Rapport

Blended toetsen: Het beste van twee werelden?

Blended toetsen: Het beste van twee werelden?

Eva Poort, Kimberley Lek, Sebastiaan de Klerk (CitoLab, Stichting Cito)
Paul van der Molen (Centrale Toetsen en Examens, Stichting Cito)

Cito | Arnhem

Inhoud

Samenvatting

Het onderwijs digitaliseert in rap tempo. Ook toetsen en examens worden steeds vaker digitaal afgenomen, zoals bijvoorbeeld de centrale examens (CE's) voor vmbo bb/kb. Voor de andere onderwijsrichtingen wordt nu ook onderzocht of de examens verder gedigitaliseerd kunnen worden. Het is echter de vraag of geheel digitale examens bij alle vakken goed passen. Daarom hebben we in dit rapport onderzocht hoe het concept van blended toetsen – toetsen die deels op papier en deels digitaal vormgegeven worden – in te zetten is voor de centrale examens. Dit hebben we gedaan op basis van literatuuronderzoek en interviews met toetsdeskundigen van de divisie Centrale Toetsen en Examens (divisie CTE, onderdeel van Stichting Cito) en (oud) docenten. In het literatuuronderzoek gingen we in op de vraag wanneer digitaal toetsen en toetsen op papier mogelijk nadelig is voor leerlingen. In de interviews focusten we op hoe blended toetsen nu al in de praktijk ingezet wordt en hoe in de toekomst de centrale examens blended vormgegeven zouden kunnen worden.

Uit het literatuuronderzoek bleek dat leerlingen vaak maar niet altijd beter presteren wanneer zij een toets op papier maken dan wanneer ze die digitaal maken. Voor de talen en maatschappijvakken bleek dit verschil iets groter en consistenter te zijn dan voor de exacte vakken, waar iets vaker geen verschillen worden gevonden. Er zijn verscheidene oorzaken die deze verschillen in leerlingprestaties kunnen verklaren; één daarvan is dat mensen informatieve teksten over het algemeen beter lijken te begrijpen wanneer ze die op papier lezen. Ook hier zijn in de literatuur verschillende mogelijke oorzaken voor te vinden. Misschien lukt het lezers minder goed om een 'mentaal landschap' te creëren van een digitale tekst, of lezen ze die oppervlakkiger omdat ze dat gewend zijn van sociale media of overschatten ze hun eigen begrip van de tekst meer als ze die digitaal hebben gelezen, waardoor ze minder geneigd zijn om de tekst nog eens goed door te lezen. Tegelijkertijd toonde het literatuuronderzoek ook aan dat het typen van examens om verschillende redenen de voorkeur heeft boven schrijven, behalve voor de exacte vakken wiskunde, scheikunde en natuurkunde. Getypte antwoorden bleken bijvoorbeeld vaak van even goede of betere kwaliteit te zijn als handgeschreven antwoorden, tenzij leerlingen bijvoorbeeld een wiskundige uitwerking moesten geven. Ook vinden studenten typen vaak fijner en beoordelen docenten liever getypte dan handgeschreven antwoorden. Het literatuuronderzoek geeft dus al een aantal redenen om de examens voor sommige vakken niet geheel te digitaliseren.

Uit de interviews bleek dat voor de meeste vakken van de vier vakgebieden de we behandeld hebben – de exacte vakken, talen, maatschappijvakken en kunstvakken – een vorm van een blended examen in te denken en zelfs aan te raden is. Hier zit tussen de vakken wel veel variatie in. Eén en dezelfde oplossing zou dus niet voor alle vakken even goed werken. Voor de kunstvakken zouden de examens bijvoorbeeld bijna geheel gedigitaliseerd kunnen worden. Papier is voor deze vakken op het moment vooral nog van toegevoegde waarde voor vakken zoals muziek waarbij leerlingen bijvoorbeeld noten op een notenbalk moeten zetten. Voor de talen en maatschappijvakken zou verdere digitalisering ook van toegevoegde waarde zijn: zowel docenten als leerlingen zouden baat hebben bij getypte examens. Voor deze examens zou, afhankelijk van het vak, papier alleen nog het beste medium zijn voor bronnen (vooral leesteksten) en uitwerkbijlages, in elk geval totdat leerlingen gewend zijn aan digitaal lezen. Voor deze drie vakgebieden lijkt het bovendien niet onaannemelijk dat blended examens alleen tijdens een transitieperiode van toegevoegde waarde zijn en dat deze examens uiteindelijk volledig gedigitaliseerd worden. Alleen de exacte vakken wiskunde, scheikunde en natuurkunde vormen hier een uitzondering op. Voor deze vakken werkt papier nu – en waarschijnlijk in de toekomst – nog het beste volgens de toetsdeskundigen én (oud)docenten. Het ligt daarom voor de hand om deze examens nu en in de toekomst nog geheel op papier af te nemen.

Definitie

De definitie van *blended toetsen* die we aanhouden in dit onderzoek is als volgt: in een blended toets worden sommige onderdelen van de toets op papier aangeboden en andere onderdelen op een device. Onder onderdelen van de toets verstaan we bijvoorbeeld de bronnen, de vraag zelf, het antwoordvel, enzovoorts. We beperken ons wat betreft devices tot computers (laptops en eventueel desktops) en gaan er verder vanuit dat een digitaal toetsplatform (waarvan Facet een voorbeeld is) gebruikt wordt om de onderdelen van de toets die via de computer aangeboden worden, aan te bieden. We laten voorlopig in het midden of het toetsplatform een internet-verbinding vereist of niet.

1 Aanleiding

De afgelopen jaren is de digitalisering van het onderwijs in een stroomversnelling terecht gekomen. Ook toetsen worden steeds vaker gedigitaliseerd, waaronder de centrale examens (CE's) voor vmbo bb/kb. Digitaal toetsen biedt veel voordelen en kansen, zoals nieuwe vraagvormen, flexibele afnamemomenten en betere voorzieningen voor leerlingen met extra ondersteuningsbehoeften. Er gaan daarom ook steeds meer stemmen op voor het verder digitaliseren van de CE's voor de overige onderwijsrichtingen (vmbo gl/tl, havo en vwo). Toch is het belangrijk om voorzichtig te blijven, gegeven het feit dat (wetenschappelijk) onderzoek uit de afgelopen decennia het beeld schetst dat toetsen op papier en toetsen op een computer niet altijd zonder meer vergelijkbaar zijn (Backes & Cowan, 2019; Gordanier et al., 2023; D. Li et al., 2017; Lynch, 2022; Steedle et al., 2020). Bovendien zijn sommige typen toetsopgaven die nu vaak voorkomen in de centrale eindexamens moeilijk te digitaliseren, zoals wiskundeopgaven waarbij leerlingen hun uitwerking moeten tonen.

De vraag rijst dus of het de beste keus is om op termijn alle CE's (volledig) digitaal af te nemen. Over de digitalisering van toetsen en examens wordt vaak gedacht in zwart en wit: een toets wordt afgenomen op papier óf digitaal. Er zijn echter ook grijstinten mogelijk, waarbij sommige onderdelen van de toets of het examen op papier worden aangeboden en andere digitaal. Denk aan een examen Nederlands waarbij leerlingen leesteksten op papier krijgen en digitaal antwoord geven. Dit soort 'blended toetsen' zou voor meerdere vakken een uitkomst kunnen bieden en daarmee (in elk geval tijdens de transitie) van toegevoegde waarde kunnen zijn. In dit onderzoek verkennen wij daarom de mogelijkheden en voordelen van blended toetsen voor de centrale examens.

2 Opzet en onderzoeksvragen

Dit rapport bestaat uit een literatuuronderzoek en een praktijkverkenning door middel van interviews met toetsdeskundigen van de divisie Centrale Toetsen en Examens (divisie CTE, onderdeel van Stichting Cito) en (oud)docenten die in de praktijk al een vorm van blended toetsen inzetten. De overkoepelende onderzoeksvraag waar we een antwoord op proberen te geven is als volgt:

Welke onderdelen van het centrale schriftelijke eindexamen zouden op papier aangeboden moeten worden en welke digitaal?

