieve context. Als demonstratie van een alternatieve benadering beschreven ze de ontwikkeling van een beslisboom voor het beoordelen van een baliegesprek. Op basis van analyse van de kenmerken van dergelijke gesprekken kwamen de auteurs tot een complex model dat het construct 'baliegesprekken' goed representeert, waardoor generaliseerbaarheid naar andere baliegesprekken aannemelijk is. Het model omvat talige en pragmatische aspecten. De beoordelaar neemt een serie binaire beslissingen die gekoppeld zijn aan de toe te kennen punten. Fulcher en collega's betogen dat het voorgesteld model meer valide is.

Beslisbomen focussen op talige en pragmatische aspecten en zijn in principe context-gebonden (O'Grady, 2019, 2025). De ontwikkeling gaat als volgt: uit een set kandidaatprestaties wordt een aantal producten verzameld dat intuïtief een range vertegenwoordigt van de diverse vaardigheidsniveaus. In overleg tussen experts worden deze

producten geordend van zwak naar sterk. De prestaties worden geanalyseerd om de kenmerken van de verschillende groepen van prestaties naar vaardigheid te typeren. De kenmerkende eigenschappen worden gebruikt voor het ontwikkelen van een beslisboom met ja/nee-vragen. Een voorbeeld van een eenvoudige beslisboom voor een spreekopdracht, afkomstig uit O'Grady (2019), is weergegeven in Figuur 1. Deze beslisboom deelt prestaties in naar vijf niveaus.

Figuur 1 *Voorbeeld van een eenvoudige beslisboom*



Hirai & Koizumi (2008) vergeleken een analytische en een (taak-specifieke) EBB-schaal in een studie waarbij 52 kandidaten en zes beoordelaars waren betrokken. De vragen die het onderzoek beoogde te beantwoorden waren: (1) Hoeveel taken zijn nodig om een betrouwbare score te kunnen geven met elk van de twee schalen?, (2) Hoe is de relatie tussen de twee schalen, en (3) Welke schaal is het meest praktisch? De kandidaten waren studenten Engels en engineering aan universiteiten in Japan. Ze voerden een naverteltaak uit op basis van korte leesteksten. De analytische schaal die in dit onderzoek werd gebruikt keek naar communicatieve efficiëntie, grammatica, vloeiendheid en uitspraak. De EBB-schaal werd voor het onderzoek geconstrueerd en onderscheidde de aspecten communicatieve efficiëntie, grammatica & woordenschat, inhoud en uitspraak, met een afzonderlijke beslisboom per beoordelingsaspect. Data werden geanalyseerd met behulp van generaliseerbaarheidstheorie. Uit de resultaten bleek dat met gebruik van beide modellen ten minste twee taken nodig waren om te komen tot een score met voldoende generaliseerbaarheid. De correlaties tussen beide schalen waren hoog maar de beoordeling met de EBB-schaal leidde tot een iets betere generaliseerbaarheid (betrouwbaarheid). De beoordelaars vonden de EBB-schaal echter niet makkelijk om te hanteren en de beoordeling met het analytisch model ging sneller.

In een onderzoek gericht op de impact van voorbereidingstijd op prestaties op taken in spreekexamens keek O'Grady (2019) onder andere naar de invloed van het beoordelingsmodel. Hij vergeleek daarbij een analytisch model en een EBB-model om spreekprestaties te beoordelen die tot stand waren gekomen met een uiteenlopende hoeveelheid voorbereidingstijd. De prestaties met een voorbereidingstijd van vijf minuten scoorden het hoogste, en die uitkomst gold voor beide beoordelingsmodellen, die in de context van dit onderzoek dus gelijk presteerden.

Het gebruik van een beslisboom is makkelijk, maar de ontwikkeling is tijdrovend. Onderzoek van Yokouchi (2025) had tot doel na te gaan of de inspanningen de moeite waard waren in termen van validiteit, betrouwbaarheid en toepasbaarheid. De data in dit onderzoek zijn spreekproducten van kandidaten van een toets Engels, genaamd VELC (Visualising English Language Competence), afgelegd door studenten in Japan. Er lagen twee taken ten grondslag aan de spreekproducten en in beide taken moesten kandidaten een mondelinge samenvatting geven op basis van

een korte leestekst. Twee beoordelingsmodellen werden ontwikkeld, een beslisboom met vijf ja/nee-vragen en een holistisch model dat vijf niveaus onderscheidde. Drie beoordelaars beoordeelden de spreekproducten met de beide modellen. De beoordelingsovereenstemming bleek voor beide modellen goed, boven de .90. De beoordelaars vonden de beslisboom makkelijker om te gebruiken. De beslisboom zou daarom de betere optie zijn, en de ontwikkeltijd waard.

4.8.3 Onderzoek naar schaallengte

- De kwaliteit van een beoordelingsmodel kan worden onderzocht met Rasch-modellen.
- Dichotomisering 'achteraf' kan leiden tot betere modelfit.

Bij valide en betrouwbare beoordeling speelt de kwaliteit van het beoordelingsmodel een rol die vaak over het hoofd wordt gezien. Met behulp van Rasch-modellen evalueerde Weaver (2006) drie verschillende beoordelingsmodellen die werden gebruikt bij schrijftaken afkomstig van een toets Engels voor toelating tot Japanse universiteiten. De schrijfproducten waren afkomstig van 1108 Japanse studenten, die drie verschillende taken maakten, alle gerelateerd aan het beschrijven van een figuur. In de eerste taak moesten de kandidaten korte vragen beantwoorden bij de figuur. De antwoorden werden beoordeeld met een driepuntschaal. De tweede taak bestond uit het afmaken van een dialoog over de figuur; deze taak werd beoordeeld met een vierpuntschaal. In de derde taak moesten de kandidaten zinnen schrijven, eveneens bij de figuur, en de taak werd beoordeeld met een vijfpuntschaal. Het onderzoek was erop gericht te bepalen of de modellen op de best mogelijke manier onderscheid konden maken tussen kandidaten van verschillende vaardigheidsniveaus, en dus goed voldoen aan de verwachtingen van het Rasch-model. Een beoordelingsmodel functioneert goed als alle schaalpunten worden gebruikt, en de oplopende schaalpunten samenhangen met oplopende vaardigheid van kandidaten. Als een beoordelingsmodel niet aan deze verwachtingen voldoet dan zijn mogelijke maatregelen het verwijderen van een categorie in de schaal of het samenvoegen van categorieën. Uit de analyses bleek dat dichotomisering van de 3- en 4 puntschalen, waarbij dus categorieën werden samengevoegd, de beste vaardigheidsschatting van kandidaten opleverde. Dat gold niet voor de vijfpuntschaal, al voldeed deze schaal ook niet helemaal aan het Raschmodel. Omdat er sprake bleek van plafondeffecten (veel kandidaten haalden de hoogste score) adviseerde de auteur om de schrijftaak uitdagender te maken.

4.8.4 Kritische reflectie

Dichotome beoordelingsmodellen worden niet op grote schaal gebruikt in highstakes toetsing. Er werd niet veel onderzoek gevonden dat gebruik maakte binaire oordelen en het meeste van dit type onderzoek kwam niet uit de context van highstakes toetsing. Het gevonden onderzoek dat afkomstig was uit de context van highstakes toetsing ging over de ontwikkeling van een alternatief voor een dichotoom model (bijvoorbeeld Deygers et al., 2011). De twee gevonden methodes van beoordeling die wel gebruik maken van binaire categorieën zijn beslisbomen en checklists. Hoewel voorbeelden van het gebruik van deze technieken in de context van taalvaardigheidstoetsen schaars zijn, valt op dat de uitwerking verschilt in mate van detaillering. Met name meer uitgebreide checklists lijken voordelen te hebben voor kandidaten in termen van de feedback op hun prestaties. Als nadeel geldt dat een hoge constructielast wordt gerapporteerd.

4.9 Holistisch beoordelen

Dit hoofdstuk bespreekt de vraag: Wat is er bekend over holistisch beoordelen? Zijn er andere examens die voor het beoordelen van Schrijven en Spreken werken met (gedeeltelijk) holistische beoordelingsmodellen? En hoe zitten die beoordelingsmodellen in elkaar? De tekst bespreekt eerst de kenmerken van holistisch beoordelen en de voor-en nadelen ten opzichte van analytische beoordeling, gevolgd door onderzoek dat de twee modellen vergelijkt. Daarna komt de rol van de beoordelaar bij holistische beoordeling aan de orde, gevolgd door onderzoek in de specifieke context van grootschalige *highstakes* taaltoetsen. De vraag 'En hoe zitten die beoordelingsmodellen in elkaar?' wordt besproken aan de hand van reviewartikelen over TOEFL (Alderson, 2009; Cumming et al., 2021; Xi et al., 2021) en IELTS (Cumming et al., 2021; Read, 2022; Xi et al., 2021). De literatuur voor deze deelvraag is gevonden met de zoekterm: "holistic rating" AND "language testing". Een overzicht van de gebruikte teksten is gegeven in Tabel 13.

Tabel 13 *Bestudeerde literatuur deelvraag 8*

1. Alaei, M. M., Ahmadi, M., & Zadeh, N. S. (2014). The impact of rater's personality traits on holistic and analytic scores: Does genre make any difference too? *Procedia-Social and Behavioral Sciences, 98*, 1240-1248.
2. Alderson, J. C. (2009). Test review: Test of English as a foreign language™: Internet-based test (TOEFL iBT®). *Language Testing, 26*(4), 621-631.
3. Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing, 27*(4), 515-535.
4. Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2011). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing, 29*(1), 91-108.
5. Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics, 11*(2).
6. Chuang, Y. Y. (2009). Foreign language speaking assessment: Taiwanese college English teachers' scoring performance in the holistic and analytic rating methods. *Asian EFL Journal, 11*(1), 150-173.
7. Chun, L., & Yunnan, X. (2018). A study on rater reliability under holistic and analytic scoring of CEPT Writing by using Generalizability Theory and Many-facet Rasch Model. In *4th International Conference on Economics, Management, Law and Education (EMLE 2018)* (pp. 1013-1019). Atlantis Press.
8. Cumming, A. Cho, Y., Burstein, J. Everson, P. & Kantor, R. (2021) Assessing academic writing. In: Xi, X., & Norris, J. M. (Eds.). (2021). *Assessing academic English for higher education admissions.* Routledge.
9. Galti, A. M., Saidu, S., Yusuf, H., & Goni, A. A. (2018). Rating scale in writing assessment: holistic vs. analytical scales: A review. *International Journal of English Research, 4*(6), 4-6.
10. Huang, J., & Han, T. (2013). Holistic or analytic? An EFL institutional writing assessment dilemma for policy makers. *Leadership Policy Quarterly, 2*(1), 1-18.
11. Izadpanah, M. A., Rakhshandehroo, F., Hoomanfard, H. M., & Mahmoudikia, M. (2014). On the consensus between holistic rating system and analytic rating system: A comparison between TOEFL IBT and Jacobs et. al. composition profile. *International Journal of Language Learning and Applied Linguistic World, 6*(1), 170-187.
12. Li, S., Wen, T., Li, X., Feng, Y., & Lin, C. (2023). Comparing holistic and analytic marking methods in assessing speech act production in L2 Chinese. *Language Testing, 40*(2), 249-275.
13. Ma, W. (2022). What the analytic versus holistic scoring of international teaching assistants can reveal: Lexical grammar matters. *Language Testing, 39*(2), 239-264.
14. Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2014). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing, 32*(1), 39-62.
15. Ono, M., Yamanishi, H., & Hijikata, Y. (2019). Holistic and analytic assessments of the TOEFL iBT® Integrated Writing Task. *JLTA Journal, 22*, 65-88.
16. Pallotti, G. (2022). Holistic and analytic assessment of functional adequacy. *TASK, 2*(1), 85-114.
17. Park, J. (2006). The study on the rater reliability of three scoring methods in assessing argumentative essays: Holistic, analytic, and multipletrait scoring methods. *Foreign Language Education Research, 9*, 63-84.
18. Park, H., & Yan, X. (2019). An investigation into rater performance with a holistic scale and a binary, analytic scale on an ESL writing placement test. *Papers in Language Testing and Assessment, 8*(2), 34-64.
19. Read, J. (2022). Test review: The international English language testing system (IELTS). *Language testing, 39*(4), 679-694.
20. Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In: Kunnan, A.J. *Fairness and validation in language assessment*, 129-152.
21. Sakyi, A. A. (2003). *A study of the holistic scoring behaviours of experienced and novice ESL instructors.* University of Toronto.
22. Wiseman, C. S. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *International Journal of Language Testing, 2*(1), 59-92.
23. Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing, 24*(2), 251-286.
24. Xi, X., Norris, J.M., Ockey, G.J., Fulcher, G. & Purpura, E. (2021). Assessing Academic Speaking. In: Xi, X., & Norris, J. M. (Eds.). *Assessing academic English for higher education admissions.* Routledge.

4.9.1 Analytische en holistische modellen

- De keuze voor een holistisch of analytisch model hangt samen met de opvatting van het te meten construct.
- Beide modellen hebben voor- en nadelen.

Zowel analytische als holistische beoordelingsmodellen worden veel gebruikt in de praktijk van highstakes taaltoetsing. Bij holistische beoordeling krijgt de leerder per taak een score voor de hele prestatie, waarbij het voornaamste criterium is of ideeën goed zijn uitgewerkt en of de boodschap overkomt. Bij een analytische beoordeling worden verschillende aspecten afzonderlijk gescoord, zoals bijvoorbeeld grammatica, woordenschat en coherentie (Galti et al., 2018; Ma, 2022). De keuze voor het ene of het andere type model hangt samen met de conceptualisering van de vaardigheid: hoewel in holistische beoordeling verschillende aspecten van een construct betrokken kunnen worden (zoals bijvoorbeeld woordgebruik of grammatica), veronderstelt een holistisch model een gelijkmatige ontwikkeling van die aspecten. Een analytisch model houdt er rekening mee dat het te meten construct meerdimensionaal is en dat ontwikkeling van verschillende aspecten van de vaardigheid (wederom bijvoorbeeld grammatica of woordgebruik) ongelijk kan verlopen (S. Li et al., 2023; J. Park, 2006). In termen van toepasbaarheid verschillen toetsontwikkelaars en onderzoekers van mening over welk type model het beste is, maar in het algemeen wordt aangenomen dat holistische modellen kunnen worden toegepast om te komen tot een algemeen oordeel over het taalniveau van leerders, terwijl analytische modellen gebruikt worden om de verschillende aspecten van taalvaardigheid in kaart te kunnen brengen (Ma, 2022).

De literatuur noemt verschillende nadelen van holistisch beoordelen: ten eerste is het niet duidelijk welke criteria beoordelaars impliciet betrekken in hun holistisch oordeel en is het mogelijk dat verschillende beoordelaars verschillende criteria hanteren (Barkaoui, 2010; Ma, 2022). Ten tweede biedt een holistisch oordeel weinig informatie voor de kandidaat en voor het onderwijs. De training kan dus -op basis van de score- niet worden aangepast op de behoeften van de leerder. Ten derde ontwikkelen leerders zich op deelaspecten van spreek- of schrijfvaardigheid niet allemaal op dezelfde manier. Met name leerders met ongelijke profielen, dus bepaalde sterktes en zwaktes in hun taalontwikkeling, zijn moeilijker te beoordelen met een holistisch model (Ma, 2022). Als voordeel van holistisch beoordelen wordt genoemd dat het de aandacht richt op wat de leerder kan, en minder op wat die niet kan. In die zin kan een holistisch model als een positief beoordelingsmodel worden gezien (Wiseman, 2012). Daarnaast gaat holistische beoordeling meestal sneller en is het dus in economisch opzicht aantrekkelijker (Wiseman, 2012; Xi, 2007).

Analytische modellen beoordelen zoals gezegd verschillende aspecten van de vaardigheid. Dit biedt de leerder -in principe- diagnostische informatie. Als nadeel van analytische modellen geldt dat ze belastend zijn voor beoordelaars die op verschillende aspecten tegelijk moeten letten. Deze hoge 'cognitieve last' kan ook inconsistente beoordeling veroorzaken. Daarnaast heeft veel onderzoek hoge correlaties tussen de verschillende deelaspecten van analytische beoordelingen gevonden, wat erop wijst dat de analytische beoordeling in psychometrische zin redundant is: als de score op één aspect in hoge mate voorspellend is voor de score op een ander aspect, is het niet nodig beide aspecten afzonderlijk te beoordelen. Tegelijk zijn er ook voordelen van analytische modellen: ze bieden een meer accuraat beeld van de vaardigheid van leerders, passend bij de opvattingen van communicatieve taalvaardigheid als meerdimensionaal construct, en ze bieden de basis voor feedback aan taalleerders (Ma, 2022).

4.9.2 Onderzoek dat holistische en analytische modellen vergelijkt

- Onderzoek dat het functioneren van holistische en analytische modellen vergelijkt, vindt meestal een voordeel voor het analytisch model: de beoordelaarsovereenstemming is hoger en de betrouwbaarheid groter.
- Scores toegekend met een analytisch model lijken ook vaak wat hoger uit te pakken.

Er is nogal wat onderzoek gedaan waarbij de twee type beoordelingsmodellen en de uitkomsten die zij genereren worden vergeleken. Deze paragraaf bespreekt een aantal van die onderzoeken. Chuang (2009) vergeleek het gebruik van twee beoordelingsmodellen voor het beoordelen van spreektaken. Het ene model was een globale schaal die zeven oplopende niveaus van spreekvaardigheid holistisch beschreef. Het werd vergeleken met een analytisch model dat de aspecten grammaticale accuraatheid, woordkeus, uitspraak, vloeiendheid en begrijpelijkheid beoordeelde op een zevenpuntschaal. Er werden vijf spreekproducten gebruikt, afkomstig van studenten Engels aan een universiteit in Taiwan. 62 docenten Engels die lesgaven aan universiteiten in Taiwan, beoordeelden dezelfde spreekproducten twee keer, met beide modellen. Daarnaast vulden ze een vragenlijst in. Uit de resultaten bleek dat de gemiddelde scores van kandidaten beoordeeld met de analytische schaal iets hoger waren dan de scores behaald met de holistische schaal, maar het verschil was niet significant. Ook kwamen de beoordelingen overeen wat betreft de

gemiddelde score per taak: de taken met de hoogste en laagste gemiddelde score waren dezelfde onder beide beoordelingscondities. De prestaties van beide modellen waren dus vergelijkbaar.

Onderzoek van Park (2006) vergeleek drie verschillende beoordelingsmodellen. Naast holistische en analytische beoordeling betrok hij ook *multiple-trait scoring* bij het onderzoek. Deze benadering zal ik bespreken in paragraaf 4.10, ik beperk me hier tot het beschrijven van Parks bevindingen met betrekking tot analytische en holistische beoordelingsmodellen. Het doel van het onderzoek was om inzicht te krijgen in het functioneren van de verschillende beoordelingsmodellen. Het holistische beoordelingsmodel was gebaseerd op TOEFL en voor het analytische model werd een bestaand model gebruikt, dat de aspecten inhoud, taalgebruik, organisatie, woordenschat, en spelling & interpunctie beoordeelde met behulp van een vierpuntschaal. Er waren vijf beoordelaars betrokken bij het onderzoek. Ze beoordeelden zestien schrijfproducten, afkomstig van Koreaanse leerders van het Engels. Uit de resultaten bleek dat de gemiddelde scores op basis van het holistische model beduidend lager waren dan die verkregen met het analytische model. Ook de beoordelaarsovereenstemming was met het analytische model duidelijk beter.

Ook J. Huang en Han (2013) vergeleken twee typen beoordelingsmodellen, met de vraag of deze tot dezelfde beoordeling en beoordelingskwaliteit zouden leiden. Ze vergeleken een bestaand holistisch model met een voor het doel ontwikkeld analytisch model. Er werden negen schrijfproducten van drie niveaus (hoog, medium en laag) beoordeeld door vijf ervaren, getrainde beoordelaars. De producten werden eerst holistisch beoordeeld en werden twee weken later herbeoordeeld met een analytische schaal. De data werden geanalyseerd met generaliseerbaarheidstheorie (G-theorie). Uit de resultaten bleek dat de scores met behulp van het analytische model duidelijk hoger waren. De generaliseerbaarheid (betrouwbaarheid) van de beoordeling was beter bij de analytische dan bij de holistische beoordeling (met een G-coëfficiënt van .90 versus .64). De conclusie was daarom dat het analytische model beter functioneerde.

Ma (2022) vergeleek de beoordeling met behulp van een bestaand holistisch model met een analytische beoordeling. Het onderzoek had tot doel te onderzoeken of analytische en holistische schalen leerders op dezelfde manier onderscheiden naar spreekvaardigheid, en welke sub-schaal in het analytische model de grootste bijdrage levert aan het bepalen van de vaardigheid. De spreekproducten waren afkomstig van een spreektoets voor klasse-assistenten met Engels als tweede taal in de VS. De toets bestond uit twaalf spreektaken, en er waren 127 kandidaten van verschillende niveaus betrokken bij het onderzoek. De holistische score was al beschikbaar, omdat dat de gangbare manier was waarop de toets beoordeeld werd. Alle spreekproducten werden herbeoordeeld door tien getrainde beoordelaars met behulp van een voor het onderzoek ontwikkeld analytisch beoordelingsmodel. Het model onderscheidde vier aspecten (1) fonetische & fonologische competentie, (2) lexicale & grammaticale competentie, (3) retorische organisatie en (4) ontwikkeling van het onderwerp. Uit de analyses bleek dat beide beoordelingsmodellen kandidaten even goed indeelden naar vaardigheid. Het aspect lexicale en grammaticale competentie leverde de grootste bijdrage aan het bepalen van de spreekvaardigheid: dit aspect had de hoogste correlatie met de holistische scores, en de verschillen in scores op grammatica van kandidaten net onder en net boven de cesuur waren significant voor dit aspect.

Ook S. Li et al. (2023) vergeleken analytische en holistische beoordeling in een spreekassessment voor Amerikaanse leerders van het Chinees. De data waren afkomstig van 70 participanten die een serie van acht voor het doel ontwikkelde spreektaken aflegden via de computer. De totale dataset bestond uit 560 responses die beoordeeld werden door vier getrainde beoordelaars. De onderzoekers ontwikkelden een holistische en een analytische schaal die dezelfde aspecten in kaart brachten. Het verschil was dat de kandidaat in de ene conditie één totaalscore kreeg en in de andere conditie afzonderlijke scores per taak. De beoordelingscriteria waren communicatieve functie, vloeiendheid, intonatie, accuraatheid, en register. In beide beoordelingsmodellen kon er gescoord worden van 1 punt (zwak) tot 4 punten (uitstekend). Data werden geanalyseerd met een Rasch-model. Uit de resultaten bleek dat de scores tussen de holistische en analytische beoordeling sterk correleerden en dat beide beoordelingen betrouwbaar waren. De analytische beoordeling was de meest betrouwbare en op basis van het analytische model vertoonden de Rasch-analyses betere modelfit voor kandidaten en items. De onderzoekers concluderen daarom dat de analytische beoordeling de voorkeur geniet in highstakes toetsen.

Een onderzoek van H. Park en Yan (2019) werd eerder besproken in paragraaf 4.8, bij de vraag naar dichotome beoordeling. Het onderzoek is ook relevant met betrekking tot holistisch beoordelen. Het vergeleek de betrouwbaarheid van een holistisch beoordelingsmodel met een dichotoom analytisch model. De dataset bestond uit 452 essays afkomstig van de English Placement Test, afgenomen aan een Amerikaanse universiteit. Negen getrainde

beoordelaars waren betrokken bij het onderzoek. Het onderzoek verliep in twee fases. In de eerste fase beoordeelden de beoordelaars met behulp van het holistische beoordelingsmodel. Het model deelde kandidaten in op vier niveaus, A, B, C en D, waarbij A het hoogste niveau is. De globale criteria zijn gericht op kwaliteit van de argumentatie en op lexicale en grammaticale aspecten ('lexico-grammar'). Naast de beoordeling namen de beoordelaars deel aan een interview en een hardop-denksessie. Uit de resultaten bleek dat de beoordelaarsovereenstemming zeer laag was. De beoordelaars bleken ook na training hun eigen uiteenlopende criteria toe te voegen aan het model. Ook focusten ze eenzijdig op de criteria voor argumentatie, omdat ze de criteria voor lexicale en grammaticale aspecten moeilijk toepasbaar vonden. Op basis van deze informatie werd een analytisch model ontwikkeld. Het analytische model splitste de elementen argumentatie en lexicale en grammaticale elementen uit in verschillende deelvragen die met ja of nee beantwoord konden worden, bijvoorbeeld: 'Is het essay makkelijk te begrijpen?'; 'Zijn er grote grammaticale fouten?' en 'Is het woordgebruik meestal passend?'. De beoordelaars gebruikten deze schaal met een nieuwe set schrijfproducten en uit de resultaten bleek dat de overeenstemming sterk verbeterd was. De overeenstemming was beter voor argumentatie dan voor lexicaal-grammaticale aspecten. De beoordelaars gaven aan dat ze de analytische schaal makkelijker vonden om te gebruiken.

In de context van een toelatingsexamen Engels als vreemde taal voor studenten van een Iraanse universiteit vergeleek ook Wiseman (2012) de beoordeling met een holistisch en een analytisch beoordelingsmodel. Ze onderzocht drie onderzoeksvragen: (1) Functioneren het holistisch en het analytisch model adequaat als zespuntschaal, dus zijn de categorieën duidelijk te onderscheiden? (2) Zijn de categorieën adequaat geordend? En (3) Welk model kan kandidaten het best indelen naar vaardigheid? Dit onderzoek maakte gebruik van schrijfproducten van 60 kandidaten. De toets werd normaalgesproken beoordeeld met een holistische schaal, die leerders indeelde op zes niveaus. Het analytische model werd voor dit onderzoek ontwikkeld; het beoordeelde de aspecten taakuitvoering, ontwikkeling van het onderwerp, organisatie, register en woordenschat, en grammaticale accuraatheid. Elk aspect werd beoordeeld op een zespuntschaal waarbij alle schaalpunten waren omschreven. De beoordelaars waren vijf ervaren en getrainde docenten Engels. Rasch-analyses werden gebruikt voor de verwerking van de data. Uit de analyses bleek allereerst dat de drie analytische criteria overeenkwamen in moeilijkheid, en de zes schaalpunten bleken adequaat geordend. Dit bevestigt de kwaliteit van de analytische schaal. Analyse van de holistische schaal maakte duidelijk dat het laagste schaalpunt niet werd gebruikt. Dat wijst erop dat de schaal beter zou functioneren als een vijfpuntschaal. Beide schalen konden kandidaten indelen naar vaardigheid maar de analytische schaal deed dat nauwkeuriger en discrimineerde dus beter.

Carr (2000) voerde analyses uit op data van de universitaire toelatingstoets UCLA ESLPE. Deze toets, ontwikkeld en afgenomen aan de University of California, Los Angeles (UCLA) in de VS, bestaat uit de onderdelen lezen, luisteren en schrijven. De schrijftoets werd aanvankelijk beoordeeld met een analytisch model, dat de aspecten inhoud, organisatie, en taalgebruik onderscheidde. Beoordelaars vonden het onderscheid tussen inhoud en organisatie lastig te maken, en dat was de reden om het model om te werken tot een holistisch model dat uitsluitend naar taalgebruik keek. De globale schaal onderscheidde zes niveaus. Carrs analyse betreft het functioneren van de analytische schaal en de holistische schaal. Daarbij keek hij naar de mate waarin de leestoets, de luistertoets en de holistische of analytische schrijfscores voorspellend waren voor de eindscore. Uit de analyses bleek dat de score voor schrijven zoals toegekend met het holistische model de beste voorspeller was voor de totaalscore, gevolgd door lezen en luisteren.

In China moeten studenten aan de universiteit een toelatingstoets Engels afleggen, de College English Placement Test. Chun en Yunnan (2018) onderzochten de betrouwbaarheid van de beoordeling van het onderdeel schrijfvaardigheid. Ze keken naar twee factoren die de score van kandidaten kunnen beïnvloeden, namelijk de beoordelaar en de beoordelingsschaal, en ze gebruikten zowel **G-theorie** als **many-faceted Raschmodellen** voor de analyse. Bij het onderzoek waren vijftien getrainde beoordelaars betrokken die 300 schrijfproducten beoordeelden met een holistisch en een analytisch beoordelingsmodel. De holistische schaal onderscheidde zes niveaus, het analytische model was gebaseerd op de Canadian Benchmarks en onderscheidde de aspecten taak (inhoud), grammatica, spelling en interpunctie, woordgebruik en structuur, alle beoordeeld met een zespuntschaal. Bij elk van de twee beoordelingsmodellen was één beoordelaar toereikend om een betrouwbaarheid boven .80 te bereiken. Beoordelaars verschilden in streng- en mildheid met beide beoordelingsmodellen, maar beoordeelden in beide condities overwegend consistent. De auteurs adviseren beoordelaars individueel te trainen of feedback te geven.

4.9.3 TOEFL iBT

- TOEFL iBT is één van de bekendste grootschalige high stakes toetsen. De onderdelen spreken en schrijven worden beoordeeld met een holistisch model en daarnaast automatisch gescoord.
- Onderzoek wijst uit dat de beoordeling van TOEFL voldoende valide en betrouwbaar is.

Een van de bekendste grootschalige highstakes spreektoetsen Engels als tweede taal is TOEFL iBT. Het onderdeel spreekvaardigheid wordt gemeten met een spreektoets die (soms) TAST wordt genoemd (*TOEFL Academic Speaking Test*). Voor de beoordeling van TAST wordt een holistisch beoordelingsmodel gebruikt, dat kijkt naar overall performance door het combineren van drie criteria: (1) presentatie (uitspraak en vloeiendheid), (2) taalgebruik (grammatica en woordenschat) en (3) ontwikkeling van het onderwerp (coherentie en volledigheid). Een link naar het holistische beoordelingsmodel vind je hier: https://www.ets.org/pdfs/toefl/TOEFL iBT-speaking-rubrics.pdf. Naast de menselijke beoordeling worden de prestaties automatisch gescoord met behulp van software, de zogenaamde SpeechRater™ (Bridgeman et al., 2011).

Vanwege de nadelen van holistische beoordeling onderzocht Xi (2007) de kwaliteit van een alternatief, analytisch model voor TAST. Het onderzoek maakte gebruik van generaliseerbaarheidstheorie om de betrouwbaarheid van het analytische model en het onderscheid tussen de schaalpunten te berekenen. Daarnaast wilde men weten in welke mate ongelijke cursistprofielen voorkwamen. Prestaties van kandidaten met ongelijke profielen zouden immers slechter te beoordelen zijn met een holistisch model. Het analytische model omvatte de criteria taakuitvoering, taalgebruik en ontwikkeling van het onderwerp elk afzonderlijk op een vierpuntschaal. Dit maakte het analytische model goed vergelijkbaar met het holistische model. De data bestond uit de volledige spreekexamens van 140 kandidaten. Het examen bestond uit acht taken met elk 45 tot 90 seconden spreektijd. Er waren veertien beoordelaars bij het onderzoek betrokken. Zij beoordeelden de prestaties eerst met de holistische en daarna met de nieuwe analytische schaal. Bij de holistische beoordeling wordt per taak een score gegeven. Het gemiddelde van de scores is de gerapporteerde score. Daarna werden beoordelaars getraind in het gebruik van het analytische model en er werd opnieuw beoordeeld. Ook vulden de beoordelaars een vragenlijst in, gericht op de vraag hoe goed de verschillende beoordelingsaspecten te onderscheiden waren. Uit het onderzoek bleek allereerst dat de beoordelaarsovereenstemming tussen de drie onderscheiden beoordelingsaspecten gelijk was. Beoordelaars vonden de analytische schaal redelijk makkelijk te gebruiken maar sommigen vonden 'taalgebruik' moeilijker te beoordelen en vonden dat het overlap vertoonde met de andere aspecten. Ook vonden sommige beoordelaars 'ontwikkeling van het onderwerp' lastig om te beoordelen. Analyses van de betrouwbaarheid lieten zien dat de beoordelaarsovereenstemming overwegend goed was. De betrouwbaarheid van de toets werd hoger door het toevoegen van meer taken (met een sterke toename tussen één en vier taken en kleine verdere toename van vier naar zes taken). Het toevoegen van meer beoordelaars (van één naar twee) leverde een kleine bijdrage aan de betrouwbaarheid. De vraag of de analytische scores voordelen bieden boven de holistische scores werd onderzocht door na te gaan welke aanvullende informatie verkregen werd met de analytische score. De scores op de drie deelaspecten bleken sterk samen te hangen. Dat betekent dat kandidaten over het algemeen geen 'ongelijk profiel' hadden en dus was er in dat opzicht weinig meerwaarde van de analytische score. Ook lijkt de score op basis van het analytische model lastiger om te rapporteren, omdat dat zou betekenen dat stakeholders het relatief belang van elk van de componenten moeten afwegen[10]. Xi (2007) concludeerde dat er niet veel reden is om van het holistisch model over te stappen op een analytisch beoordelingsmodel.