Deze vraag proberen we voor vier 'vakgebieden' (de exacte vakken, talen, maatschappijvakken en kunstvakken) te beantwoorden. Daarbij gaan we dus ook in op de vraag of blended toetsen voor elk vak(gebied) geschikt dan wel aan te raden is en of blended toetsen voor elk vak(gebied) op dezelfde manier ingezet zou moeten worden. Door het literatuuronderzoek en de interviews met elkaar te combineren komen we tot een mogelijk toekomstbeeld van hoe het concept van blended toetsen ingezet zou kunnen worden voor de CE's.

3 Literatuuronderzoek: Wanneer is digitaal toetsen mogelijk nadelig voor leerlingen en wanneer juist voordelig?

Er is voor zover bij ons bekend geen wetenschappelijke onderzoeksliteratuur over blended toetsen. Er bestaan wel blended toetsen zoals het centraal schriftelijk en praktisch examen (cspe) waarin opdrachten voorkomen waar de leerling zowel met papier als met de computer moet werken en de compex-examens die tussen 2003 en 2010 zijn afgenomen. Maar er lijkt nog niet eerder een systematische wetenschappelijke studie gedaan te zijn naar de meerwaarde van die blended vorm (ten opzichte van geheel op papier of geheel digitaal). In het literatuuronderzoek gaan we daarom in op de volgende onderzoeksvraag:

Wanneer is toetsen op papier of digitaal mogelijk nadelig voor leerlingen en wanneer juist voordelig?

Hierbij kijken we naar wat uit (wetenschappelijk) onderzoek bekend is over de verschillen in leerlingprestaties (1) tussen toetsen op papier en digitaal, (2) tussen lezen op papier en van een (computer)scherm en (3) tussen schrijven met de hand en typen. Ons doel is om zo een indicatie te krijgen van welke onderdelen van het examen het beste op papier aangeboden zouden kunnen worden en welke digitaal. We leggen de focus op de prestaties van leerlingen, omdat hier binnen de drie thema's het meeste onderzoek naar is gedaan. We laten verschillen in betrouwbaarheid, validiteit en dergelijke dus achterwege. De onderzoeken die we bespreken komen grotendeels uit wetenschappelijke tijdschriften en rapporten van internationale toetsinstituten en toetsautoriteiten; ook behandelen we een aantal onderzoeken uit conferentiebijdragen en proefschriften.

3.1 Verschillen tussen toetsen op papier en digitaal

In het eerste deel van het literatuuronderzoek kijken we dus naar de verschillen in prestaties tussen het afnemen van toetsen en examens op papier (in de Engelstalige literatuur *paper-based testing*, afgekort PBT) en op de computer (*computer-based testing*, CBT). We maken een onderscheid tussen toetsen voor de talen en maatschappijvakken, die vaak met lange tekstbronnen werken en vaak ook langere antwoorden vereisen, en toetsen voor de exacte vakken, waar de bronnen vaker grafisch zijn en de antwoorden regelmatig speciale symbolen en notaties vereisen.

Zoals gezegd worden de centrale examens voor vmbo bb/kb al digitaal afgenomen. Voor zover bij ons bekend zijn er geen onderzoeken gedaan naar de verschillen tussen die digitale afnames en de papieren afnames daarvoor; daarom gaat al het onderzoek dat we hier bespreken over buitenlandse toetsen en examens.

3.1.1 Toetsen voor de taal- en maatschappelijke vakken

Wat betreft toetsen voor de taal- en maatschappijvakken lijkt uit onderzoek een klein voordeel voor PBT naar voren te komen, maar er is veel verdeeldheid. Er zijn zowel onderzoeken – en een meta-analyse – die concluderen dat er geen verschillen zijn tussen PBT en CBT

(o.a. D.-H. Kim & Huynh, 2008; Toroujeni, 2022; Wang et al., 2008; W. Yu & Iwashita, 2021) als onderzoeken die aantonen dat papieren toetsen beter gemaakt worden (o.a. Backes & Cowan, 2019; Gordanier et al., 2023; Gu et al., 2021; Jeong, 2014; Liu et al., 2016; Proctor et al., 2019; Støle et al., 2020) en onderzoeken die het omgekeerde concluderen (o.a. D. Li et al., 2017; Proctor et al., 2019; Steedle et al., 2020).

De meest recente meta-analyse die we hebben kunnen vinden, bijvoorbeeld, concludeert dat er geen verschil is tussen digitale en papieren leestoetsen. Wang et al. (2008) namen in hun meta-analyse 11 artikelen met in totaal 36 experimenten op waarin de onderzoekers een vergelijking maakten tussen de papieren en computerversie van een leestoets. (De auteurs geven helaas niet meer informatie over wat voor leestoetsen het om gaat.) Op basis van deze 36 experimenten berekenden Wang et al. (2008) een klein maar niet significant gestandaardiseerd effect (d_w) van 0,04 ten gunste van papieren toetsen. Deze inmiddels al wat oude meta-analyse suggereert dus dat het voor een leestoets niet uitmaakt of een kandidaat die op papier of op de computer maakt.

Meer recentere onderzoeken laten juist vaak zien dat een papieren toets beter is voor het toetsen van de talen en maatschappijvakken dan een digitale toets (Backes & Cowan, 2019; Gordanier et al., 2023; Gu et al., 2021; Liu et al., 2016; Proctor et al., 2019; Støle et al., 2020). Backes en Cowan (2019), bijvoorbeeld, analyseerden de resultaten van de *Partnership for Assessment of Readiness for College and Careers* (PARCC) toets. Deze toets toetst verschillende vaardigheden en vakgebieden, waaronder ook taalvaardigheid (*English Language Arts*, ELA). In 2015 en 2016 werd die toets in de Amerikaanse staat Massachusetts zowel op papier als digitaal aangeboden. Bijna 400.000 leerlingen in *grades* 3 t/m 8 (groep 5 t/m 8 en de eerste twee jaren van de middelbare school) maakten toen óf de digitale óf de papieren variant. Backes en Cowan (2019) vonden dat leerlingen die de toets digitaal maakten significant lager scoorden dan leerlingen die de toets op papier maakten. Voor het taalonderdeel ging het om een verschil van 0,24 standaard deviaties (SDs). Hoewel dit verschil volgens de richtlijnen van Cohen (1988) klein is, heeft het volgens de auteurs in de praktijk weldegelijk grote gevolgen: op basis van de methode van Lipsey et al. (2012) berekenen zij dat een verschil van 0,24 SDs gelijk staat aan een leeropbrengst van tussen de 4,2 en 11,0 maanden scholing (afhankelijk van in welk leerjaar de leerling zit en gebaseerd op een schooljaar van 9 maanden). Wel moet bij de resultaten van Backes en Cowan (2019) een kanttekening geplaatst worden: de digitale variant van de PARCC-toets was niet één-op-één identiek aan de papieren variant.[1] Hierdoor kan het verschil in scores tussen de papieren en digitale variant groter of kleiner uitgevallen zijn dan wanneer beide varianten volledig identiek waren geweest.