Bridgeman et al. (2011) voerden valideringsonderzoek uit naar de beoordeling van de TOEFL iBT spreektoets door de scores op basis van het holistisch model en Speechrater™ te vergelijken met een maat voor communicatieve adequaatheid. Ze gebruikten hiervoor naïeve beoordelaars, studenten die niet getraind waren als beoordelaar van spreekvaardigheid. Eerder onderzoek naar deskundige en naïeve beoordelaars richtte zich meestal op verschillen in beoordeling, maar niet op daadwerkelijk bewijs van communicatieve effectiviteit. Daarnaast was het belangrijke om de destijds nieuwe automatische scores door Speechrater™ te onderzoeken op de relatie met communicatieve effectiviteit. De opzet van het onderzoek was als volgt: de spreekproducten van 184 kandidaten werden voorgelegd aan in totaal 555 studenten, die geen ervaren TOEFL-beoordelaars waren. De studenten beoordeelden elk de communicatieve adequaatheid van twaalf spreekproducten, door meerkeuzevragen te beantwoorden over inhoud van de producten. Deze vragen konden ze alleen maar beantwoorden als ze de uiting hadden begrepen. De scores op

10 TOEFL iBT rapporteert een niveau, niet een zakslaagbeslissing. Stakeholders bepalen welk niveau zij vereisen voor hun doel, bijvoorbeeld toelating tot een opleiding. De stakeholders moeten dus meer dan in de staatsexamencontext in staat zijn om de uitslag van de kandidaat te interpreteren.

basis van de beoordeling door studenten werden vergeleken met de scores door de getrainde TOEFL-beoordelaars en door Speechrater™. De scores voor communicatieve adequaatheid bleken sterk samen te hangen met de scores toegekend door de beoordelaars, die een holistisch model gebruikten. De scores hingen ook samen met de analytische maten die Speechrater™ in kaart brengt, maar die relatie was minder sterk. Het onderzoek leverde een onderbouwing voor de validiteit van het TOEFL iBt onderdeel Spreken als maat voor communicatieve competentie, maar maakte ook duidelijk dat de automatische score niet volledig in staat was alle aspecten van communicatieve competentie op te pikken.

Ook Ockey et al. (2014) keken naar het onderdeel spreekvaardigheid van TOEFL iBT. Zij vergeleken de uitkomsten van de TOEFL spreektoets met die van een assessment van spreekvaardigheid aan een Japanse universiteit. De 'universiteitstoets' bestond uit drie onderdelen, een groepsgesprek met drie kandidaten, een individuele taak waarin een plaatje en een grafiek moesten worden beschreven, en een voorbereide presentatie in tweetallen. De taken werden analytisch beoordeeld op een negenpuntschaal, op de drie criteria uitspraak, vloeiendheid en woordenschat & grammatica. Daarnaast waren er taak-specifieke criteria: 'interactie' voor de eerste taak, 'de vaardigheid in het beschrijven' bij de tweede taak, en 'presentie' voor de derde taak. Het onderzoek had tot doel de relatie te onderzoeken tussen de TOEFL-scores en de analytische scores. Men wilde weten of TOEFL alle relevante aspecten van communicatieve competentie beoordeelt. De spreekprestaties in dit onderzoek waren afkomstig van 226 studenten Engels in Japan. Er waren 12 beoordelaars betrokken bij het universitaire assessment en de TOEFL-toetsen werden beoordeeld door gecertificeerde TOEFL-beoordelaars. De betrouwbaarheid van de drie onderdelen van het universitaire assessment was respectievelijk .78, .81 en .72, de betrouwbaarheid van TOEFL was .82. Correlaties van TOEFL met de onderdelen van het assessment luidden als volgt: .76 voor de groepsdiscussie, .73 voor de beschrijftaak en .68 voor de presentatie. Verder vond het onderzoek matige tot hoge correlaties tussen de verschillende beoordelingsaspecten en TOEFL. De correlaties tussen de criteria uitspraak, vloeiendheid en woordenschat & grammatica was hoog en de correlaties tussen TOEFL en de taak-specifieke criteria was matig. Ockey en collega's concluderen dat TOEFL een breed construct van spreekvaardigheid meet, maar dat de scores sterker samenhangen met linguïstische maten als vloeiendheid, uitspraak, woordenschat en grammatica dan met aspecten als interactionele competentie en presentatievaardigheden.

Ono et al. (2019) onderzochten de beoordeling van de TOEFL iBT geïntegreerde schrijftaak. In deze taak moeten kandidaten een tekst produceren op basis van een korte leestekst en een luisterfragment. 36 leerders van het Engels aan een universiteit in Japan maakten de schrijftaak en zeven beoordelaars, docenten Engels, beoordeelden de prestaties met zowel het holistische beoordelingsmodel van TOEFL als met een voor eerder onderzoek ontwikkeld analytisch model (Yang & Plakans, 2012, zie ook deelvraag 1). Het analytisch model keek naar de aspecten inhoud, organisatie, taalgebruik en letterlijk brongebruik. Zowel de TOEFL-schaal als de analytische schaal kent vijf schaalpunten. De beoordelaars beoordeelden eerst analytisch, daarna holistisch, en vulden een vragenlijst in. Uit de resultaten bleek dat de betrouwbaarheid van beide modellen goed was, respectievelijk α=.89 voor het holistisch model en α=.90 voor het analytisch model. De analytische en de holistische score correleerden significant en de score met het analytische model was iets hoger. De aspecten inhoud en organisatie correleerden hoog met de overall scores maar de beoordeling van het aspect brongebruik bleek problematisch en de correlaties waren laag. Uit de vragenlijsten bleek dat beoordelaars verschillende opvattingen hanteren over wat adequaat brongebruik is in een geïntegreerde schrijftaak. In de vragenlijsten gaven beoordelaars verder aan dat ze de analytische beoordeling tijdrovend vonden en dat ze de beslissing voor het toekennen van één van de verschillende schaalpunten vaak lastig vonden.

Ook Izadpanah et al. (2014) keken naar de TOEFL iBT schrijftoets. Zij wilden weten of scores op een analytische schaal de scores op de toets beoordeeld met een holistische model voorspellen. Het onderzoek werd uitgevoerd aan een universiteit in Iran en de producten van het TOEFL onderdeel schrijven van 160 studenten werden beoordeeld door drie beoordelaars die het onderzoek ook uitvoerden. De analytische schaal was een model ontwikkeld door Jacobs in 1981. Het model onderscheidt niet alleen verschillende aspecten maar kent ook een gewicht toe aan elk aspect. De aspecten zijn inhoud (30%), taalgebruik (25%), organisatie (20%), woordenschat (20%) en spelling en interpunctie (5%). De schrijfproducten werden beoordeeld met beide modellen en een multiple regressie analyse werd uitgevoerd om de relatie tussen de twee beoordelingsmodellen te onderzoeken en na te gaan welke aspecten de holistische score voorspellen. De scores op de beide modellen correleerden hoog en beide modellen waren succesvolle voorspellers van elkaar. Uit de analyses bleek verder dat taalgebruik de beste voorspeller was voor de holistische score, gevolgd

door spelling en inhoud. Woordenschat en organisatie waren geen significante voorspellers. De schrijfproducten werden op grond van beide beoordelingen verdeeld naar drie vaardigheidsniveaus, laag, midden en hoog. De indeling van producten van midden en hoog niveau correleerden hoog, maar het analytisch model bleek de vaardigheid van de kandidaten van laag niveau te onderschatten. De beoordelaars legden onterecht te veel nadruk op taalgebruik bij het beoordelen van deze producten.

4.9.4 Beoordelaars

- Beoordelaars kunnen hun eigen criteria hanteren bij het gebruik van een holistisch model.
- Beoordelingstraining en individuele feedback is belangrijk.

Een van de zorgen over het gebruik van holistische beoordelingsmodellen is dat beoordelaars mogelijk elk hun eigen criteria hanteren of hun eigen weging aanbrengen in criteria. Om zicht te krijgen op de denkprocessen van beoordelaars voerde Sakyi (2000) een onderzoek uit met behulp van hardop-denkprotocollen. De beoordelaars in dit onderzoek waren zes ervaren docenten Engels. Zij beoordeelden elk twaalf essays van leerders van het Engels, met behulp van een holistisch beoordelingsmodel. Tijdens het beoordelingsproces articuleerden zij hun afwegingen en gedachten. Op grond van een kwalitatieve analyse van de responsen vond de onderzoeker dat er vier beoordelaarsstijlen te onderscheiden waren. Het eerste type was met name gericht op fouten in de tekst, het tweede type focuste op inhoud, specifiek op de uitwerking van ideeën, het derde type is gebaseerd op een persoonlijke emotionele reactie op de inhoud van de tekst, en het vierde type focuste op de beoordelingsvoorschriften. Daarbij waren twee subtypen te onderscheiden, waarbij het eerste afging op de eerste indruk en het tweede de hele tekst las alvorens tot een oordeel te komen. Het onderzoek biedt daarmee ondersteuning voor de gedachte dat beoordelaars een holistisch model op verschillende manieren kunnen toepassen.

In een vervolgonderzoek keek Sakyi (2003) naar de verschillen in beoordelingsgedrag tussen ervaren en pasgetrainde beoordelaars, allen docenten Engels als tweede taal. In dit onderzoek werden zes benchmarkschrijfproducten van TOEFL beoordeeld door vijf ervaren beoordelaars en twaalf onervaren beoordelaars. De beoordelaars beoordeelden met het holistische TOEFL-model en voerden ook een paarsgewijze vergelijking uit (zie ook deelvraag 9). De analyse richtte zich op de aspecten waar beoordelaars op letten, op de criteria die ze gebruikten om de verschillende niveaus te onderscheiden, op de strategieën die ze gebruikten om tot een beslissing te komen, en op de verschillen in aanpak tussen gevorderde en nieuwe beoordelaars. Ook dit onderzoek vond de eerder gerapporteerde beoordelingsstijlen (Sakyi, 2000). De onervaren beoordelaars waren meer gefocust op inhoud en begrijpelijkheid en de ervaren beoordelaars keken sterker naar taalgebruik en organisatie van de tekst. Alle beoordelaars vonden het lastig om het onderscheid tussen de verschillende niveaus goed te maken, wat leidde tot lage beoordelaarsovereenstemming. Sakyi benadrukt het belang van beoordelaarstraining bij het gebruik van een holistisch beoordelingsmodel.

Met behulp van **multilevel modeling** keek Barkaoui (2010) naar de bijdrage van beoordelaars en tekst op holistische scores. Multilevel modeling is een manier om data te analyseren waarin mensen (of gegevens) deel uitmaken van groepen. Het model houdt rekening met die groepsstructuur, waarbij subgroepen ‘genest’ kunnen zijn binnen groepen. In dit onderzoek beoordeelden 31 ervaren beoordelaars en 29 nieuwe, onervaren beoordelaars 128 schrijfproducten van leerders van het Engels als tweede taal, zowel met een holistisch en een analytisch model. De schrijfproducten werden verdeeld in pakketjes van 24 teksten en elke tekst werd door meerdere beoordelaars beoordeeld, met zowel de holistische als de analytische schaal. De holistische en de analytische schaal hanteerden dezelfde evaluatiecriteria, en hanteerden elk een negenpuntschaal. Bij de holistische beoordeling moest één score worden toegekend aan het schrijfproduct en met de analytische schaal werden producten beoordeeld op communicatieve kwaliteit, organisatie, argumentatie, linguïstische accuraatheid en linguïstische adequaatheid. Daarnaast werd ook tekstlengte in de analyse betrokken. Uit de resultaten bleek dat er grote variatie was in de wijze waarop holistische en analytische scores samenhingen. Tekstlengte had een positieve relatie met de holistische scores, maar de relatie tussen de holistische scores enerzijds en de analytische criteria anderzijds kende grote variatie. Verder bleken ervaren beoordelaars lagere scores te geven dan nieuwe beoordelaars. Ook hechtten de ervaren beoordelaars meer waarde aan linguïstische accuraatheid. De onervaren beoordelaars hechtten meer waarde aan argumentatie. De beoordelaarsovereenstemming was groter tussen de ervaren beoordelaars, de oordelen van de nieuwe beoordelaars liepen meer uiteen.

Alaei en collega's probeerden te begrijpen waarom beoordelaars verschillen en keken daarbij naar de invloed van persoonlijkheidskenmerken op holistische en analytische beoordeling van schrijfproducten, en bovendien naar de interactie met genre (Alaei et al., 2014). De beoordelaars waren 31 Iraanse docenten Engels als tweede taal in Iran. Zij beoordeelden vier teksten, één die oorzaak en gevolg beschreef, een beschrijving, een opiniërende en een argumenterende tekst. Dat deden ze eerst impressionistisch-holistisch en daarna met behulp van een analytische schaal ontwikkeld door Jacobs in 1981, eveneens gebruikt door (Izadpanah en collega's (2014), zie paragraaf 4.9.3. Het model onderscheidt niet alleen verschillende aspecten maar kent ook een gewicht toe aan elk aspect. De aspecten zijn inhoud (30%), taalgebruik (25%), organisatie (20%), woordenschat (20%) en spelling en interpunctie (5%). Daarnaast vulden de beoordelaars een persoonlijkheidstest in met 60 items, die de aspecten openheid, consciëntieusheid, extraversie, vriendelijkheid en neuroticisme in kaart brengt. Het onderzoek vond geen significante correlaties tussen de vijf dimensies van persoonlijkheidskenmerken en de holistische scores op de vier tekstsoorten. Bij de vergelijking van de persoonlijkheidskenmerken en de analytische scores werd een significante positieve relatie gevonden tussen vriendelijkheid en inhoud: vriendelijke beoordelaars gaven hogere scores voor inhoud. Er werd ook een positieve relatie gevonden tussen consciëntieusheid en woordenschat: consciëntieuze beoordelaars gaven hogere scores voor woordenschat. De auteurs bepleiten beoordelaarstraining om individuele verschillen tussen beoordelaars te verkleinen.

Onderzoek van Pallotti (2022) vergeleek de dimensies van een holistisch model gericht op het beoordelen van functionele adequaatheid (FA) met in totaal elf analytische (CAF) maten, waarvan de meeste automatisch werden gescoord. Het doel van het onderzoek was te kijken welke analytische maten voorspellend waren voor de holistische beoordeling. Of andersom: waarop beoordelaars onbewust letten als ze een holistisch oordeel geven over FA. De data waren spreek- en schrijfproducten van jonge leerders van het Italiaans (N=217). De leerlingen moesten een stomme film navertellen en er daarna over schrijven. De prestaties werden beoordeeld op functionele adequaatheid, verder uitgesplitst naar inhoud, begrijpelijkheid, en coherentie. Pallotti vond dat de analytische maten maar een deel van de variantie van de holistische scores verklaarden. Het aantal secundaire ideeën (informatieve details) en lexicale diversiteit waren de beste voorspellers. FA werd geoperationaliseerd in drie dimensies. De dimensie inhoud werd het best voorspeld door het aantal en de uitwerking van ideeën. Begrijpelijkheid werd (deels) voorspeld door woordgebruik. Pallotti betoogt dat de resultaten belangrijk zijn bij de ontwikkeling van globale schalen: ontwikkelaars moeten zich bewust zijn van de manier waarop holistische beoordeling van FA en analytische maten samenhangen.

4.9.5 Kritische reflectie

Bij het beoordelen van productieve vaardigheden is er altijd interactie tussen de vaardigheid van de kandidaat, de taak, de beoordelaar en het gehanteerde beoordelingsmodel. We zijn geïnteresseerd in de vaardigheid van de kandidaat en moeten 'ruis' als gevolg van taak, beoordelaar en model zo veel mogelijk reduceren. Dit deelonderzoek richtte zich op holistische beoordelingsmodellen en het gebruik en de bruikbaarheid ervan. Uit de gelezen teksten viel allereerst een enorme variatie binnen de gebruikte modellen op. Analytische modellen verschillen in het aantal aspecten dat wordt onderscheiden, op welke aspecten zij focussen, hoeveel schaalpunten er worden gebruikt en of de verschillende schaalpunten worden beschreven of juist weer op gevoel, min of meer holistisch worden toegekend. Holistische modellen geven één score voor de taak, maar kunnen toch ook weer verschillende beoordelingsaspecten omschrijven, zoals bijvoorbeeld gebeurt in de TOEFL-schalen voor de beoordeling van spreektaken, waarbij beoordelaars op basis van een algemene beschrijving én een beschrijving van uitspraak en vloeiendheid, taalgebruik, en ontwikkeling van het onderwerp één best passende score toekennen. Zie https://www.ets.org/pdfs/toefl/TOEFL iBT-speaking-rubrics.pdf. Als we dit beoordelingsmodel vergelijken met het beoordelingsmodel van schrijven eveneens van TOEFL, dan vallen daarbij ook de grote verschillen in het operationaliseren van holistische beoordeling op. Zie https://www.ets.org/pdfs/toefl/TOEFL iBT-writing-rubrics.pdf

Ondanks de grote verscheidenheid in manieren waarop holistische en analytische modellen geoperationaliseerd werden, was er een duidelijke trend in de geraadpleegde literatuur, in het voordeel van analytische modellen. De conclusie van dit deelonderzoek is daarom: alle beoordelingsmodellen hebben nadelen maar holistische modellen hebben meer nadelen dan analytische modellen.

Een veelgenoemd voordeel van analytische modellen is dat ze de basis bieden voor feedback aan kandidaten. Hoewel de examens Spreken en Schrijven worden beoordeeld met een analytisch model, wordt tot op heden aan de kandidaten alleen een totaalscore gerapporteerd en krijgen zij geen informatie over hun scores op de verschillende aspecten. Vooral kandidaten die het examen moeten overdoen zouden gebaat zijn bij meer gedetailleerde informatie

over hun prestaties. Inmiddels wordt bij het staatsexamen Nt2 het verstrekken van feedback aan kandidaten dan ook voorbereid.

4.10 Andere beoordelingsmodellen

De laatste onderzoeksvraag luidt: Zijn er nog andere beoordelingsmodellen die worden gebruikt bij summatieve taalexamens? Wat zijn voor- en nadelen, wat zijn de ervaringen, hoe functioneren ze in de praktijk? De vraag is erg breed en het was niet makkelijk om de relevante zoekterm(en) te vinden. Met de zoekterm "rating scales" AND ("assessing" OR "testing") AND ("speaking" OR "writing" OR "productive skills") AND ("advantages" OR "disadvantages") werd een databankzoekactie uitgevoerd. Helaas bleken er erg veel overlappende resultaten met de literatuur gevonden bij deelvraag 5, namelijk algemene literatuur over het beoordelen van productieve vaardigheden met aandacht voor modellen en beoordelaars.

Uit het bestuderen van de literatuur voor de huidige deelvraag en de deelvragen 5 t/m 8 bleek dat er twee typen beoordeling worden besproken in de literatuur die nog niet aan niet aan bod waren gekomen in dit onderzoek. De eerste is *multiple trait scoring* en de tweede is paarsgewijze vergelijking (*comparative judgement*). Daarom is aanvullend gezocht met de zoekopdrachten "Multiple trait scoring" AND ("speaking" OR "writing" OR "productive skills") AND ("assessing" OR "testing") AND ("comparative judgement") AND ("speaking" OR "writing" OR "productive skills") AND ("assessing" OR "testing"). Beide manieren van beoordelen worden niet veelvuldig gebruikt, en zeker niet in de specifieke context van grootschalige highstakes toetsing van tweede taal. Het was dan ook niet mogelijk om twintig relevante teksten te vinden over deze onderwerpen. Aanvullend zijn nog enkele teksten gevonden door gebruik te maken van de referenties in de gelezen teksten. De gebruikte teksten zijn te vinden in Tabel 14.

Tabel 14 *Bestudeerde literatuur deelvraag 9*

1. Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes.*
2. Benton, T., & Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking? *Research Matters 26*. UCLES.
3. Bramley, T. (2015). *Investigating the reliability of Adaptive Comparative Judgment.* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.
4. East, M., & Young, D. (2007). Scoring L2 writing samples: Exploring the relative effectiveness of two different diagnostic methods. *New Zealand Studies in Applied Linguistics, 13*(1), 1-21.
5. Han, C. (2021). Analytic rubric scoring versus comparative judgment: A comparison of two approaches to assessing spoken-language interpreting. *Meta, 66*(2), 337-361.
6. Paquot, M., Rubin, R., & Vandeweerd, N. (2022). Crowdsourced adaptive comparative judgment: A community-based solution for proficiency rating. *Language Learning, 72*(3), 853-885.
7. Park, J. (2006). The study on the rater reliability of three scoring methods in assessing argumentative essays: Holistic, analytic, and multiple-trait scoring methods. *Foreign Language Education Research, 9*, 63-84.
8. Sickinger, R., Brunfaut, T., & Pill, J. (2025). Comparative Judgement for evaluating young learners' EFL writing performances: Reliability and teacher perceptions of holistic and dimension-based judgements. *Language Testing, 42*(2), 137-166.
9. Sims, M. E., Cox, T. L., Eckstein, G. T., Hartshorn, K. J., Wilcox, M. P., & Hart, J. M. (2020). Rubric rating with MFRM versus randomly distributed comparative judgment: A comparison of two approaches to second-language writing assessment. *Educational Measurement: Issues and Practice, 39*(4), 30-40.
10. Thwaites, P., Vandeweerd, N., & Paquot, M. (2025). Crowdsourced comparative judgement for evaluating learner texts: How reliable are judges recruited from an online crowdsourcing platform?. *Applied Linguistics, 46*(4), 611-628.
11. Von Zansen, A., & Huhta, A. (2022, December). Developing automated feedback on spoken performance: Exploring the functioning of five analytic rating scales using many-facet rasch measurement. In *Diversity of Methods and Materials in Digital Human Sciences: Proceedings of the Digital Research Data and Human Sciences DRDHum Conference 2022, December 1-3, Jyväskylä, Finland* (pp. 211-229). University of Jyväskylä.
12. Wheadon, C., Barmby, P., Christodoulou, D. & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England, Assessment in Education: Principles, Policy & Practice, 27:1, 46-64.

4.10.1 *Multiple trait scoring*

- Multiple trait modellen beoordelen verschillende aspecten van de taakuitvoering afzonderlijk.
- Naast aspecten van taalvaardigheid kan bijvoorbeeld communicatieve effectiviteit of communicatieve kwaliteit worden beoordeeld.
- Het verschil met analytische modellen, die immers ook vaak aspecten als inhoud of taakuitvoering beoordelen, is vaag.

Multiple trait beoordelingsmodellen lijken op analytische modellen: beoordelaars scoren verschillende aspecten van de prestatie afzonderlijk (J. Park, 2006). Het verschil is dat een analytisch model aspecten van één latente trek in kaart brengt: spreek- of schrijfvaardigheid. Een *multiple trait* model kan meerdere constructen meten (Kuiken & Vedder, 2021). Park (2006) voegt daaraan toe dat een *multiple trait* model taak-specifiek is. *Multiple trait* scoring benadrukt daarmee de contextafhankelijke dimensie van taalvaardigheid en de effectiviteit van communicatie binnen contexten. In de praktijk is het onderscheid overigens moeilijk te maken: een beoordelingsmodel met beoordeling van het aspect inhoud kan beschouwd worden als een *multiple trait* model omdat inhoud immers geen taalvaardigheid is. De specifieke aanvullende zoekopdracht leverde dan ook heel weinig op. Slechts enkele onderzoekers typeren hun beoordelingsmodel als een *multiple trait* model en de kenmerken van beschreven modellen overlappen met de analytische modellen die onder andere bij deelvraag 8 aan de orde kwamen.

Het onderscheid tussen een analytisch beoordelingsmodel en een *multiple trait* model wordt wél gemaakt door East en Young (2007). Zij vergeleken een *multiple trait* model voor de beoordeling van schrijfvaardigheid met een analytisch beoordelingsmodel. Beide modellen werden onderzocht om een eenvoudige binaire model dat in de onderzochte context werd gebruikt te vervangen. Het binaire model was een korte checklist met ja/nee vragen. Deze checklist was niet in staat om de sterke en zwakke punten van leerders aan het licht te brengen en bood zo weinig basis voor feedback. Er werden daarom twee nieuwe modellen getest. De nieuw ontwikkelde modellen waren (1) een multiple trait model, dat keek naar taakuitvoering, effectieve communicatie en taal, en (2) een analytisch model dat keek naar verschillende aspecten van communicatieve competentie: cohesie & coherentie, woordgebruik, grammaticale accuraatheid, spelling & interpunctie en register. Er werden met elk van de modellen dertig schrijfproducten beoordeeld door twee beoordelaars. Uit analyses bleek dat beide modellen goed functioneerden, met adequate beoordelaarsovereenstemming boven .80. Beide modellen waren in staat kandidaten in te delen naar vaardigheid. De auteurs spraken geen voorkeur uit voor één van de modellen.

J. Park (2006) vergeleek drie verschillende beoordelingsmodellen. Naast holistische en analytische beoordeling, besproken bij deelvraag 8, betrok hij ook een *multiple trait* beoordelingsmodel. Het *multiple trait* model beoordeelde de aspecten inhoud, organisatie, taal, en samenhang op een vierpuntschaal. Vijf beoordelaars beoordeelden 16 argumenterende teksten met drie verschillende modellen. Uit de resultaten bleek dat de scores toegekend met het holistisch model lager waren dan de scores toegekend met de beide andere modellen. De beoordelaarsovereenstemming was het hoogst met het *multiple trait* model en de beoordelaars hadden ook een voorkeur voor dit model. De overeenstemming tussen de aspecten van het analytisch model, dat de aspecten inhoud, taalgebruik, organisatie, woordenschat, en spelling & interpunctie beoordeelde met behulp van een vierpuntschaal, en het *multiple trait* model was hoog, wat suggereert dat de modellen uitwisselbaar zijn.

Barkaoui (2008) vergeleek het functioneren van een holistisch en een *multiple trait* model voor het beoordelen van schrijfvaardigheid. Bij het onderzoek waren 29 ervaren en getrainde en 31 nieuwe beoordelaars betrokken. Het *multiple trait* model was een negenpuntschaal en onderscheidde de aspecten communicatieve kwaliteit, organisatie, argumentatie, linguïstische accuraatheid en linguïstische beheersing. Het holistisch model hanteerde negen omschreven niveaus. De beoordelaars beoordeelden 24 schrijfproducten met de twee beoordelingsmodellen. In de data-analyse werd gekeken naar beoordelaarsovereenstemming en interne consistentie van beoordelaars. Uit de resultaten bleek dat dat de holistische beoordeling leidde tot betere beoordelaarsovereenstemming. Het *multiple trait* model leidde tot betere interne consistentie, met name bij de onervaren beoordelaars. De auteur betoogt dat het *multiple trait* model de betere optie is.

4.10.2 *Paarsgewijze vergelijking*

- Paarsgewijze vergelijking (CJ) lijkt een efficiënte manier van beoordelen.
- CJ kan leiden tot hoge betrouwbaarheid.
- Betrouwbaarheid kan worden gemanipuleerd door het aantal vergelijkingen per product te verhogen.
- Zowel experts als leken kunnen CJ uitvoeren.

Paarsgewijze vergelijking (in het Engels aangeduid als *comparative judgement*, afgekort CJ) is een proces waarbij meerdere beoordelaars paren van schrijf- of spreekprestaties ordenen naar niveau. Ze moeten daarbij steeds de vraag beantwoorden welk van de twee producten afkomstig is van de meer vaardige taalgebruiker. De beoordelingen worden verzameld in een digitaal systeem. Door herhaalde paarsgewijze vergelijking, met meerdere beoordelaars per prestatie, ontstaat er consensus over de rangorde. Paarsgewijze vergelijking representeert zo de collectieve expertise van beoordelaars. Met behulp van onder andere Rasch-modellen kan een vaardigheidsschaal worden gemodelleerd en kan bovendien de consistentie van beoordelaars in kaart worden gebracht (Han, 2021; Wheadon et al., 2020). Recent is er veel aandacht voor paarsgewijze vergelijking als beoordelingsmethode maar er zijn nog niet veel studies die de methode hebben onderzocht in de context van T2-leren (Sickinger et al., 2025).

Een grootschalige toepassing van CJ vond plaats in een onderzoek in het primair onderwijs in Engeland (Wheadon et al., 2020). Primair onderwijs is wellicht enigszins buiten de scope van het huidige onderzoek, maar het onderzoek is interessant vanwege de schaalgrootte. Bij dit onderzoek waren schrijfproducten (T1) van 55.599 basisschoolleerlingen uit de leerjaren 1 t/m 6 betrokken. Docenten moesten de schrijfproducten van hun leerlingen uploaden in een digitaal systeem, zie www.nomoremarking.com, en het systeem zorgde ervoor dat docenten niet hun eigen leerlingen beoordeelden. Er werden ankerteksten aan de pool van te beoordelen teksten toegevoegd om de consistentie van beoordelaars te kunnen monitoren. Uit het onderzoek bleek dat de betrouwbaarheid van de vergelijkingen zeer hoog was, enigszins variërend per leerjaar, met .85 als de laagste overeenstemming voor de producten van leerlingen uit het hoogste leerjaar. De tijd die leerkrachten kwijt waren met beoordeling via CJ verschilde ook per leerjaar. Voor het vergelijken van teksten van eerstejaars leerlingen was gemiddeld minder dan 11 seconden nodig, wat betekent dat leerkrachten 330 vergelijkingen kunnen maken in een uur. Voor het vergelijken van prestaties van leerlingen uit klas 6 was gemiddeld 41 seconden nodig, wat betekent dat een leerkracht 88 vergelijkingen per uur kan maken. De methode wordt daarmee niet erg tijdrovend geacht.

Paarsgewijze vergelijking is normaalgesproken gebaseerd op een holistisch oordeel. Sickinger et al. (2025) onderzochten de vraag of analytische criteria iets konden toevoegen aan de kwaliteit van CJ. Het onderzoek vond plaats in het voortgezet onderwijs in Oostenrijk, en de schrijfprestaties waren afkomstig van 300 jonge leerders van het Engels. In een landelijk onderzoek naar onderwijskwaliteit werd een schrijftoets afgenomen met een analytisch model. Omdat de beoordeling tijdrovend was, werd in dit onderzoek gekeken of CJ een goed alternatief was, dat een vergelijkbare uitslag zou kunnen rapporteren, even betrouwbaar was, en eenvoudiger toepasbaar. Er waren 27 docenten betrokken bij het proces van paarsgewijs vergelijken en alle docenten voerden drie taken uit: ze voerden twee keer 78 paarsgewijze vergelijkingen uit, met en zonder analytische dimensies, en ze beoordeelden 30 teksten met behulp van een analytisch model. De aspecten die werden beoordeeld in zowel het analytisch model als in de paarsgewijze vergelijking met dimensies waren: taakuitvoering, coherentie & cohesie, grammatica en woordenschat. Het onderzoek vond dat CJ tot hoge betrouwbaarheid, op of boven .90 leidde. De prestaties van de beoordelaars waren overwegend goed bij de paarsgewijze vergelijking met een enkele inconsistente beoordelaar. Met gebruik van het analytisch model werden er meer inconsistente beoordelaars gevonden. De correlaties tussen holistische en analytische CJ waren hoog. Beoordelaars vonden de analytische CJ moeilijker dan de holistische en ook moeilijker dan de beoordeling met het analytische model. De analytische CJ kostte meer tijd dan de holistische, maar beide vormen van CJ gingen wel sneller dan de beoordeling met het analytische model.