De papieren variant en computervariant van de toets die Gordanier et al. (2023) analyseerden waren daarentegen nagenoeg identiek. Zij analyseerden de scores van een soortgelijke gestandaardiseerde toets als de PARCC die afgenomen wordt in de Amerikaanse staat South Carolina. Het enige verschil tussen de papieren en digitale variant daarvan was dat in de hogere leerjaren de digitale toets een aantal *technology-enhanced items* bevatte (met bijvoorbeeld een digitale rekenmachine in plaats van een fysieke rekenmachine). In 2015, 2016 en 2017 werd ook die toets zowel op papier als digitaal aangeboden. Ook weer bijna 400.000 leerlingen in *grades* 3 t/m 8 maakten toen de toets, die onder andere een taalonderdeel (*English Language Arts*, ELA) en een onderdeel *social studies* bevatte (vergelijkbaar met maatschappijleer/-kunde/-wetenschappen in Nederland). De digitale variant werd voor alle onderdelen van de toets

1 De papieren en digitale variant verschilde op de volgende punten van elkaar: de itemset, de presentatie van bronteksten en de manier waarop antwoord gegeven moest worden zowel op gesloten vragen (digitaal kon bijvoorbeeld ook gesleept worden) en open vragen (digitaal was een spellingscheck beschikbaar).

significant slechter gemaakt dan de papieren variant van de toets. Voor het taalonderdeel bleek het om een verschil van 0,09 SDs te gaan. Hoewel de auteurs het zelf niet doen, kunnen we op basis van de methode van Lipsey et al. (2012) weer uitrekenen dat het hier zou gaan om een leeropbrengst van tussen de 2,0 en 3,4 maanden. Omdat de papieren en digitale varianten identiek waren, is dit waarschijnlijk een betere indicatie van de 'ware' grootte van het verschil in toetsscores tussen papieren en digitale taaltoetsen dan de 0,24 SDs waar Backes en Cowan (2019) op uitkwamen. Voor *social studies* hing het verschil samen met het leerjaar: voor leerlingen van de basisschool was het verschil met 0,002 SDs niet significant, terwijl het verschil van 0,03 SDs (een leeropbrengst van ongeveer 1 maand) voor leerlingen uit de eerste twee jaren van de middelbare school wel significant was. Dit suggereert dat het effect van digitaal toetsen ook groter kan worden naarmate de materie complexer wordt.

Bovendien vonden Gordanier et al. (2023) dat leerlingen uit arme huishoudens de digitale variant nog slechter maakten dan de rest van de leerlingen. Voor leerlingen uit arme huishoudens was het effect van CBT voor het taalonderdeel drie keer zo groot als voor de groep andere leerlingen (0,15 SDs vergeleken met 0,05 SDs). Voor *social studies* bleek het zelfs zo te zijn dat alleen leerlingen uit arme gezinnen een nadeel ondervonden van CBT van 0,05 SDs. Dit verschil tussen leerlingen uit arme huishoudens en de rest van de leerlingen kan volgens de auteurs deels verholpen worden door scholen betere digitale voorzieningen te geven. Op scholen met betere digitale voorzieningen was het verschil tussen CBT en PBT voor het taalonderdeel (en de onderdelen wiskunde en natuurwetenschappen, die we later in detail bespreken) namelijk kleiner. Echter, voor leerlingen uit arme huishoudens bleek dat niet genoeg te zijn om het gat tussen CBT en PBT volledig te dichten, zeker voor leerlingen die thuis geen beschikking hadden tot de nieuwste digitale devices. De auteurs concluderen dat digitaal toetsen het al bestaande gat tussen kinderen uit arme huishoudens en rijkere huishoudens dus kan vergroten.

Onderzoek van Li et al. (2017) en Steedle et al. (2020) toont echter aan dat de ACT-toets over het algemeen beter digitaal gemaakt wordt dan op papier. Ook de ACT-toets bestaat uit meerdere onderdelen, waaronder een onderdeel begrijpend lezen (*Reading*) en een onderdeel over taalverzorging (*English*), die we hier bespreken. In het onderzoek van Li et al. (2017) worden twee studies beschreven, één uit 2014 waaraan meer dan 5.500 leerlingen meededen en één uit 2015 met meer dan 3.000 leerlingen. Uit de analyse bleek dat de online versie in beide studies voor alle onderdelen beter werd gemaakt dan de papieren versie, maar dit was vooral het geval voor de leestoets. Daar was het gestandaardiseerde verschil tussen de online en papieren versie 0,33 SDs in de eerste studie en 0,17 SDs in de tweede. Voor het onderdeel over taalverzorging waren de gestandaardiseerde verschillen 0,15 SDs en 0,17 SDs. Ook bleek dat leerlingen die de toets online deden minder vragen niet beantwoordden dan op papier.

Dit laatste kwam ook uit de analyse van Steedle et al. (2020). Sinds de studie van Li et al. (2017) is de ACT-toets van toetsplatform gewisseld. Steedle en zijn collega's hebben daarom opnieuw onderzocht of de scores van de online en papieren versie vergelijkbaar zijn. Daarvoor hebben ze drie studies uitgevoerd met meer dan 3.500 leerlingen (in oktober 2019), meer dan 6.000 leerlingen (december 2019) en meer dan 6.500 leerlingen (februari 2020). Ook uit deze analyses bleek dat de leestoets en taalverzorgingstoets online beter gemaakt werden dan op papier, met gestandaardiseerde effecten van tussen de 0,16 SDs en 0,22 SDs voor begrijpend lezen en 0,10 SDs en 0,13 SDs voor taalverzorging. ook bleek dus dat leerlingen die de toets online deden minder vragen niet beantwoordden dan op papier, vooral vragen die aan het einde van leestoets en taalverzorgingstoets zaten. Het lijkt er dus op dat de digitale versie van de toets minder tijd kostte om in te vullen, waardoor leerlingen meer van de vragen konden beantwoorden.

Ten slotte is er nog één groot onderzoek dat het vermelden waard is, namelijk dat van Proctor en collega's (Proctor et al., 2019). Zij bekeken in drie studies de resultaten op de *SAT Suite of*

Assessments van meer dan 5.000 leerlingen die de SAT deden, meer dan 2.000 leerlingen die de PSAT 10 deden en meer dan 3.500 leerlingen die de PSAT 8/9 deden. Ze vonden dat er geen belangrijke verschillen waren wat betreft de scores tussen de digitale en papieren versie, behalve voor de toets begrijpend lezen. Dat onderdeel werd in de digitale versie beter gemaakt met een effectgrootte van tussen de 0,13 SDs en 0,18 SDs. Volgens de auteurs ligt de oorzaak van dit verschil bij de zogenoemde *Command of Evidence*-items. Dat zijn items waarbij leerlingen uit de gegeven opties moesten kiezen welk deel van de tekst het beste bewijs leverde voor hun antwoord op de vorige vraag (vergelijkbaar met citeervragen in de centrale examens). De auteurs gaan helaas niet dieper in op waarom deze items makkelijker waren in de digitale versie. Echter, bij de voorbeelden die ze in de bijlage hebben opgenomen valt op dat, hoewel in beide versies antwoord gegeven wordt door één van de vier opties bij een meerkeuzevraag te selecteren, in de digitale versie de antwoordopties ook in het tekstfragment in geel gemarkeerd zijn en in de papieren versie niet. Hierdoor was het waarschijnlijk makkelijker of in elk geval sneller om het juiste antwoord te vinden.

We sluiten dit onderdeel af met een onderzoek van Gu et al. (2021), een wat kleiner maar wel interessant onderzoek. Zij bekeken de resultaten van 565 universiteitsstudenten op de *HEIghten CT*-toets, een toets die het kritisch denkvermogen meet. Kandidaten krijgen vragen over maatschappelijke en ethische vraagstukken, geschiedenis en vragen waarbij ze vanuit verschillende culturele oogpunten dienen te denken. Ook wordt hun numerieke geletterdheid en begrijpend lezen getoetst en moeten ze een schrijfopdracht maken. Gu et al. (2021) vonden dat de digitale variant van deze toets beduidend moeilijker was en meer tijd kostte dan de papieren variant waardoor kandidaten de digitale variant vaker niet af konden maken. Hier lagen volgens de onderzoekers verschillende redenen aan ten grondslag. Ten eerste bevat de toets veel lange tekstbronnen die in de digitale versie niet op één scherm paste, waardoor er gescrold moest worden – iets waar we later op terugkomen. Bovendien was het digitaal niet mogelijk om delen van de tekst te markeren. Ten tweede werden vragen die bij dezelfde tekst hoorden op papier op één pagina getoond, maar digitaal op verschillende schermen gepresenteerd. Welke van deze mogelijke verklaringen ook daadwerkelijk een rol speelde, konden de onderzoekers echter niet met zekerheid zeggen.

3.1.2 Toetsen voor de exacte vakken

Toetsen voor de taal- en maatschappijvakken lijken dus vaak beter gemaakt te worden op papier dan digitaal. In 2015 hebben collega's van CTE in opdracht van het College voor Toetsen en Examens een literatuurstudie gedaan waarin ze keken naar de verschillen in prestaties op papieren en digitale rekenen-wiskundetoetsen (Kuhlemeier et al., 2015). In dat onderzoek keken ze specifiek naar de effecten van (a) de afnamemodus (op papier of op de computer), (b) het toekennen van *partial credit* aan antwoorden, (c), de mogelijkheid tot *item review*, (d) het gebruik van kladpapier en (e) de invoer van antwoorden op de computer versus op papier. Hierbij merken ze op dat vooral veel onderzoek gedaan is naar het eerste punt, wat voor ons onderzoek ook van grootste interesse is. Op dat gebied komen ze tot de conclusie dat er weinig aanleiding is om te veronderstellen dat het effect van de afnamemodus op de prestaties van leerlingen groot is: de gemiddelde effectgrootte is klein. Hier voegen ze aan toe dat als er bij rekenen-wiskundetoetsen een significant verschil wordt gevonden tussen de papieren en digitale variant, dat dat dan wel vaker ten gunste van PBT is dan CBT. Ook merken ze op het effect sterk af lijkt te hangen van het itemtype: moduseffecten zijn kleiner voor meerkeuzevragen dan andere itemtypen.

Onderzoek dat sinds 2015 is gedaan lijkt vaker aan te wijzen dat toetsen voor de exacte vakken beter op papier gemaakt worden, maar zoals Kuhlemeier et al. (2015) ook opmerken is er nog steeds veel verdeeldheid. In het eerder genoemde onderzoek van Backes en Cowan (2019) is ook

het verschil tussen de papieren en digitale versie van het wiskundeonderdeel van de PARCC-toets berekend. Dit kwam uit op een verschil van 0,10 standaard deviaties (SDs). Dit is dus kleiner dan het verschil voor het taalonderdeel, maar vertegenwoordigt nog steeds een verschil in leeropbrengst van tussen de 1,3 en 5,3 maanden volgens de methode van Lipsey et al. (2012). In een eerdere vergelijking tussen de digitale en papieren versie van de PARCC-toets vonden Liu et al. (2016) daarnaast dat ook het geometrie-onderdeel van deze toets beter op papier gemaakt werd.

Ook Gordanier et al. (2023) concludeerden dat het wiskundeonderdeel van de toets die zij analyseerden beter op papier gemaakt werd dan op de computer. Voor wiskunde bleek net als voor *social science* het effect samen te hangen met het leerjaar: het verschil voor leerlingen op de basisschool was slechts 0,003 SDs en niet significant, maar voor leerlingen op de middelbare school was met 0,04 SDs wel significant. Volgens de methode van Lipsey et al. (2012) vertegenwoordigt dat een leeropbrengst van tussen de 1,1 en 1,6 maanden. Hierbij moet weer vermeld worden dat de papieren en digitale variant van de toets die Gordanier et al. (2023) analyseerden identiek waren, terwijl dit voor Backes en Cowan (2019) niet het geval was. Daarom ligt het verschil van Gordanier et al. (2023) waarschijnlijk dichter bij de 'ware' grootte van het verschil tussen papieren en digitale toetsen.

Voor het onderdeel natuurwetenschappen kwamen Gordanier et al. (2023) uit op een verschil van 0,04 SDs (een leeropbrengst van tussen de 1,1 en 1,6 maanden). Hoewel het verschil tussen basisschoolleerlingen en middelbareschoolleerlingen niet significant was, was het effect van digitaal toetsen wel groter voor leerlingen op de middelbare school (0,03 SDs vergeleken met bijna 0,05 SDs). Zeker voor wiskunde en mogelijk ook voor natuurwetenschappen lijkt digitaal toetsen dus nadeliger te worden naarmate de materie ingewikkelder wordt. Ook bleek voor zowel wiskunde als natuurwetenschappen dat het effect van digitaal toetsen niet kleiner werd in het tweede of derde jaar dat de toets (ook) digitaal afgenomen werd. Ten slotte toonde de analyse aan dat leerlingen uit arme huishoudens een groter nadeel ondervonden van digitaal toetsen (voor wiskunde: 0,07 SDs; voor natuurwetenschappen: 0,07 SDs), dan leerlingen uit rijkere huishoudens (voor wiskunde: geen verschil; voor natuurwetenschappen: 0,02 SDs). Voor wiskunde was het dus net als voor *social studies* het geval dat alleen leerlingen uit armere huishoudens de digitale toets slechter maakten dan de papieren toets. Ook voor de exacte vakken kan digitaal toetsen dus het al bestaande gat tussen kinderen uit arme huishoudens en rijkere huishoudens vergroten.

Li et al. (2017) en Steedle et al. (2020) hebben net als Backes en Cowan (2019) en Gordanier et al. (2023) ook gekeken naar de onderdelen wiskunde en natuurwetenschappen van hun toets, de ACT-toets. Li et al. (2017) hadden zoals eerder besproken twee studies uitgevoerd. Alleen uit de eerste studie uit 2014 bleek dat leerlingen de digitale variant van het onderdeel natuurwetenschappen significant beter maakten dan de papieren variant, met een verschil van 0,19 SDs. In de tweede studie uit 2015 was dit niet het geval. Het wiskundeonderdeel werd in beide jaren even goed gemaakt op papier als digitaal. De analyse van Steedle et al. (2020) toont hetzelfde aan: het wiskundeonderdeel werd alleen in de tweede van hun drie studies beter digitaal gemaakt, maar het verschil was met 0,05 SDs niet heel groot. Het onderdeel natuurwetenschappen werd in de eerste en derde van de drie studies beter digitaal gemaakt, met een verschil van 0,07 en 0,12 SDs. Verder bleek ook weer dat leerlingen die de toets online deden minder vragen niet beantwoordden dan op papier. Dit bleek ook uit het onderzoek van Proctor et al. (2019) voor de open wiskundevragen voor de *SAT Suite of Assessments* (voor de gesloten vragen was er geen verschil in percentages leerlingen die de vraag wel/niet beantwoordden). Proctor et al. (2019) vonden bovendien ook dat er geen significante verschillen waren tussen de digitale en papieren variant van de *SAT Suite of Assessments* voor wiskunde.

3.1.3 Belangrijke factoren en mogelijke verklaringen

De auteurs van deze studies gaan vaak zelf niet dieper in op waarom de papieren of digitale variant van de toets die zij geanalyseerd hebben beter gemaakt zou worden. Als we echter zelf naar de verschillende toetsen kijken, dan valt bijvoorbeeld op dat de ACT en SAT alleen uit meerkeuzevragen bestaan, terwijl in de PARCC ook soms open vragen voorkomen, zowel in het onderdeel *English language arts* als in het wiskunde-onderdeel. Bovendien zijn de teksten in de PARCC tussen de 600 en meer dan 2000 worden lang en daarmee over het algemeen langer dan de teksten uit de ACT en SAT, die tussen de 750 en 1000 woorden lang zijn. Ten slotte wordt in het wiskunde-onderdeel van de PARCC vaker gewerkt met figuren, tabellen en andere grafische elementen als bronnen en vereisen sommige antwoorden uitwerkingen met speciale symbolen of moeten er grafieken getekend worden. In de ACT en SAT daarentegen zijn de contexten korter en vaker (in het geval van de ACT zelfs meestal) in de vorm van een tekstuele uitleg. Er zitten dus redelijk grote verschillen tussen de PARCC aan de ene kant en de ACT en SAT die zouden kunnen verklaren waarom de PARCC beter gemaakt worden op papier en de ACT en SAT vaak beter digitaal. Voor de talen en maatschappijvakken zou een verklarende factor de lengte van de teksten kunnen zijn. Voor de exacte vakken zou de verklaring kunnen liggen in de uitwerking van de antwoorden – leerlingen zijn gewend speciale symbolen en notaties te schrijven in plaats van te typen – en manipulatie van de bronnen – op papier kun je makkelijker een figuur draaien, dubbelvouwen of hulplijnen tekenen dan digitaal.