Sims et al. (2020) vergeleken CJ met een traditioneel analytisch beoordelingsmodel en keken daarbij naar het beoordelingsgedrag van ervaren en onervaren beoordelaars. Er waren twee groepen van acht beoordelaars betrokken bij het onderzoek, elk bestaand uit vier ervaren en vier onervaren beoordelaars. Elke groep werd gevraagd om 60 schrijfproducten te beoordelen door middel van CJ en met het beoordelingsmodel. Voor elke tekst was ook een expertbeoordeling beschikbaar die was verkregen met behulp van een beoordelingsmodel. De auteurs rapporteerden een hoge mate van betrouwbaarheid voor de CJ-beoordeling, van .89 en hoger. De expertbeoordelaars beoordeelden iets betrouwbaarder dan de onervaren beoordelaars.

In de context van onderzoek (en dus niet in de context van het beoordelen van vaardigheid van kandidaten) keken Thwaites et al. (2025) naar de inzet van leken bij het paarsgewijs vergelijken van schrijfproducten. Het onderzoek was een replicatie van een studie (Paquot et al., 2022), waaruit was gebleken dat taalkundigen die door middel van *crowd sourcing* waren gerekruteerd (en die dus geen getrainde beoordelaars waren) in staat bleken paarsgewijze vergelijkingen uit te voeren met een hoge mate van betrouwbaarheid (.95). In het huidige onderzoek werden geen taalkundigen maar leken (N=37) ingezet die werden gerekruteerd via een *crowd sourcing* platform genaamd Prolific.

De enige eis aan de beoordelaars was dat ze moedertaalsprekers van het Engels waren. De beoordelaars voerden elk tien paarsgewijze vergelijkingen uit op teksten die waren geschreven in de context van de TOEFL schrijftoets. Van deze prestaties waren ook de originele scores beschikbaar. Er werd een softwaretool gebruikt om de vergelijkingen uit te voeren en beoordelaars kregen de instructie per paar teksten de vraag te beantwoorden welk van de twee teksten geschreven was door de meest gevorderde leerder van het Engels. Uit de data-analyse bleek dat de betrouwbaarheid van de beoordelingen hoog was (.94) en dat de rangorde zoals aangebracht door de leken hoog correleerde met de TOEFL-scores (r=.90). De uitkomsten suggereren dat paarsgewijze vergelijking makkelijk uitvoerbaar is, zelfs voor ongetrainde beoordelaars die geen taalkundigen zijn.

Niet alle onderzoek naar CJ vindt ondersteuning voor de methode. Han (2021) vergeleek paarsgewijze vergelijking met een analytisch beoordelingsmodel. De data voor dit onderzoek waren afkomstig van een mondeling vertaal-examen voor tolken Engels-Chinees. De beoordelaars waren twintig bachelorstudenten en twintig masterstudenten, allen met Mandarijn-Chinees als T1 en Engels als T2. De beoordelaars beoordeelden 40 opgenomen vertalingen, eerst door middel van paarsgewijze vergelijking en drie weken later met een analytisch model. Het analytische model beoordeelde drie aspecten op een achtpuntschaal: volledigheid, vloeiendheid en accuraatheid. Uit eerder onderzoek was gebleken dat deze analytische schaal goed functioneerde. Voor de 40 vertalingen waren ook de scores beschikbaar die studenten oorspronkelijk hadden gekregen en die waren toegekend met een holistisch model. Uit vergelijking van de paarsgewijze vergelijking en de scores met het analytisch model met de oorspronkelijk toegekende scores bleken hoge correlaties, alle boven .80. Dat bevestigt de validiteit van de beide onderzochte beoordelingstechnieken. Uit analyses van de beoordelingsovereenstemming bleek dat het analytisch model leidde tot betere overeenstemming. Beide modellen presteerden gelijk in termen van beoordelaarsconsistentie en inconsistente beoordelaars kwamen in beide condities voor. De bachelor studenten (onervaren beoordelaars) hadden meer tijd nodig voor het uitvoeren van hun paarsgewijze vergelijkingen dan de masterstudenten. Na het beoordelen vulden de beoordelaars een vragenlijst in en daaruit bleek dat de beoordelaars bij het gebruik van het analytische model zekerder waren van hun oordeel. Het onderzoek suggereert dat een analytisch model de beste optie is voor beoordeling van spreekproducten van studenten in een vertaalopleiding.

Een specifieke manier van paarsgewijze vergelijking wordt aangeduid als ***Adaptive Comparative Judgement*** **(ACJ)**. Het doel hiervan is om de tijd van beoordelaars zo efficiënt mogelijk te gebruiken door specifiek die prestaties aan beoordelaars voor te leggen waarover nog geen duidelijk oordeel kan worden geveld. ACJ heeft in die zin overeenkomstige kenmerken met adaptief toetsen. Het verschil is echter dat in een adaptieve toets de itemparameters geschat zijn voor afname van de toets. Bij ACJ is dat niet het geval. Bramley (2015) onderzocht de betrouwbaarheid van ACJ door het uitvoeren van een aantal simulaties. Betrouwbaarheid in CJ wordt uitgedrukt als de **Scale Separation Index, SSR**. Onderbouwd door de data uit de simulaties betoogde Bramley dat de SSR op basis van ACJ een vertekend beeld geeft. De bias wordt groter naarmate het aantal vergelijkingen waarop de schaal gebaseerd is groter wordt. Bramley concludeerde dat de SSR “is at best misleading and at worst worthless as an indicator of scale reliability whenever a CJ study has involved a significant amount of adaptivity” (p. 15). De betrouwbaarheid kan worden vergroot door correlaties met de oordelen van een tweede groep beoordelaars of door het toevoegen van ankerprestaties.

Paarsgewijze vergelijking wordt doorgaans geassocieerd met hoge betrouwbaarheid maar de resultaten van onderzoek zijn niet eenduidig. Om meer vat te krijgen op de kenmerken die de betrouwbaarheid van CJ beïnvloeden, voerden Verhavert et al. (2019) een meta-analyse uit op 49 onderzoeken die gebruik hadden gemaakt van CJ. Betrouwbaarheid van CJ kan worden gemanipuleerd door enerzijds het aantal vergelijkingen per prestatie (het aantal malen dat een prestatie wordt vergeleken met de prestaties van verschillende andere kandidaten) en anderzijds het aantal beoordelaars. Daarnaast heeft ook expertise van de beoordelaars en de te beoordelen vaardigheid mogelijk invloed op de betrouwbaarheid van CJ. De data voor dit onderzoek waren afkomstig van een online CJ assessment tool, genaamd *Digital Platform for the Assessment of Competences* (D-PAC; www.d-pac.be). Er waren data beschikbaar van 49 onderzoeken, afkomstig uit primair en voortgezet onderwijs, hoger onderwijs, onderzoek en selectie.

In 32 onderzoeken werd geschreven tekst beoordeeld, tien onderzoeken hadden betrekking op spreekprestaties, vijf op beeld en twee op portfolio’s. Uit analyses bleek dat alleen het aantal vergelijkingen per product grote impact heeft op de betrouwbaarheid van CJ. Andere variabelen zoals taaktype of expertise van beoordelaars hadden geen invloed. De onderzoekers waren ook in staat het vereiste aantal vergelijkingen in relatie tot betrouwbaarheid te schatten: voor

een betrouwbaarheid van .70 zijn gemiddeld 13 vergelijkingen per product nodig, voor een betrouwbaarheid van .80 ongeveer 20 en voor een betrouwbaarheid van .90 zijn tussen de 26 en 37 vergelijkingen nodig.

In een enigszins kritisch artikel onderbouwden Benton en Gallacher (2018) de stelling dat CJ 'slechts' een vorm is van herhaald beoordelen[11] met een eenvoudig beoordelingsmodel en een geavanceerd statistisch model. Herhaald beoordelen en geavanceerde statistiek is ook mogelijk bij het gebruik van traditionele beoordelingsmodellen. Zo is het bijvoorbeeld mogelijk om streng- en mildheid van beoordelaars te verrekenen in scoretoekenning en uiteraard is ook de inzet van een grote hoeveelheid beoordelaars mogelijk. Voor dit onderzoek gebruikten ze data uit twee eerdere onderzoeken. De data waren afkomstig van het onderdeel Engels van de GCSE, een eindexamen in de UK. De dataset was gebruikt in twee eerdere onderzoeken en de gegevens uit die onderzoeken waren beschikbaar. Het ene onderzoek werd uitgevoerd door Bramley en Vitello (2018), en vergeleek CJ en ACJ. Bij dit onderzoek waren 34 beoordelaars betrokken. Het andere onderzoek werd in 2015 uitgevoerd door Child, Munro en Benton, en onderzocht een traditioneel model. Bij dit onderzoek waren 17 beoordelaars betrokken. Bovendien was de oorspronkelijk score die leerlingen hadden ontvangen op basis van één examinator beschikbaar. Er waren dus in totaal 18 traditionele oordelen. De data werden gemodelleerd en geanalyseerd: er werd gekeken naar het toevoegen van meer beoordelaars in het traditionele model, naar de streng- en mildheid van beoordelaars en er werd een paarsgewijze vergelijking gesimuleerd op basis van de traditionele oordelen. De resultaten onderbouwden de stelling dat de kracht van CJ ligt in de inzet van meerdere beoordelaars en een sterk statistisch model, en dat dezelfde technieken kunnen worden gebruikt bij traditionele beoordeling, met dezelfde uitkomsten.

4.10.3 Kritische reflectie

Deze paragraaf besprak allereerst multiple trait scoring. Hoewel de literatuur deze manier van beoordelen noemt als een vorm die afwijkt van analytische beoordeling, lijken de verschillen uiterst mager en kan in principe elk analytisch model dat ook inhoud beoordeelt als een multiple trait model worden beschouwd. Het is echter zo dat veel gangbare analytische modellen, waaronder die die bij het staatsexamen Nt2 worden gebruikt, het aspect inhoud taak-specifiek beoordelen zonder de term 'multiple trait-model' te omarmen. In dit geval lijkt het meer te gaan om terminologische voorkeuren dan om een wezenlijk andere manier van beoordelen.

Die 'wezenlijk andere manier van beoordelen' vinden we wel bij het paarsgewijs vergelijken, een manier om kandidaat-producten te ordenen naar vaardigheid. De beoordeling van de twee producten gebeurt impressionistisch en holistisch, en veel training lijken beoordelaars niet nodig te hebben. Onderzoek vond dat bij het maken van voldoende vergelijkingen (waarbij het product van de kandidaat meermaals en door meerdere beoordelaars wordt vergeleken met steeds andere producten), paarsgewijs vergelijken leidt tot hoge betrouwbaarheid, wat een sterk punt is van de methode.

11 Bij CJ worden producten vergeleken met een relatief groot aantal andere producten. Zeg dat dit aantal 15 is, dan wordt het product dus 15 keer beoordeeld. Anders verwoord: de beoordeling wordt 15 keer herhaald.

5 Conclusies en aanbevelingen

5.1 Conclusies

- De algemene conclusie van dit onderzoek is dat de huidige vormgeving van het staatsexamen Nt2 gebaseerd is op zeer verdedigbare keuzes.

Dit literatuuronderzoek beantwoordt in totaal tien onderzoeksvragen met betrekking tot toetsing van taalvaardigheid in tweede talen. Toegespitst op inhoud, vormgeving en beoordeling van met name de productieve vaardigheden richtte het onderzoek zich op de ontwikkelingen die er in dit vakgebied in de afgelopen 25 jaar zijn geweest. In deze periode zijn er natuurlijk verschillende maatschappelijke en technologische ontwikkelingen geweest. Relevante maatschappelijke ontwikkelingen zijn onder andere veranderde opvattingen over migratie. Dit heeft ook geleid tot het gebruik van taaltoetsen in de context van verblijfsrecht, en daarmee een nieuwe dimensie toegevoegd aan het begrip 'highstakes'. Vragen over validiteit en betrouwbaarheid van toetsen zijn daardoor relevanter dan ooit. Daarnaast hebben technologische ontwikkelingen in de laatste decennia nieuwe mogelijkheden gecreëerd voor het toetsen van taalvaardigheid, ontwikkelingen die recent door de opkomst van AI (buiten de scope van dit onderzoek) mogelijk nog worden versneld.

Een eerste deelvraag die het huidige onderzoek bespreekt betreft geïntegreerd toetsen. Uit het onderzoek bleek dat grootschalige highstakes toetsen zoals TOEFL gebruik maken van de integratie van vaardigheden, waarbij korte lees- en luisterteksten worden aangeboden als input voor spreek- en schrijftaken. De inputteksten zijn zeer kort, en naast deze geïntegreerde spreek- en schrijfopdrachten worden de vaardigheden lezen en luisteren afzonderlijk getoetst. Geïntegreerd toetsen van vaardigheden is dus mogelijk in een highstakes context. Als geïntegreerd toetsen overwogen zou worden om de validiteit van taken te vergroten, dan overtuigt de uitwerking zoals bij TOEFL niet: de inputteksten zijn zeer kort en de toetstaken zijn hiermee geen goede afspiegeling van taalgebruik in academische contexten. Daarnaast lijkt de ontwikkellast van dit type geïntegreerde taken hoog. De tweede deelvraag ging dieper in op vormen van integratie. Er werden enkele toetsvormen besproken waarin vaardigheden integratief worden getoetst, echter gebaseerd op een achterhaald theoretisch model. Daarnaast werd aandacht besteed aan de integratie van verschillende niveaus in één toets. Veel grote toetsen voor het Engels rapporteren een schaalscore aan de kandidaat, in plaats van een zak-slaagbeslissing. Afnemers bepalen welke score in hun context nodig is voor toegang tot bijvoorbeeld een opleiding. Voor de kandidaat heeft dit model mogelijk voordelen.

De derde deelvraag betrof het toetsen van spreekvaardigheid en de vraag of er alternatieven waren voor de computerafname. De paragraaf maakte duidelijk dat sommige highstakes examens nog altijd gebruik maken van een interview, waarbij een examinator praat met de kandidaat. Waar de computerafname wellicht vragen oproept over de validiteit van de toets, roept het interviewexamen grote vragen op over de betrouwbaarheid: de beoordelaar heeft de dubbele rol van examinator en beoordelaar en de kandidaat is afhankelijk van die ene beoordeling. Uiteraard is een computerafname ook efficiënter. De vierde deelvraag besprak lezen van scherm en van papier. Het onderzoek maakte duidelijk dat lezen van papier nog altijd voordelen biedt voor de lezer, met name als het gaat om het lezen van complexe informatieve teksten. Deze worden beter begrepen bij het lezen van papier. De verschillen zijn in de laatste jaren wel kleiner geworden.

Het tweede deel van het literatuuronderzoek, met de deelvragen 5 t/m 9, besprak verschillende beoordelings-modellen voor het beoordelen van productieve vaardigheden: spreken en schrijven. Dit deel vestigde de aandacht op de beoordelaar als bron van 'ruis', als gevolg van streng- of mildheid, inconsistentie, ervaring, of wijze van interpretatie van de scoringsvoorschriften. Tevens bleek dat er een grote variatie aan beoordelingsmodellen wordt gebruikt bij het beoordelen van productieve vaardigheden. Voor deelvraag 6, de vraag naar positieve beoordelingsmodellen, werd onderzoek bestudeerd dat keek naar complexiteit, accuraatheid en vloeiendheid. Het onderzoek liet zien dat deze maten samenhang vertonen met de vaardigheid van kandidaten. CAF-maten kunnen

zeer gedetailleerd in kaart worden gebracht met behulp van software. Deze software pikt echter niet alle aspecten van communicatieve competentie op.