Een andere belangrijke factor in deze kwestie is die van gewenning. Uit het onderzoek van Backes en Cowan (2019) bleek dat het verschil tussen CBT en PBT voor taalvaardigheid in het tweede jaar na invoering van de digitale variant ongeveer half zo klein was als in het eerste jaar. Voor wiskunde bleek dat het verschil in het tweede jaar ook kleiner was, met ongeveer een derde. Leerlingen die dus voor de tweede keer de digitale variant van de PARCC maakten, waren daar al wat meer aan gewend, maar presteerden nog steeds slechter dan leerlingen die de toets op papier maakten. Gordanier et al. (2023) vonden hier echter geen bewijs voor. Toch sluiten ook zij niet uit dat wennen aan digitale toetsen meer tijd nodig heeft. En hoewel Li et al. (2017) en Steedle et al. (2020) er geen analyse aan besteden, blijkt uit bestudering van hun effectgroottes ook dat de effecten in het tweede of derde jaar ook over het algemeen kleiner zijn dan in het eerste jaar nadat de digitale variant van de ACT beschikbaar kwam. Het lijkt er dus op dat leerlingen kunnen wennen aan het digitaal maken van toetsen; mogelijk speelt hoe digitaal vaardig ze zijn daar ook een rol in.

3.1.4 Samenvatting

Al met al lijkt het er dus op dat leerlingen om verschillende mogelijke redenen beter presteren wanneer zij een toets op papier maken dan wanneer ze die digitaal maken. Voor de talen en maatschappijvakken lijkt dit verschil iets groter en consistenter te zijn dan voor de exacte vakken, waar iets vaker geen verschillen worden gevonden. Bovendien suggereert de studie van Gordanier et al. (2023) dat leerlingen uit armere huishoudens een groter nadeel ondervinden van de switch naar digitaal toetsen dan hun klasgenoten. De al bestaande prestatiekloof tussen deze twee groepen kan dus mogelijk nog verder vergroot worden door over te stappen op digitale eindexamens. Hier moet echter wel een kanttekening bij geplaatst worden: Backes en Cowan (2019) lieten zien dat het verschil tussen PBT en CBT in het tweede jaar na invoering van digitaal toetsen kleiner was dan in het eerste jaar. Er is dus mogelijk sprake van een gewennings-periode die in elk geval langer lijkt te zijn dan één jaar.

Voor de centrale eindexamens in de Nederlandse context is voor zover bij ons bekend niet onderzocht of de overstap op digitale afname voor vmbo bb/kb nadelig of voordelig voor de leerlingen heeft uitgepakt. Op basis van het onderzoek dat we hier besproken hebben, valt dit

echter niet uit te sluiten. Het is daarom aan te raden om hier statistisch onderzoek naar te doen, als mogelijk, eventueel aangevuld met kwalitatief onderzoek door leerlingen en docenten uitgebreid te interviewen.

3.1.5 Wat we nog niet weten

We hebben mogelijke verklaringen besproken voor de verschillen in leerlingprestaties tussen digitale en papieren toetsen. We weten echter niet welke hiervan inderdaad belangrijk zijn en waarom. Daarnaast weten we ook niet hoe lang de gewenningsperiode zou duren, als er inderdaad zo'n periode is. Backes en Cowan (2019) stellen dat leerlingen meer dan een jaar nodig hebben om te wennen, maar waar de bovengrens ligt is onduidelijk. Ook is niet bekend waar leerlingen precies aan moeten wennen en of er groepen leerlingen zijn die een langere gewenningsperiode nodig hebben of aan bepaalde aspecten van digitaal toetsen meer moeten wennen dan andere leerlingen. Het onderzoek van Gordanier et al. (2023) liet al zien dat leerlingen uit arme huishoudens meer benadeeld worden door de digitalisering van toetsen dan leerlingen uit rijkere huishoudens, dus dit valt niet uit te sluiten. Of ze op termijn toetsen uiteindelijk even goed op papier als digitaal zullen (kunnen) maken, is echter op basis van deze onderzoeken ook nog niet duidelijk.

3.2 Verschillen tussen lezen op papier en van een scherm

Door de digitalisering van onze samenleving komt het steeds vaker voor dat we van een scherm lezen in plaats van op papier. Ook in de klas en met de digitalisering van de examens steeds vaker in toetscontexten. Al sinds de jaren '90 doen onderzoekers onderzoek naar de mogelijke gevolgen van lezen van een scherm op tekstbegrip, leessnelheid en andere factoren. En natuurlijk naar de mogelijke verklaringen hiervoor. In dit onderdeel van het literatuuronderzoek gaan we hier dieper op in.

3.2.1 De samenhang tussen het leesmedium en tekstbegrip en leessnelheid

De studies op dit onderzoeksgebied laten ook geen eenduidig beeld zien: sommige onderzoeken vonden een verschil in tekstbegrip en/of leessnelheid ten gunste van lezen van een scherm, andere ten gunste van lezen van papier en weer anderen vonden geen verschil. Tot voor kort was het dus onduidelijk of er verschillen zijn tussen lezen van een scherm en lezen van papier. Uit drie recente meta-analyses blijkt namelijk dat lezen van papier over het algemeen tot een beter tekstbegrip leidt dan lezen van een scherm (Clinton, 2019; Delgado et al., 2018; Kong et al., 2018). Delgado et al. (2018) laten bijvoorbeeld zien dat het gestandaardiseerde effect (Hedge's *g*) op basis van 56 vergelijkingen uit 38 studies 0,21 ten gunste van papier is.[2] Op basis van een net andere selectie van 29 onderzoeken met in totaal 33 vergelijkingen vond Clinton (2019) een gestandaardiseerde effect van 0,25 en Kong et al. (2018) vonden ook weer een

2 Delgado et al. (2018) hebben *between-participants* en *within-participants* onderzoeken apart geanalyseerd. In *between-participants* onderzoeken leest één groep deelnemers van een scherm en een andere van papier. In *within-participants* onderzoeken leest elke deelnemer één tekst van een scherm en één van papier. Het gemiddelde effect wordt in deze twee soorten onderzoeken op een andere manier berekend: in *between-participants* onderzoeken berekent men Hedge's *g* en in *within-participants* onderzoeken de *mean change index*, d_c. Uit Delgado et al.'s (2018) analyse bleek het effect van leesmedium van dezelfde grootte te zijn voor beide typen onderzoek (Hedge's g = -0,21 en d_c = -0,21). In de analyse van de *within-participants* onderzoeken hadden ze 18 vergelijkingen uit 16 studies met meegenomen.

gemiddeld effect van 0,21 op basis van 17 studies met 47 vergelijkingen. Deze drie meta-analyses laten dus hetzelfde beeld zien: als het aankomt op tekstbegrip, is lezen van papier gemiddeld genomen beter dan lezen van een scherm.

De effecten die deze drie meta-analyses rapporteren – g = -0,21, g = -0,25 en g = -0,21 – zijn volgens Cohen's (1988) richtlijnen kleine effecten. Delgado et al. (2018) plaatsen het echter weer in context: de jaarlijkse groei in leesvaardigheid op de basisschool in Groot-Brittannië ligt tussen g = 0,08 (in groep 8) en g = 0,55 (in groep 3) (Luyten et al., 2017) en het gemiddelde effect van leesinterventies is geschat op g = 0,45 (Scammacca et al., 2015). Het effect van leesmedium op tekstbegrip is volgens Delgado en collega's dus vergelijkbaar in grootte met ongeveer twee derde van de jaarlijkse groei in leesvaardigheid op de basisschool en ongeveer half zo groot als het effect dat interventies gemiddeld hebben op het verbeteren van de leesvaardigheid van leerlingen. In de dagelijkse onderwijs- en toetspraktijk zijn dit dus wel degelijk effecten van belang.

Wat betreft de effecten van leesmedium op leessnelheid zijn de resultaten minder eenduidig. Zowel Clinton (2019) als Kong et al. (2018) vonden dat het leesmedium geen significant effect heeft op de snelheid waarmee lezers lazen. Toch lazen lezers in Clinton's (2019) meta-analyse gemiddeld genomen iets sneller van papier, hoewel het gemiddelde verschil klein was (g = 0,08 op basis van 14 vergelijkingen uit 13 studies). Zij concludeert op basis hiervan dat dit zou kunnen betekenen dat lezers efficiënter lezen van papier dan van een scherm. Hun tekstbegrip is immers beter, terwijl ze de tekst ongeveer net zo snel gelezen hebben als wanneer ze die van een scherm zouden hebben gelezen. De resultaten van Kong et al. (2018) tonen echter dat we voorzichtig moeten zijn met deze suggestie. In hun studie was het effect van leesmedium op leessnelheid ook niet significant, maar lazen lezers juist langzamer van papier (g = -0,48 op basis van 19 vergelijkingen uit 8 studies). Op basis van hun resultaten kunnen we dus niet concluderen dat lezers efficiënter lezen van papier. Sterker nog, het zou dus juist zo kunnen zijn dat het tekstbegrip van lezers die van een scherm lezen slechter is omdat ze de tekst (te) snel lezen.

3.2.2 Belangrijke factoren en mogelijke verklaringen

Als we teruggaan naar het effect van leesmedium op tekstbegrip, valt het op dat dit effect samenhangt met een aantal factoren. Uit de analyses van Delgado et al. (2018) en Clinton (2019) bleek bijvoorbeeld dat het effect van leesmedium groter en alleen significant was voor informatieve teksten (g = 0,27 en g = 0,32, respectievelijk), en niet voor verhalende teksten (g = 0,01 en g = 0,04, respectievelijk)[3]. Bovendien bleek uit de meta-analyse van Delgado et al. (2018) dat het effect van het medium op tekstbegrip groter was wanneer teksten onder tijdsdruk gelezen werden. Het effect is dan gemiddeld 0,26, terwijl het slechts 0,09 is wanneer mensen op hun eigen tempo de teksten lezen. In de context van de centrale eindexamens, als leerlingen informatieve teksten onder tijddruk moeten lezen, kan dit dus gevolgen hebben voor hun begrip van deze teksten en daarmee hun toetsscores.

Volgens Delgado et al. (2018) was het effect van leesmedium ook alleen significant als er van een computerscherm gelezen werd (g = 0,23) en niet als er een tablet of ander *hand-held* device werd gebruikt (g = 0,12), hoewel het verschil tussen deze twee soorten devices niet significant was. Ook lijkt het erop dat het effect ongeveer twee keer zo groot is als lezers moeten scrollen om verder te lezen (g = 0,25) dan wanneer dat niet nodig is (g = 0,13). Hierbij moet echter

3 Twee meta-analyses over het effect van leesmedium specifiek op het lezen van verhalende teksten laten hetzelfde zien (Furenes et al., 2021; Schwabe et al., 2022).

opgemerkt worden dat dit verschil ook weer niet significant was én dat het een effect van tekstlengte in vermomming kan zijn. Delgado et al. (2018) hebben namelijk geen rekening gehouden met het feit dat scrollen vaker nodig is bij lange teksten dan korte teksten, hoewel ze wel vonden dat het effect van leesmedium op tekstbegrip niet verschilde tussen lange (g = 0,25) en korte teksten (g = 0,22). Hoewel niet significant, sluit dit effect wel goed aan bij wat de *cognitive map theory* genoemd wordt (Eklundh, 1992; Hou et al., 2017; Mangen et al., 2013; Piolat et al., 1997; Wästlund, 2007; Wästlund et al., 2005) Deze theorie stelt het lastiger is voor lezers om een mentaal 'landschap' van de tekst te creëren als die van een scherm gelezen wordt, vooral als in de digitale tekst gescrold moet worden. Omdat hun mentale 'landschap' van een digitale tekst slechter is, is het onder andere moeilijker om informatie (antwoorden) terug te vinden in te tekst. Mogelijk doet scrollen op die manier dus af aan hun tekstbegrip.

Ook Clinton's (2019) analyse ondersteunt deze theorie op een indirecte wijze. Zij heeft in haar analyse gekeken naar het type vraag dat over de tekst gesteld werd om het tekstbegrip van de lezer in beeld te brengen. Die heeft ze opgedeeld in vragen die gingen over de tekst zelf, zoals over details die direct in de tekst terug te vinden zijn (*literal understanding*), en vragen waarbij lezers connecties moeten maken tussen delen van de tekst of tussen de tekst en hun algemene kennis (*inferential understanding*). Voor beide soorten vragen vond zij dat lezen van papier beter was, hoewel dat voordeel iets kleiner was voor *inferential understanding* (g = -0,26) dan voor *literal understanding* (g = -0,33). Zij knoopt het zo aan elkaar: het is voor lezers waarschijnlijk lastiger om een mentaal 'landschap' van de tekst te maken als zij van een scherm lezen. Dit zou vooral het beantwoorden van vragen waarop het antwoord direct in de tekst terug te vinden is (*literal understanding*) bemoeilijken.

Eén mogelijke reden waarom lezen van een scherm nadelig voor iemands tekstbegrip is, is dus dat het mogelijk lastiger is om een mentaal 'landschap' van de tekst te creëren. Een andere reden heeft te maken met de manier waarop mensen teksten van een scherm lezen, vermoedelijk oppervlakkiger. Dit wordt ook wel de *shallowing hypothesis* genoemd (Annisette & Lafreniere, 2017; Carr, 2010). Door de komst van met name sociale media, maar ook de ontwikkelingen in traditionele media en andere vormen van communicatie komen mensen steeds meer in aanraking met grote hoeveelheden berichten. Door die overweldigende hoeveelheid berichten – en de algoritmen van de platforms waarop ze geplaatst worden – worden mensen gestimuleerd om op een oppervlakkige manier met al die berichten om te gaan. Daardoor associëren mensen lezen van een scherm in het algemeen mogelijk steeds meer met oppervlakkig lezen (Delgado et al., 2018), wat mogelijk gevolgen heeft voor hun tekstbegrip. De analyse van Delgado et al. (2018) lijkt daar ook (op indirecte wijze) bewijs voor te geven: het effect van het leesmedium op tekstbegrip is door de tijd heen licht gegroeid (met g = 0,01 per jaar). Studies met een recentere publicatiedatum vonden ietwat grotere verschillen dan studies van langer geleden. Het effect van het leesmedium lijkt dus niet kleiner te worden naarmate mensen meer ervaring opdoen met lezen van een scherm. Dit komt ook overeen met het feit dat zowel in hun analyse als die van Clinton (2019) het effect van het leesmedium niet samenhing met leeftijd: jonge mensen die vermoedelijk *digital natives* waren, waren niet beter (maar ook niet slechter) in lezen van een scherm dan oudere mensen die niet waren opgegroeid met technologie.