Deelvraag 7 besprak dichotome beoordelingsmodellen. Een deel van het gevonden onderzoek suggereert dat het mogelijk is om een fijnmazig dichotoom model te ontwikkelen (denk aan 30 of meer beoordelingsaspecten) waarmee taalvaardigheid betrouwbaar kan worden beoordeeld. Beknopte modellen met minder aspecten functioneerden minder goed. De ontwikkel- en beoordelingslast lijkt daarmee hoog. Deelvraag 8 ging dieper in op holistische beoordelingsmodellen. Uit het onderzoek bleek dat holistische beoordeling betrouwbaar kan zijn, maar bij vergelijking met een analytisch beoordelingsmodel blijkt meestal een voordeel voor het laatste. Analytische modellen leiden vaak tot hogere betrouwbaarheid en bieden de mogelijkheid voor feedback aan kandidaten. De laatste deelvraag besprak beoordeling via paarsgewijze vergelijking, een innovatieve wijze van beoordeling. Met deze methode kan een hoge mate van betrouwbaarheid worden bereikt als schrijf- of spreekproducten door een panel van beoordelaars worden vergeleken met voldoende andere producten. Voorbeelden van grootschalige toepassing in een toetscontext zijn interessant.

De algemene conclusie van dit onderzoek is dat de huidige vormgeving van het staatsexamen Nt2 gebaseerd is op zeer verdedigbare keuzes. Op basis van de literatuur is er weinig reden om af te wijken van het model van vier vaardigheden (deelvraag 1 en 2). De digitale afname van de spreekexamens heeft, zowel op het gebied van betrouwbaarheid als op het gebied van efficiëntie duidelijke voordelen boven een examen op basis van een mondeling interview (deelvraag 3). Lezen van scherm leidt nog altijd tot gereduceerd tekstbegrip ten opzichte van lezen van papier, met name als lezers meer complexe informatieve teksten moeten lezen, wat de keuze rechtvaardigt om bij de staatexamens Nt2 de leesteksten op papier aan te bieden aan kandidaten. Met betrekking tot beoordelingsmodellen biedt het analytische beoordelingsmodel voordelen boven andere gangbare modellen, met name op het gebied van betrouwbaarheid.

5.2 Aanbevelingen

Hoewel de algemene conclusie is dat het staatsexamen Nt2 nog goed aansluiten bij hedendaagse opvattingen van taalvaardigheidstoetsing, heeft het onderzoek ook geleid tot enkele aandachtspunten en aanbevelingen ter verbetering van de kwaliteit. Dat geldt ten eerste voor feedback aan kandidaten. Eén van de belangrijke voordelen van analytische beoordelingsmodellen, is dat deze de basis bieden voor feedback aan kandidaten over hun prestaties. Bij het staatsexamen Nt2 wordt echter aan kandidaten tot op heden alleen een totaalscore gerapporteerd. Kandidaten die zakken krijgen daarmee geen informatie over de vraag aan welke aspecten van hun taalvaardigheid ze vooral nog moeten werken. Inmiddels wordt het geven van feedback aan gezakte kandidaten voorbereid, een belangrijke service aan kandidaten die het examen moeten herkansen. Hiermee kunnen kandidaten zich beter voorbereiden op een volgend examen.

Inspiratie voor verdere verbetering van de examenervaring voor kandidaten is te vinden bij de Amerikaanse taaltoets COPI. Deze computertoets geeft de kandidaat op verschillende punten autonomie. Ten eerste kan de kandidaat het niveau kiezen waarop die een opgave wil maken en daarna een vervolgopgave kiezen die makkelijker, even moeilijk of moeilijker is dan de vorige. Een algoritme zorgt ervoor dat de kandidaat genoeg opgaven krijgt om een niveau te bepalen. Ten tweede kan de kandidaat het onderwerp van de taak kiezen. Bovendien kan de kandidaat binnen grenzen bepalen hoeveel denktijd die nodig heeft. Doordat de kandidaat zeggenschap heeft over de afname, wordt stress gereduceerd (Kenyon & Malone, 2010). Het is aan te bevelen om na te gaan in welke opzichten de autonomie van kandidaten bij het staatsexamen Nt2 kan worden vergroot.

Een opvallend verschil tussen het staatsexamen Nt2 en bekende taaltoetsen voor het Engels (TOEFL, IELTS) is, dat het staatsexamen Nt2 een zak-slaagbeslissing aan de kandidaat rapporteert, terwijl de grote toetsen voor het Engels ervoor kiezen een score te rapporteren, met een verwijzing naar een bijbehorend niveau (geen van deze toetsen gebruikt het ERK als referentiekader). Het is aan afnemers, zoals scholen en universiteiten, om vervolgens te bepalen welke schaalscore zij nodig achten voor toelating. Hoewel dit meer vergt van stakeholders, kan het belangrijke voordelen hebben voor de kandidaat: een document waarop staat dat de kandidaat gezakt is, vertegenwoordigt voor de kandidaat geen waarde. Echter, als een kandidaat die een programma II-examenonderdeel aflegt en daarop een score van 495 behaalt daarmee zou kunnen aantonen dat het niveau B1 wordt beheerst, dan heeft dat wel

waarde. Het B1-niveau geeft de kandidaat toegang tot opleidingen in het mbo en indien van toepassing voldoet de kandidaat ermee aan een deel van de inburgeringseisen. IELTS en TOEFL volgen deze rapportagesystematiek voor alle vaardigheden. Onderzoek van Hulstijn en collega's (2012) liet in een Nederlandse context zien dat twee sets spreekopdrachten, die ontwikkeld waren om specifiek B1 of B2 te meten, beide even goed in staat waren leerders op B1- en B2-niveau te onderscheiden. Dat biedt in elk geval voor spreken enige onderbouwing van de mogelijkheden tot integratie. Uiteraard verdient het cognitieve profiel van de programma I-kandidaat aandacht als bij het staatsexamen Nt2 zou worden besloten integratie van taalniveaus verder te verkennen.

Een tweede cluster van aanbevelingen heeft betrekking op de efficiëntie van de examens. Ook hierbij was het met name onderzoek naar de taaltoetsen TOEFL iBT en IELTS, dat leidt tot vragen over de efficiëntie van de onderdelen Spreken en Schrijven van het staatsexamen Nt2. In een taaltoets schatten we de vaardigheid van een kandidaat. Een goede schatting is reproduceerbaar en zoveel mogelijk vrij van ruis (construct-irrelevante variatie). Ruis kan worden veroorzaakt door beoordelaars (die mild, streng, of inconsistent kunnen beoordelen) of door taken (die bijvoorbeeld als gevolg van vertrouwdheid met het onderwerp voor de ene kandidaat makkelijker zijn dan voor de andere kandidaat van hetzelfde taalvaardigheidsniveau). Deze bronnen van ruis kunnen worden gereduceerd door het inzetten van een voldoende groot aantal taken en door het inzet van meerdere beoordelaars, met hogere betrouwbaarheid tot gevolg.

Aan beide voorwaarden wordt bij het staatsexamen Nt2 ruimschoots voldaan: zowel voor spreken als voor schrijven en zowel voor programma I als voor programma II worden tien of meer taken ingezet. Door gesegmenteerd beoordelen worden ook beoordelaarseffecten sterk gereduceerd, omdat een groot panel van beoordelaars betrokken is bij de beoordeling van elk examen van individuele kandidaten. Daarbij wordt elke opgave door twee beoordelaars beoordeeld. Bij vergelijking met de grote toetsen voor het Engels valt op hoeveel groter de inzet is bij het staatexamen Nt2. Bijvoorbeeld voor Schrijven geven zowel IELTS als TOEFL twee taken, waarbij in totaal 400 (IELTS) en 450 (TOEFL) woorden geschreven moeten worden. Voor spreken geeft IELTS drie, en TOEFL vier taken. Onderzoek naar het optimale aantal taken en beoordelaars, zoals uitgevoerd door Lee (2006), Xi (2007) en Plakans et al. (2019) laat echter twee dingen zien: de impact van het aantal taken op de betrouwbaarheid is veel groter dan de impact van het aantal beoordelaars (dit bleek ook uit onderzoek van Bachman, 2002; Frost et al., 2020; H. T. D. Huang et al., 2018; Shin & Ewert, 2015). Bovendien is er een grens aan de impact van toevoegen van meer items na een bepaalde hoeveelheid items. De betrouwbaarheid stijgt sterk als er twee items worden ingezet in plaats van één, maar de impact van een zesde of zevende item op de betrouwbaarheid is veel kleiner. Het is daarom aan te bevelen voor zowel Schrijven als Spreken te laten evalueren of de huidige inzet van beoordelaars en opgaven optimaal is, als betrouwbaarheid van de toets en efficiëntie van ontwikkel-, afname en beoordelingskosten worden afgewogen.

Met focus op beoordeling verdient wellicht ook het voorbeeld van TOEFL iBT Spreken en Schrijven aandacht. Deze examens worden beoordeeld door een menselijke beoordelaar en daarnaast automatisch gescoord. De computer beoordeelt linguïstische aspecten zoals uitspraak, spelling, en grammatica, en doet dat altijd meer consistent dan een menselijke beoordelaar dat zou kunnen doen. De menselijke beoordelaar kijkt ook naar taalgebruik en verder naar inhoud en begrijpelijkheid. De inzet van een automatische scoring naast een menselijke beoordeling combineert de sterke kanten van beide beoordelingen en is (na ontwikkeling) uiteraard efficiënter dan het inzetten van twee menselijke beoordelaars. Nader onderzoek naar automatische scoring is daarom te overwegen. Meer algemeen lijkt het raadzaam ontwikkelingen op het gebied van de inzet van grote taalmodellen in de context van toetsing kritisch en nieuwsgierig te volgen.

5.3 Beperkingen van het onderzoek

Voor dit onderzoek zijn ruim 200 teksten bestudeerd over toetsing van taalvaardigheid. Daarbij moet voor ogen worden gehouden dat alle literatuur gezocht is met Engelse zoektermen en dat alle gelezen publicaties in het Engels waren geschreven. Hierdoor zijn wellicht sommige toetsen uit de buurlanden, zoals het Duitse Goethe-examen of de Franse DALF en DELF-examens onderbelicht gebleven. Een opvallend kenmerk van de onderzoeksliteratuur is dat een zeer groot deel van de het bestudeerde onderzoek participanten inzet die in de literatuur als WEIRD worden aangeduid: Western, Educated, Industrialized, Rich en Democratic. Dat is niet typisch voor het huidige onderzoek maar een breder verschijnsel in de onderzoeksliteratuur (Andringa & Godfroid, 2020). Daarnaast is taalonderzoek in de meeste gevallen uitgevoerd onder leerders van het Engels (Plonsky, 2023). Voor het huidige onderzoek komt

daarbij dat de focus van onderzoek zeer sterk ligt op taaltoetsing in academische contexten. Dat maakt dat enige voorzichtigheid moet worden betracht bij het generaliseren van conclusies.

Een inhoudelijke beperking van het huidige onderzoek is dat de focus van de onderzoeksvragen sterk lag op het toetsen en beoordelen van de productieve vaardigheden Spreken en Schrijven, terwijl de vaardigheden Lezen en Luisteren onderbelicht zijn gebleven. Mogelijk zou de vraag ‘zijn de examens nog up-to-date?’ ook voor de receptieve vaardigheden in de toekomst nog beantwoord kunnen worden.

Begrippenlijst

Begrip	Uitleg
Adaptive comparative judgement	Een vorm van *comparative judgement* waarbij presta-ties gecontroleerd worden voorgelegd aan beoordelaars, op basis van de overeenstemming die bleek uit eerdere beoordelingen.
Beoordelaarsovereenstemming	De mate waarin oordelen van beoordelaars overeenkomen.
Beslisboom	Een beoordelingsmodel dat gebruik maakt van ja/nee-vragen.
CAF	Complexity, Accuracy, Fluency. Deze maten worden gebruikt om de kwaliteit van taalproductie te beschrijven in linguïstische termen.
CAT	Computer Adaptive Test
CBT	Computer Based Test
CMT	Computer Mediated Test
CJ	Comparative judgement
Clozetoets	Toetsvorm op basis van weggelaten woorden in een leestekst.
CMD	Cognitive Diagnostic Modeling
Cognitive Diagnostic Modeling	Statistisch model voor de analyse van beoordeling met EDD-schalen
C-test	Toetsvorm op basis van weggelaten woorden in een leestekst. Een C-toets geeft beginletters van het weggelaten woord.
Comparative judgement	Paarsgewijze vergelijking, een innovatieve manier van beoordelen waarbij de beoordelaars steeds twee prestaties met elkaar vergelijken en de vraag beantwoorden welke van de twee afkomstig is van de meest vaardige taalleerder.
Dependability	Generaliseerbaarheid, binnen G-theorie de term die wordt gebruikt voor betrouwbaarheid
EBB-schaal	Empirically derived, Binary-choice, Boundary-definition scale. Een beslisboom voor de beoordeling van spreek- of schrijfproducten.
EDD-schaal	Empirically-derived Descriptor-based Diagnostic checklist (EDD), een checklist voor het beoordelen van schrijfvaardigheid.
Geïntegreerde toets	Een toets, doorgaans voor de beoordeling van spreken of schrijven, met taken die tevens een beroep doen op lezen en/of luisteren.
Generaliseerbaarheid	Betrouwbaarheid
Generaliseerbaarheidstheorie	Een uitbreiding van de klassieke toetstheorie die de verschillende bronnen van meetfouten (zoals verschillen tussen beoordelaars of items) analyseert om te bepalen hoe generaliseerbaar (betrouwbaar) een test-score is. Variantieanalyse wordt gebruikt om de bijdrage van elke bron aan de totale fout te kwantificeren.
G-theorie	Generaliseerbaarheidstheorie
Inferenties	In een leestoets zijn vragen die inferenties vereisen de vragen waarbij de kandidaat een conclusie moet trekken op basis van beschikbare informatie, bijvoorbeeld door tekstfragmenten te vergelijken. Het antwoord op de vraag staat niet letterlijk in de tekst.
Integratieve toets	Cloze-toets of gatentekst, toets waarin de aspecten grammatica, woordenschat en spelling in samenhang worden getoetst.
Interbeoordelaarsbetrouwbaarheid	De mate van overeenstemming tussen de beoordelingen van twee of meer onafhankelijke beoordelaars.
Intrabeoordelaarsbetrouwbaarheid	De mate waarin een beoordelaar consistent dezelfde resultaten behaalt bij herhaalde beoordeling van dezelfde data.

Begrip	Uitleg
Lexeem	Eenheid van betekenis die niet verder vereenvoudigd kan worden. Bijvoorbeeld: het lexeem (en woord) 'hoog' kunnen we verbuigen tot 'omhoog', 'verhogen', etc.
Many-Faceted Raschmodel	Een statistische benadering die wordt gebruikt om spreek- en schrijf-vaardigheidstoetsen te evalueren, met name gericht op de variabiliteit die wordt geïntroduceerd door verschillende 'facetten', zoals beoordelaars, taken en kandidaten.
Multilevel modeling	Een procedure om geneste data te analyseren (zoals cursisten binnen groepen)
Onafhankelijke taken	Taken die één productieve vaardigheid meten en geen input bieden in de vorm van een lees- of luistertekst
Paarsgewijze vergelijking	Een innovatieve manier van beoordelen waarbij de beoordelaars steeds twee prestaties met elkaar vergelijken en de vraag beantwoorden welke van de twee afkomstig is van de meest vaardige taalleerder. Engels: comparative judgement, CJ.
PBLA	Portfolio-based language assessment
Retrofit	Het gebruik en eventueel aanpassing van een bestaande toets voor een nieuw doel of een nieuwe populatie.
RLW-taak	Read, Listen, Write. Geïntegreerde schrijftaak op basis van een lees- en een luistertekst
Ruis	Construct-irrelevante variatie, die een toetsscore onbedoeld beïnvloedt, zoals strengheid van een beoordelaar of onbekendheid van de kandidaat met het onderwerp van een taak.
Scale separation index	De maat voor betrouwbaarheid bij paarsgewijze vergelijking
Taakeffecten	Het fenomeen dat toetstaken niet eenduidig makkelijk of moeilijk zijn voor kandidaten, bijvoorbeeld als gevolg van vertrouwdheid met het onderwerp. Taakeffecten beïnvloeden de score ongeacht de vaardigheid van de kandidaten en beïnvloeden zo de betrouwbaarheid en generaliseerbaarheid van de toets-scores.
Task-based language assessment	Toetsing binnen een communicatieve benadering van taalvaardigheid die de taak ziet als focus van toetsing. Taakuitvoering wordt zowel inhoudelijk als linguïstisch beoordeeld.
TBLA	Task-based language assessment
Terugslageffect	De invloed van een toets op de inrichting van het onderwijs
T-unit	Kortste grammaticaal aanvaardbare zin

Referenties

Abbott, M., & Lee, K. K. (2025). Lessons Learned From a High-stakes Portfolio-Based Language Assessment Initiative for LESLLA Learners. *LESLLA Symposium Proceedings, 20*(1), 35–51. https://doi.org/10.5281/zenodo.16366764

Abbott, M., & Lee, K. K. (2023). Beginning adult literacy learners, portfolio's and self-regulated learning. *LESLLA Symposium Proceedings 17 (1),* 41–56. https://doi.org/10.5281/zenodo.8075583

Alaei, M. M., Ahmadi, M., & Zadeh, N. S. (2014). The Impact of Rater's Personality Traits on Holistic and Analytic Scores: Does Genre Make any Difference too? *Procedia-Social and Behavioral Sciences, 98*, 1240–1248.