Ten slotte voegt Clinton (2019) nog een derde mogelijke verklaring aan het verhaal toe. Uit haar analyse bleek dat lezers hun begrip van de tekst meer overschatten als ze van een scherm lezen. Dit verschil in kalibratie, zoals dat genoemd wordt, kwam gemiddeld uit op een Hedge's g van 0,20. Alle lezers denken dus over het algemeen dat ze een tekst beter begrijpen dan eigenlijk het geval is, maar lezers die die tekst van een scherm hebben gelezen overschatten hun begrip van de tekst meer dan lezers die de tekst op papier hebben gelezen. Dit kan het effect van lees-medium ook verklaren. Als lezers hun begrip van een tekst meer overschatten wanneer ze van een scherm lezen, zijn ze misschien minder geneigd om tijdens het lezen hun leesgedrag aan te

passen. Het gevolg hiervan zou kunnen zijn dat ze de tekst minder goed lezen, waardoor hun daadwerkelijke begrip uiteindelijk ook slechter is.

3.2.3 Onderzoek sinds de drie meta-analyses

Onderzoek dat sinds de drie meta-analyses is gepubliceerd, bevestigt nogmaals dat tijdsdruk een belangrijke factor is. Delgado en Salmerón (2021) lieten universiteitsstudenten een informatieve tekst lezen van papier of een scherm en onder tijdsdruk of niet. Tijdens het lezen vulden de studenten op vaste momenten korte vragenlijsten in om te meten hoe vaak hun gedachten afdwaalden. Alleen wanneer ze onder tijdsdruk lazen, rapporteerden de studenten die van een scherm lazen vaker dat hun gedachten afdwaalden dan de studenten die van papier lazen. En de studenten die onder tijdsdruk van een scherm lazen begrepen de tekst het minst goed. De auteurs concluderen daarom dat lezen van een scherm onder tijdsdruk studenten stimuleert om onoplettender te lezen. Dit past volgens de auteurs ook weer bij de *shallowing hypothesis*. Het tekstbegrip en de mate van gedachtendwalen van de studenten die *zonder* tijdsdruk lazen hing verder niet samen met het leesmedium.

Een onderzoek van Stiegler-Balfour et al. (2023) laat verder zien dat leesvaardigheid zelf ook een belangrijke factor is. In hun experiment lazen studenten met een hoog en laag leesvaardigheidsniveau twee informatieve teksten op papier, op een computer of op een iPad. Voor studenten met een hoog leesvaardigheidsniveau maakte het leesmedium niet veel uit: hun tekstbegrip was hetzelfde ongeacht het medium, hoewel ze wel iets langzamer lazen van een computerscherm van dan papier of een iPad. De studenten met een lagere leesvaardigheid lazen echter sneller van een scherm (computer of iPad) dan van papier en hun tekstbegrip was ook slechter wanneer ze van een scherm lazen. Ook deze resultaten passen weer binnen de *shallowing hypothesis*. Bovendien toont dit onderzoek aan dat lezen van een scherm voor sommige lezers equivalent kan zijn aan lezen van papier, maar voor anderen niet.

Latini et al. (2019) onderzochten of het leesdoel nog van belang is. Zij lieten studenten twee informatieve teksten lezen. Deze teksten lazen ze óf allebei van papier, allebei van een scherm óf één op papier en één van een scherm. Verder kreeg de helft van de studenten de opdracht de teksten te lezen alsof ter voorbereiding op een examen, de andere helft alsof ze voor hun plezier lazen. De studenten beantwoordden daarna op een computer een aantal vragen over de teksten. Het tekstbegrip was voor alle studenten hetzelfde ongeacht het leesmedium en leesdoel. Vooral interessant is hun bevinding dat de studenten die beide teksten van papier lazen ter voorbereiding op een examen langere antwoorden op de vragen gaven. Studenten die één of beide tekst(en) van een scherm lazen, deden dit niet. De studenten die van papier lazen ter voorbereiding op een examen besteedden dus meer tijd aan het beantwoorden van de vragen. Ook liet de analyse zien dat deze studenten vaker allebei de teksten in hun antwoorden gebruikten. (Tijdens het beantwoorden van de vragen konden de studenten de teksten niet meer inzien.) Hoewel het voor hun tekstbegrip dus niet uitmaakte van welk medium de studenten lazen, had lezen op papier toch voordelen voor hun antwoorden op de toets. Bovendien suggereert dit onderzoek dat lezen van papier gunstiger is als lezers meer dan één tekst in hun antwoord moeten gebruiken.

De meeste onderzoeken die we tot nu toe besproken hebben, zijn gedaan met volwassenen, vaak bachelorstudenten. Er zijn beduidend minder onderzoeken gedaan met kinderen en tieners, maar die bevestigen wel dat het leesmedium ook voor hen van belang is (o.a. Halamish & Elbaz, 2020; Mangen et al., 2013; Ronconi et al., 2022; Støle et al., 2020). Ronconi et al. (2022), bijvoorbeeld, lieten 13-jarige scholieren twee korte informatieve teksten lezen, één op papier en één op een tablet. Over elke tekst beantwoordden de leerlingen drie open vragen, één over het belangrijkste idee van de tekst, één over de belangrijkste punten en één waarin ze gevraagd

werden alle overige informatie te geven die ze zich konden herinneren. Het leesmedium had geen effect op de snelheid waarmee de leerlingen lazen. Wel bleek dat leerlingen het belangrijkste idee van de tekst minder goed begrepen als ze de tekst van een scherm hadden gelezen dan op papier. De andere twee vormen van tekstbegrip werden niet beïnvloed door het leesmedium. Ten slotte vonden de onderzoekers dat leerlingen van 13 jaar oud, net als volwassenen, hun tekstbegrip overschatten als zij van een scherm lazen – vooral de jongens – en dat deze overschatting het effect van leesmedium op tekstbegrip volledig verklaarde. Dat wil zeggen, leerlingen die van een scherm lazen overschatten hun tekstbegrip meer en beantwoordden daardoor de vragen over de tekst minder goed.

3.2.4 Samenvatting

Wat betreft tekstbegrip kunnen we op basis van de besproken studies de conclusie trekken dat papier de voorkeur heeft. Hier zijn wel een aantal condities aan verbonden: papier lijkt vooral (of alleen) beter te zijn ten opzichte van computerschermen (en niet e-readers of tablets), voor informatieve teksten (en niet voor fictie) en wanneer onder tijdsdruk gelezen moet worden. Ook lijkt het verschil tussen lezen van papier en een scherm afhankelijk te zijn van of er in de digitale tekst gescrold moet worden en van het leesvaardigheidsniveau van de lezer. Het effect van het leesmedium lijkt bovendien langzaam groter te worden (met g = 0,01 per jaar). Op dit vlak lijkt dus geen sprake te zijn van een gewenningsperiode. Wat betreft leessnelheid zijn de resultaten niet eenduidig: soms lezen lezers sneller van papier, soms van een scherm en soms zijn er geen verschillen. Als lezers op papier tot een beter begrip van de tekst komen in dezelfde of kortere leestijd, dan zou dit erop kunnen wijzen dat lezen van papier ook efficiënter is dan lezen van een scherm. Omdat er weinig overeenstemming is tussen de verschillende onderzoeken, kunnen we hier echter geen stevige conclusie uit trekken.

Er zijn grofweg drie mogelijke verklaringen van dit effect te vinden in de literatuur. De eerste wordt de *cognitive map theory* genoemd en stelt dat het voor lezers moeilijker is om een mentaal landschap te creëren van een digitale tekst dan een tekst op papier, vooral als in de digitale tekst gescrold moet worden. Daardoor is het moeilijker om informatie terug te vinden in de tekst. De tweede theorie gaat ervanuit dat mensen digitale teksten oppervlakkiger lezen omdat ze de digitale wereld associëren met sociale media, clickbait, korte (kranten)artikelen en dergelijke. Doordat ze daar zo'n enorme hoeveelheid aan informatie tegenkomen die ze niet allemaal tot in de puntjes kunnen doorspitten, schieten ze bij aanzien van een digitale tekst als het ware automatisch in een modus van vluchtig lezen. De derde en laatste verklaring legt de oorzaak bij het inschattingsvermogen van de lezer. Lezers lijken hun begrip van een tekst minder goed in te kunnen schatten wanneer ze die tekst van een scherm hebben gelezen en kunnen hun leesgedrag hierdoor mogelijk niet voldoende bijsturen om tot een even goed begrip van de tekst te komen.