Alderson, J. C. (2009). Test review: Test of English as a foreign language™: Internet-based test (TOEFL iBT®). *Language Testing, 26*(4), 621–631. https://doi.org/10.1177/0265532209346371

Alikulova, S. A. (2020). Alikulova, S. A. (2020). A COMPARISON OF COMPUTER ASSISTED AND FACE-TO-FACE SPEAKING ASSESSMENT: PERFORMANCE, PERCEPTIONS, ANXIETY, AND COMPUTER ATTITUDES. *Экономика и Социум, 11*(78), 61–64.

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics, 40*, 134–142. https://doi.org/10.1017/S0267190520000033

Ardin, P., & Hendryanti, R. (2023). (2023, May). Investigating the Speaking Examiners' Challenges in a Synchronous Online Test. *Proceedings of the 20th AsiaTEFL-68th TEFLIN-5th INELTAL Conference (ASIATEFL 2022), 749*, 22.

Artunç, E. K., & Hart, D. O. (2020). Interactional competence in paired speaking tests: A study on proficiency-based pairings. *System, 89*.

Aziz, M., Hassan, A., & Aljamili, O. (2024). The effect of paper reading versus screen reading on the inferential reading performance among university ESL learners. *Jurnal Arbitrer, 11*(1), 85–95.

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671–704.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453–476.

Banerjee, J., Lestari, S. B., & Rossi, O. (2021). Choosing test format and task types. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing.* Routledge/Taylor & Francis.

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing, 26*, 5–19. https://doi.org/10.3389/fpsyg.2020.00330

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes.* University of Toronto.

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing, 27*(4), 515–535.

Benton, T., & Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking? *Research Matters, 26*, 22–28.

Birjandi, P., & Bagherkazemi, M. (2011). From Face-to-Face to Paired Oral Proficiency Interviews: The Nut Is yet to Be Cracked. *English Language Teaching, 4*(2), 169–175.

Bouazza, K. (2023). *The Effect of Reading Medium on EFL Students' Reading Comprehension Reading on Paper versus Reading on Screen: A Case of Master Two Students at Mascara University.* Mascara University.

Bramer, W. M., Rethlefsen, M. L., Kleijnen, J., & Franco, O. H. (2017). Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Systematic Reviews 2017 6:1*, *6*(1), 1–12. https://doi.org/10.1186/S13643-017-0644-Y

Bramley, T. (2015). *Investigating the reliability of Adaptive Comparative Judgment. Cambridge Assessment Research Report.*

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2011). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing, 29*(1), 91–108. https://doi.org/10.1177/0265532211411078

Brunfaut, T. (2023). Future challenges and opportunities in language testing and assessment: Basic questions and principles at the forefront. *Language Testing, 40*(1), 15–23. https://doi.org/10.1177/02655322221127896

Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics, 28*(1), 147–164. https://doi.org/10.1111/ijal.12196

Bulté, B., Housen, A., & Pallotti, G. (2025). Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning, 75*(2), 533–574.

Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics, 11*(2).

Chapelle, C. A. (2021). Validity in Language Assessment. In: Winke, P. & Brunfaut, T. (2021) T. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing* (1st ed., pp. 11–201). Routledge/ Taylor & Francis.

Chapelle, C. A., & Lee, H.-W. (2022). *Conceptions of validity. IThe Routledge Handbook of Language Testing.* (G. Fulcher & L. Harding (eds.); 2nd ed., pp. 17–31). Routledge/Taylor & Francis2.

Cho, Y., Rijmen, F., & Novák, J. (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBTTM integrated writing tasks. *Language Testing, 30*(4), 513–534.

Chou, M. H. (2013). Teacher interpretation of test scores and feedback to students in EFL classrooms: A comparison of two rating methods. *Higher Education Studies, 3*(2), 86–95.

Chuang, Y. Y. (2009). Foreign language speaking assessment: Taiwanese college English teachers' scoring performance in the holistic and analytic rating methods. *Asian EFL Journal, 11*(1), 150–173.

Chun, L., & Yunnan, X. (2018). A study on rater reliability under holistic and analytic scoring of CEPT Writing by using generalizability theory and Many-facet Rasch model. *In 4th International Conference on Economics, Management, Law and Education (EMLE 2018),* 1013–1019.

Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading, 42*(2), 288–325. https://doi.org/10.1111/1467-9817.12269

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

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing, 8*(2), 73–83.

Cumming, A., Cho, Y., Burstein, J., Everson, P., & Kantor, R. (2021). Assessing academic writing. In X. Xi & J. M. Norris (Eds.), *Assessing Academic English for Higher Education Admissions.* (pp. 107–151). Routledge/Taylor & Francis.

Deygers, B. (2017). *Assessing high-stakes assumptions. A longitudinal mixed-methods study of university entrance language tests, and of the policy that relies on them.* KU Leuven.

Deygers, B., Van den Branden, K., & Van Gorp, K. (2018). University entrance language tests: A matter of justice. Language Testing. *Language Testing, 35*(4), 449–476.

Deygers, B., Van Gorp, K., Joos, S., & Luyten, L. (2011). Rating scale design: a comparative study of two analytic rating scales in a task-based test. *Exploring Language Frameworks. Proceedings of the ALTE Kraków Conference.*, 271–287.

East, M., & Young, D. (2007). Scoring L2 writing samples: Exploring the relative effectiveness of two different diagnostic methods. *New Zealand Studies in Applied Linguistics, 13*(1), 1–21.

Effatpanah, F., Baghaei, P., & Boori, A. A. (2019). Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia, 9*(1), 12. https://doi.org/10.1186/s40468-019-0090-y

Elliott, L. J., Ljubijanac, M., & Wieczorek, D. (2019). The effect of screen size on reading speed: A comparison of three screens to print. *International Conference on Applied Human Factors and Ergonomics*, 103–109.

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: a narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology, 11*(330).

Fischer, J., Chouissa, C., Dugovičová, S., & Virkkunen-Fullenwider, A. (2011). *Guidelines for task-based university language testing.* Council of Europe Publishing.

Frost, K., Clothier, J., Huisman, A., & Wigglesworth, G. (2020). Responding to a TOEFL iBT integrated speaking task: Mapping task demands and test takers' use of stimulus content. *Language Testing, 37*(1), 133–155.

Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing, 29*(3), 345–369.

Fulcher, G. (2000). The 'communicative' legacy in language testing. *System, 28*(4), 483–497.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, 26*(1), 123–144. https://doi.org/10.1177/0265532208097339

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance Decision Trees. *Language Testing, 28*(1), 5–29. https://doi.org/0.1177/0265532209359514

Fulcher, G., & Harding, L. (2022). *The Routledge Handbok of Language Testing* (G. Fulcher & L. Harding (eds.); 1st ed.). Routledge/Taylor & Francis.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. *Computer-Based Assessment of Foreign Language Speaking Skills, 29*(51).

Galti, A. M., Saidu, S., Yusuf, H., & Goni, A. A. (2018). Rating scale in writing assessment: Holistic vs. Analytical scales: A review. *International Journal of English Research, 4*(6), 4–6.

Garcia Laborda, J., & Fernandez Alvarez, M. (2021). Multilevel language tests: Walking into the land of the unexplored. *Language Learning & Technology, 25*(2), 1–25.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*(4), 507–531. https://doi.org/10.1177/0265532209340188

Ghafournia, N. (2015). Standard assessments: Merits and demerits and the alternative assessments. *Asian Social Science, 11*(13), 166–173. https://doi.org/10.5539/ass.v11n13p166

Green, A. (2006). Watching for Washback: Observing the Influence of the International English Language Testing System Academic Writing Test in the Classroom. *Language Assessment Quarterly, 3*(4), 333–368. https://doi.org/10.1080/15434300701333152

Hamid, M. O., & Hoang, N. T. (2018). Humanising Language Testing. *The Electronic Journal for English as a Second Language, 22*(1), 1–20.

Han, C. (2021). Analytic rubric scoring versus comparative judgment: A comparison of two approaches to assessing spoken-language interpreting. *Meta, 66*(2), 337–361.

Harsch, C., & Malone, M. E. (2021). Language proficiency frameworks and scales. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing.* (pp. 33–44). Routledge/Taylor & Francis.

Heyworth, F. (2006). The Common European Framework. *ELT Journal, 60*(2), 181–183.

Hidri, S. (2018). Discrete point and integrative testing. In *The TESOL encyclopedia of English language teaching.* John Wiley & Sons. https://doi.org/doi. org/10.1002/9781118784235. eelt0375.

Hirai, A., & Koizumi, R. (2008). Validation of the EBB scale: A case of the story retelling speaking test. *JLTA Journal Kiyo, 11*, 1–20.

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY Access. *Language Assessment Quarterly, 13*(4), 359–376.

Hoo, S. H. (2016). Self- and peer-assessment of speaking. *SALT, 16*(1). https://doi.org/10.7916/salt.v16i2.1257

Huang, H. T. D., & Hung, S. T. A. (2018). Investigating the strategic behaviors in integrated speaking assessment. *System, 78*, 201–212.

Huang, H. T. D., Hung, S. T. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing, 35*(1), 27–49.

Huang, J., & Han, T. (2013). Holistic or analytic?–An EFL institutional writing assessment dilemma for policy makers. *Leadership Policy Quarterly, 2*(1), 1–18.

Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing, 29*(2), 203–221. https://doi.org/10.1177/0265532211419826

Ildikó, C. (2002). Is testing speaking in pairs disadvantageous for students? A quantitative study of partner effects on oral test scores. *NOVELTY-BRITISH COUNCIL HUNGARY, 9*(1), 22–45.

Isaacs, T. (2016). Assessing speaking. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 131–146). De Gruyter Mouton.

Isler, C., & Aydın, B. (2021). Developing and validating a computerized oral proficiency test of English as a foreign language (Coptefl). *International Journal of Assessment Tools in Education, 8*(1), 38–66.

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly, 3*(2), 151–169. https://doi.org/10.1207/s15434311laq0302_4

Iwashita, N., May, L., & Moore, J. P. (2021). Operationalising interactional competence in computer-mediated speaking tests. In M. R. Salaberry & A. R. Burch (Eds.), *Assessing speaking in context: expanding the construct and its applications.* (pp. 283–302). Multilingual Matters.

Izadpanah, M. A., Rakhshandehroo, F., Hoomanfard, H. M., & Mahmoudikia, M. (2014). On the Consensus Between Holistic Rating System and Analytic Rating System: A Comparison between TOEFL IBT and Jacobs et. al. Composition Profile. *International Journal of Language Learning and Applied Linguistic World, 6*(1), 170–187.

Jenkins, J., & Leung, C. (2019). From mythical 'standard' to standard reality: The need for alternatives to standardized English language tests. *Language Teaching, 52*(1), 86–110.

Jeong, Y. J., & Gweon, G. (2021). Advantages of print reading over screen reading: A comparison of visual patterns, reading performance, and reading attitudes across paper, computers, and tablets. *International Journal of Human–Computer Interaction, 37*(17), 1674–1684.

Jin, Y. (2023). Test-taker insights for language assessment policies and practices. *Language Testing, 40*(1), 193–203.

Joo, M. J. (2008). *Korean university students attitudes to and performance on a face-to-face interview (FTFI) and a computer administered oral test (CAOT).* Institute of Education, University of London.

Kane, M. T. (2022). Articulating a validity argument. In G. Fulcher & L. Harding (Eds.), *The Routlegde Handbook of Language Testing* (pp. 32–47). Routledge/Taylor & Francis.

Karim, S., & Haq, N. (2014). An assessment of IELTS speaking test. *International Journal of Evaluation and Research in Education, 3*(3), 152–157.

Kenyon, D., & Malone, M. (2010). Investigating examinee autonomy in a computerized test of oral proficiency. *Computer-Based Assessment of Foreign Language Speaking Skills*, 1–27.

Khadawardi, H. (2021). Transfer of print-based academic reading strategies to on-screen reading. *International Journal of English Language Education, 9*(1), 60–105.

Khamboonruang, A. (2022). Building an initial validity argument for binary and analytic rating scales for an EFL classroom writing assessment: Evidence from Many-Facets Rasch Measurement. *REFLections, 29*(3), 675–699.

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified Model. *Language Testing, 28*(4), 509–541. https://doi.org/10.1177/0265532211400860

Knoch, U. (2022). Assessing writing. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (pp. 236–253). Routledge/Taylor & Francis.

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing, 38*(4), 602–626. https://doi.org/10.1177/0265532221994052

Kong, Y., Seo, Y. S., & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education, 123*, 138–149.

Kühn, B., & Pérez Cavana, M. L. (2012). *Perspectives from the European Language Portfolio. Learner autonomy and self-assessment.* Routledge/Taylor & Francis.

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *International Review of Applied Linguistics and Language Teaching., 45*(3), 261–284. https://doi.org/10.1515/IRAL.2007.012

Kuiken, F., & Vedder, I. (2011). Task complexity and linguistic performance in L2 writing and speaking: The effect of mode. In *Second language task complexity* (pp. 91–104). John Benjamins Publishing Company.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31*(3), 329–348. https://doi.org/10.1177/0265532214526174

Kuiken, F., & Vedder, I. (2016). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing, 34*(4). https://doi.org/10.1177/0265532216663991

Kuiken, F., & Vedder, I. (2019). Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. *International Journal of Applied Linguistics, 29*(2), 192–210. https://doi.org/10.1111/ijal.12256

Kuiken, F., & Vedder, I. (2021). Scoring approaches. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing* (pp. 125–134). Routledge/Taylor & Francis.

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing, 33*(3), 319–340.

Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*(2), 131–166.

Lewkowicz, J. A. (2000). Authenticity in language testing: some outstanding questions. *Language Testing, 17*(1), 43–64.

Li, J. (2014). The role of reading and writing in summarization as an integrated task. *Language Testing in Asia, 4*(3). https://doi.org/10.1186/2229-0443-4-3

Li, S., Wen, T., Li, X., Feng, Y., & Lin, C. (2023). Comparing holistic and analytic marking methods in assessing speech act production in L2 Chinese. *Language Testing, 40*(2), 249–275.

Lizunova, I. V., van der Weel, A., Garcia-Febo, L., & Pshenichnaya, E. V. (2022). Reading on paper and screens: advantages, disadvantages, and digital divide. *Annals of Library and Information Studies (ALIS), 69*(1), 34–43.

Lowe, J., & Yu, X. (2009). Computer assisted testing of spoken English: A study of the SFLEP college English oral test system in China. *Journal of Systemic, Cybernetics and Informatics, 7*(3), 33–38.

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing, 34*(4), 493–511. https://doi.org/10.1177/0265532217710675

Lukácsi, Z. (2021). Developing a level-specific checklist for assessing EFL writing. *Language Testing, 38*(1), 86–105. https://doi.org/10.1177/0265532220916703

Luxia, Q. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142–173. https://doi.org/10.1191/0265532205lt300oa

Lynch, B. (2001). The ethical potential of alternative language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty. Essays in honour of Alan Davies* (pp. 228–239).

Ma, W. (2022). What the analytic versus holistic scoring of international teaching assistants can reveal: Lexical grammar matters. *Language Testing, 39*(2), 239–264. https://doi.org/10.1177/02655322211040020

Macaro, E. & Cohen, A. D. (2007). An introduction to systematic reviewing. In A. D. Cohen & E. Macaro (Eds.), *Language Learner Strategies. Thrity years of research and practice.* (pp. 163–164). Oxford University Press.

McKay, T. H., & Plonsky, L. (2021). Reliability analyses. Estimating error. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of second language acquisition and language testing* (pp. 468–482). Routledge/Taylor & Francis.

McLeod, M. (2021). Investigating test purpose pluralism and test retrofitting in high-stakes language proficiency testing. *Canadian Journal for New Scholars in Education, 12*(1), 85–93.

Mendhakar, A., Sierak, K., Chana, K., & Leder, H. (2025). Does the reading platform matter? An eye-tracking study. *AILA Review, 38*(1), 105–128.

Mora, J. C., Mora-Plaza, I., & Bermejo Miranda, G. (2024). (2024). Speaking anxiety and task complexity effects on second language speech. *International Journal of Applied Linguistics, 34*(1), 292–315. https://doi.org/10.1111/ijal.12494

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 International Results in Reading.* https://doi.org/10.58150/PIRLS_2016_data.

Nakatsuhara, F., Khabbazbashi, N., & Inoue, C. (2022). Assessing speaking. In G. Fulchner & L. Harding (Eds.), *The Routledge Handbook of Language Testing.* Routledge.

Norris, J. M. (2012). Purfoseful language assessment: selecting the right alternative test. *English Teahing Forum, 38*(3).

O'Grady, S. (2019). The impact of pre-task planning on speaking test performance for English-medium university admission. *Language Testing, 36*(4), 505–526. https://doi.org/10.1177/0265532219826604

O'Grady, S. (2025). Task design and rater effects in task-based language assessment. *TESOL Journal, 16*(1).

Ocal, T., Durgunoglu, A., & Twite, L. (2022). Reading from screen vs reading from paper: does it really matter?. *Journal of College Reading and Learning, 52*(2), 130–148.

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2014). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing, 32*(1), 39–62. https://doi.org/10.1177/0265532214538014

Onal, A. (2022). Oral Proficiency Interview (OPI) for the Assessment of Speaking. *The Eurasia Proceedings of Educational and Social Sciences, 28*, 11–18.

Ono, M., Yamanishi, H., & Hijikata, Y. (2019). Holistic and analytic assessments of the TOEFL iBT® Integrated Writing Task. *JLTA Journal, 22*, 65–88. https://doi.org/10.20622/jltajournal.22.0_65

Pallotti, G. (2021). Measuring Complexity, Accuracy, and Fluency. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing* (pp. 201–211). Routledge/Taylor & Francis.

Pallotti, G. (2022). Holistic and analytic assessment of functional adequacy. *TASK, 2*(1), 85–114. https://doi.org/10.1075/task.21014.pal

Paquot, M., Rubin, R., & Vandeweerd, N. (2022). Crowdsourced adaptive comparative judgment: A community-based solution for proficiency rating. *Language Learning, 72*(3), 853–885.

Park, H., & Yan, X. (2019). An investigation into rater performance with a holistic scale and a binary, analytic scale on an ESL writing placement test. *Papers in Language Testing and Assessment, 8*(2), 34–64.

Park, J. (2006). The study on the rater reliability of three scoring methods in assessing argumentative essays: Holistic, analytic, and multiple-trait scoring methods. *Foreign Language Education Research, 9*, 63–84.

Peras, I., Klemenčič Mirazchiyski, E., Japelj Pavešić, B., & Mekiš Recek, Ž. (2023). Digital versus paper reading: A systematic literature review on contemporary gaps according to gender, socioeconomic status, and rurality. *European Journal of Investigation in Health, Psychology and Education, 13*(10), 1986–2005.

Pill, J., & Smart, C. (2021). Raters. Behavior and training. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing.* (pp. 135–144). Routledge/Taylor & Francis.

Plakans, L. (2022). Writing integrated tasks. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (2nd ed.). Routledge/Taylor & Francis.

Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing, 36*(2), 161–179.

Plonsky, L. (2023). Sampling and generalizability in Lx research: A second-order synthesis. *Languages, 8*(1), 1–13. https://doi.org/10.3390/languages8010075

Poe, M., & Elliot, N. (2019). Evidence of fairness: Twenty-five years of research in assessing writing. *Assessing Writing, 42.*

Purpura, J. E. (2009). The impact of large-scale and classroom-based language assessment on the individual. In L. Taylor & C. J. Weir (Eds.), *Language Testing Matters. Investigating the wider social and educational impact of assessment.* (pp. 301–325). ALTE.

Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. Language Testing. *Language Testing, 21*(1), 28–52. https://doi.org/10.1191/0265532204lt273oa

Quaid, E., & Barrett, A. (2021). Interpretations of spoken utterance fluency in simulated and face-to-face oral proficiency interviews. *Language Education & Assessment, 4*(1), 1–18.

Rajendram, S., Larson, E., & Sinclair, J. (2019). International graduate students' perspectives on high-stakes English tests and the language demands of higher education. *Language and Literacy, 21*(4), 68–92. https://doi.org/10.20360/langandlit29428

Read, J. (2022). Test review: The international English Language Testing System (IELTS). *Language Testing, 39*(4), 679–694.

Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics, 37*(6), 828–848.

Révész, A., Kourtali, N.-E., & Mazgutova, D. (2017). Effects of Task Complexity on L2 Writing Behaviors and Linguistic Complexity. *Language Learning. 67*(1), 208–241. https://doi.org/10.1111/lang.12205

Rezazadeh, M., Tavakoli, M., & Rasekh, A. E. (2011). The Role of Task Type in Foreign Language Written Production: Focusing on Fluency, Complexity, and Accuracy. *International Education Studies, 4*(2), 169–176. https://doi.org/10.5539/ies.v4n2p169

Richardson, J. T. (2022). *The legibility of serif and sans serif typefaces: Reading from paper and reading from screens.* Springer Nature.

Riemenschneider, A., Weiss, Z., Schröter, P., & Meurers, D. (2024). The Interplay of Task Characteristics, Linguistic Complexity, and Language Proficiency in High-Stakes English as a Foreign Language Writing. *TESOL Quarterly, 58*(2), 775–801. https://doi.org/10.1002%2Ftesq.3254

Rifiyani, A. J., Widowati, D. R., Ni'mah, D., & Ubaidillah, M. F. (2022). To read on-screen or printed texts? A quantitative study into EFL students' reading achievement. *International Journal of Education and Learning, 4*(3), 191–201.

Roever, C., & Ikeda, N. (2022). What scores from monologic speaking tests can (not) tell us about interactional competence. *Language Testing, 39*(1), 7–29.

Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing, 35*(3), 331–355.

Romova, Z., & Andrew, M. (2011). Teaching and assessing academic writing via the portfolio: Benefits for learners of English as an additional language. *Assessing Writing, 16*(2), 111–122. https://doi.org/10.1016/j.asw.2011.02.005

Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing, 37*(1), 31–53.

Runge, A., Goodwin, S., Attali, Y., Poe, M., Mulcaire, P., Lo, K. L., & LaFlair, G. T. (2025). A multi-stage interactive writing task for the assessment of English language writing proficiency. *Language Testing, 42*(4), 423–446.

Safari, F., & Ahmadi, A. (2023). Developing and evaluating an empirically-based diagnostic checklist for assessing second language integrated writing. *Journal of Second Language Writing, 60.*

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment.* (pp. 129–152). University of Orlando.

Sakyi, A. A. (2003). *A study of the holistic scoring behaviours of experienced and novice ESL instructors.* University of Toronto.

Salaberry, M. R., & Burch, A. R. (2021). *Assessing speaking in context: expanding the construct and its applications* (M. R. Salaberry & A. R. Burch (eds.)). Multilingual Matters.

Sardari, S., Ahmadi, A., & Ravand, H. (2025). Creating an empirically derived checklist to diagnose second language learners' integrated writing ability. *Language Testing in Asia, 15*(1), 15–52.

Sawaki, Y. (2022). Computer-based testing. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (pp. 530–544). Routledge/Taylor & Francis.

Sayin, B. A. (2015). (2015). Exploring anxiety in speaking exams and how it affects students' performance. *International Journal of Education and Social Science, 2*(12), 112–118.

Schildt, L., Deygers, B., & Weideman, A. (2024). Language testers and their place in the policy web. *Language Testing, 41*(2), 338–356. https://doi.org/10.1177/02655322231191133

Schmitt, N., Ng, J. W. C., & Garras, J. (2011). The word associates format: Validation evidence. *Language Testing, 28*(1), 105–126.

Sheybani, E., & Zeraatpishe, M. (2018). On the dimensionality of reading comprehension tests composed of text comprehension items and cloze test items. *International Journal of Language Testing, 8*(1), 12–26.

Shi, X., Ma, X., Du, W., & Gao, X. (2023). Diagnosing Chinese EFL learners' writing ability using polytomous cognitive diagnostic models. *Language Testing, 41*(1), 109–134. https://doi.org/10.1177/02655322231162840

Shin, S. Y., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores? *Language Testing, 32*(15), 259–281.

Shirazi, M. A. (2013). Using an analytic dichotomous evaluation checklist to increase inter-and intra-rater reliability of EFL writing evaluation. *Iranian Journal of Applied Linguistics (IJAL), 16*(1), 25–57.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests.* Routledge.

Shohamy, E. (2006). *Language policy. Hidden agendas and new approaches.* Taylor & Francis.

Sickinger, R., Brunfaut, T., & Pill, J. (2025). Comparative Judgement for evaluating young learners' EFL writing performances: Reliability and teacher perceptions of holistic and dimension-based judgements. *Language Testing, 42*(2), 137–166. https://doi.org/10.1080/0969594X.2019.1700212

Sims, M. E., Cox, T. L., Eckstein, G. T., Hartshorn, K. J., Wilcox, M. P., & Hart, J. M. (2020). Rubric rating with MFRM versus randomly distributed comparative judgment: A comparison of two approaches to second-language writing assessment. *Educational Measurement: Issues and Practice, 39*(4), 30–40. https://doi.org/10.1111/emip.12329

Suvorov, R., & Hegelheimer, V. (2014). Computer-assisted language testing. In A. J. Kunnan (Ed.), *The companion to language assessment.* John Wiley & Sons. https://doi.org/10.1002/9781118411360.wbcla083

Tabari, M. A., & Wang, Y. (2022). Assessing linguistic complexity features in L2 writing: Understanding effects of topic familiarity and strategic planning within the realm of task readiness. *Assessing Writing, 52.* https://doi.org/10.1016/j.asw.2022.100605

Thwaites, P., Vandeweerd, N., & Paquot, M. (2025). Crowdsourced comparative judgement for evaluating learner texts: How reliable are judges recruited from an online crowdsourcing platform? *Applied Linguistics, 46*(4), 611–628.

Timpe-Laughlin, V., Sydorenko, T., & Dombi, J. (2024). Human versus machine: Investigating L2 learner output in face-to-face versus fully automated role-plays. *Computer Assisted Language Learning, 37*(1–2), 149–178.

Tüchler, A. F. (2024). (2024). *Reading Print vs. on Screen: Exploring Reading Comprehension, Information Location Recall, and Reading Habits Across Media.* Iniversity of Linz.

Turner, C. E., & Upshur, J. A. (1995). Some Effects of Task Type on the Relation between Communicative Effectiveness and Grammatical Accuracy in Intensive ESL Classes. *TESL Canada Journal, 12*(2), 18–31.

Van den Branden, K. (2021). Measuring task-based performance. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing.* (pp. 316–325). Routledge/Taylor & Francis.

Van der Zanden, M., & Dirkx, M. (2024). Een breed erkende startkwalificatie. Het staatsexamen Nt2 in heden, verleden en toekomst. *Les Online, 42.*

Van Gorp, K., & Deygers, B. (2013). (2013). Task-based language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 578–593). Wiley.

Vasylets, O., Gilabert, R., & Manchón, R. M. (2017). The effects of mode and task complexity on second language production. *Language Learning, 67*(2), 394–430. https://doi.org/10.1111/lang.12228

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliaility of coparative judgement. *Assessment in Education, 26*(5), 541–562. https://doi.org/10.1080/0969594X.2019.1602027

Wagner, E. (2022). Assessing listening. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (pp. 223–235). Routledge/Taylor & Francis.

Wagner, E. (2020). (2020). Duolingo English test, revised version July 2019. *Language Assessment Quarterly, 17*(3), 300–315. https://doi.org/10.1080/15434303.2020.1771343

Walsh, G. (2016). Screen and paper reading research–A literature review. *Australian Academic & Research Libraries, 47*(3), 160–173.

Weaver, C. (2006). Evaluating the use of rating scales in a high-stakes Japanese university entrance examination. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 4.*

Wei, W., & Zheng, Y. (2017). An investigation of integrative and independent listening test tasks in a computerised academic English test. *Computer Assisted Language Learning, 30*(8), 864–883.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281–300. https://doi.org/10.1191/0265532205lt309oa

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice, 27*(1), 46–64. https://doi.org/10.1080/0969594X.2019.1700212

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing, 35*(2), 161–192. https://doi.org/10.1177/0265532216686999

Winke, P., & Brunfaut, T. (2021). *The Routledge Handbook of Second Language Acquisition and Language Testing.* (P. Winke & T. Brunfaut (eds.)). Routledge/Taylor & Francis.

Wiseman, C. S. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *International Journal of Language Testing, 2*(1), 59–92.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing, 24*(2), 251–286.

Xi, X., Norris, J. M., Ockey, G. J., Fulcher, G., & Purpura, J. E. (2021). Assessing academic speaking. In X. Xi & J. M. Norris (Eds.), *Assessing Academic English for Hiher Education Admissions.* Routledge.

Yan, X., Kim, H. R., & Kim, J. Y. (2020). Dimensionality of speech fluency: Examining the relationships among complexity, accuracy, and fluency (CAF) features of speaking performances on the Aptis test. *Language Testing, 38*(4), 485–510. https://doi.org/10.1177/0265532220951508

Yang, H. C., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly, 46*(1), 80–103. https://doi.org/10.1002/tesq.6

Yokouchi, Y. (2025). Revisiting the effectiveness of a performance decision tree-style rubric compared to a grid-style rubric. *Language Testing in Asia, 8*. https://doi.org/10.1186/s40468-024-00338-5

Yoon, H.-J., & Burton, D. (2021). Measuring L2 writing. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing.* (pp. 295–304). Routledge/Taylor & Francis.

Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? *Language Testing, 25*(4), 521–551.

Yu, J., Zhou, X., Yang, X., & Hu, J. (2022). Mobile-assisted or paper-based? The influence of the reading medium on the reading comprehension of English as a foreign language. *Computer Assisted Language Learning, 35*(1–2), 217–245.

Yujia, Z. H. O. U. (2012). *Test-takers' affective reactions to a computer-delivered speaking test and their test performance.* retrieved 01/08/2012 from http://cblle. tufs. ac. jp/assets/files/publications/working_papers_09/section/295-310. pdf.

Zhang, D., & Koda, K. (2017). Assessing L2 vocabulary depth with word associates format tests: issues, findings, and suggestions. *Asian-Pacific Journal of Second and Foreign Language Education, 2*(1).

Zhu, X., Li, X., Yu, G., Cheong, C. M., & Liao, X. (2016). Exploring the relationships between independent listening and listening-reading- writing tasks in Chinese language testing: Toward a better understanding of the construct underlying integrated writing tasks. *Language Assessment Quarterly, 13*(3), 167–185. https://doi.org/10.1080/15434303.2016.1210609

손영아. (2013). *The Interlocutor Proficiency Effect on Test-taker Performance in a Paired Oral Assessment.* 서울대학교 대학원.

Dit onderzoeksrapport is een uitgave van het CvTE.
Het onderzoek is uitgevoerd door Kaatje Dalderop
Onderwijsadvies