Voor de centrale eindexamens in de Nederlandse context betekent dit dus dat overstappen op digitale examens mogelijk nadelig zou kunnen zijn voor examens waarvoor veel gelezen moet worden, zoals de taalexamens en examens maatschappijkunde/-wetenschappen en in ietwat mindere mate de examens voor vakken zoals geschiedenis en filosofie. De *cognitive map theory* zou bovendien voorspellen dat vooral citeervragen moeilijker zouden worden voor de leerlingen, zeker als in de teksten gescrold moet worden, wat nu het geval is voor de examens vmbo bb/kb in Facet. Bovendien suggereert het onderzoek van Latini et al. (2019) dat leerlingen mogelijk meer moeite zullen hebben om meerdere teksten te combineren in één antwoord.

3.2.5 Wat we nog niet weten

Natuurlijk zijn er ook nog dingen die we niet weten. Zo is het bijvoorbeeld nog niet helemaal duidelijk of de lengte van de tekst uitmaakt. Delgado et al. (2018) vonden dat het effect van leesmedium op tekstbegrip niet verschilde tussen lange en korte teksten. Tegelijkertijd leek het effect van leesmedium wel groter te zijn voor teksten waarin gescrold moest worden dan voor teksten waarbij dat niet nodig was en dat waren vaker lange teksten dan korte teksten. Ook is het niet duidelijk of het effect van leesmedium wellicht groter is voor sommige typen vragen (bijvoorbeeld open vragen of citeervragen) dan voor andere typen (bijvoorbeeld meerkeuzevragen en andere gesloten vragen). Ten slotte weten we ook nog niet of het mogelijk is om digitale leesstrategieën aan te leren. Dat wil zeggen, kunnen we leerlingen leren om digitaal net zo goed te lezen als op papier, of gebeurt dit op termijn misschien zelfs wel vanzelf?

3.3 Verschillen tussen schrijven met de hand en typen

In dit derde en laatste onderdeel van het literatuuronderzoek kijken we naar de verschillen tussen schrijven met de hand en typen. We kijken naar verschillen in kwaliteit van de antwoorden, hoe de antwoorden door beoordelaars beoordeeld worden en of het schrijfproces zelf anders is afhankelijk van of er getypt of geschreven wordt. Ook staan we stil bij de perspectieven van leerlingen en beoordelaars.

3.3.1 Is schrijven of typen beter?

Op het eerste gezicht lijkt deze kwestie simpel te zijn. Op de vraag, 'Is het beter om een antwoord te typen of om het met de hand te schrijven?' lijkt het antwoord redelijk snel 'typen' te zijn. Zowel Stoukou et al. (2023), Shin et al. (2023) en Lynch (2022) komen op basis van een review van de literatuur tot de conclusie dat getypte essays vaak van dezelfde of betere kwaliteit zijn dan geschreven essays. Het is weliswaar waar dat een aantal vooral vroege studies uit de jaren '60 tot en met '90 (Marshall & Powers, 1969; Mazzeo & Harvey, 1988; Powers et al., 1994) en studies met oudere volwassenen (e.g. Chen et al., 2011) vonden dat geschreven essays betere scores ontvingen dan getypte essays. Echter, meer recentere onderzoeken – en ook een aantal oudere – tonen aan dat getypte essays vaak juist hogere scores ontvangen (Canz et al., 2020; Ceka & O'Geen, 2019; Goldberg et al., 2003; Jin & Yan, 2017; Russell & Haney, 1997; Russell & Plati, 2001; Stoukou et al., 2023) of dat getypte en geschreven essays dezelfde scores krijgen (Hillier & Lyon, 2019; Horkay et al., 2006; MacCann et al., 2002), ook wanneer studenten/leerlingen zelf mogen kiezen of ze willen typen of schrijven (Mogey et al., 2010; Mogey & Hartley, 2013).

De kwestie of typen of schrijven beter is, is echter niet zo simpel als het lijkt. De score die een essay krijgt wordt grofweg door twee dingen bepaald: de kwaliteit van de essay zelf en de beoordelaar die de essay nakijkt. Niet alle van de genoemde onderzoeken hebben de kwaliteit en beoordeling los van elkaar onderzocht. Drie recente studies illustreren dit probleem goed. Barkaoui en Knouzi (2018), Brunfaut et al. (2018) en Kim et al. (2018) analyseerden essays geschreven door internationale studenten als onderdeel van drie verschillende toetsen Engelse taalvaardigheid (de online versie van de TOEFL, de *Integrated Skills in English*-toets van Trinity College London en de *English Placement Test* van de University of Illinois at Urbana-Champaign). Hoewel de getypte essays vaak langer, complexer, samenhangender en inhoudelijk volledige waren dan de geschreven essays, vonden geen van deze drie studies verschillen tussen de scores die handgeschreven essays en getypte essays kregen. Dit wekt de suggestie dat de beoordelaars bevooroordeeld waren ten gunste van de geschreven essays of ten nadele van de getypte essays.

3.3.2 Hoe een antwoord beoordeeld wordt, hangt niet alleen af van de kwaliteit van dat antwoord

Ook de studie van Canz et al. (2020) benadrukt dat de score van een essay niet alleen afhangt van de kwaliteit van de essay, maar ook de beoordelaar. Canz et al. (2020) selecteerden een set van 430 korte essays geschreven door leerlingen uit de onderbouw van het Duitse voortgezet onderwijs. Die handgeschreven essays transcribeerden ze stuk voor stuk, zodat van elke essay ook een getypte variant beschikbaar was. Vervolgens lieten ze een team van getrainde beoordelaars beide versies beoordelen. De 'getypte' essays – die dus identiek waren aan de handgeschreven originelen – scoorden toch hoger dan de handgeschreven originelen zowel op inhoud als op stijl. Dit toont aan dat ook deze beoordelaars, ondanks de intensieve training die ze volgens Canz et al. (2020) gedaan hadden, toch bevooroordeeld waren, in dit geval ten gunste van getypte essays.

Onderzoek van Mogey et al. (2010) bevestigt nogmaals dat beoordelaars getypte en geschreven essays soms anders beschouwen. Zij lieten 38 studenten een essay schrijven of typen (studenten mochten zelf kiezen) en zetten vervolgens deze essays om naar de andere vorm, zodat ook weer van elke essay een geschreven en getypte variant beschikbaar was. Vijf beoordelaars beoordeelden vervolgens ieder één variant van elke essay. Een analyse op basis van alle essays (originelen en transcripties) liet zien dat de handgeschreven essays beter beoordeeld werden dan de getypte essays. Echter, een analyse op basis van alleen de originelen liet zien dat de origineel getypte essays beter werden beoordeeld dan de origineel handgeschreven essays. De beoordelaars in deze studie waren dus licht bevooroordeeld ten gunste van de geschreven essays.

Uit een ouder onderzoek van Yu et al. (2004) bleek ook al hetzelfde: geschreven essays kregen hogere scores dan getypte essays. Zij opperen als verklaring hiervoor dat beoordelaars mogelijk minder streng zijn bij het beoordelen van handgeschreven essays, omdat ze getypte essays beschouwen als meer 'af' dan handgeschreven essays en dus de lat onbewust hoger leggen. Ook Rankin (2015) vond dat beoordelaars sommige aspecten van een essay anders beschouwen afhankelijk van of de tekst getypt of handgeschreven is. Verkeerd gespelde woorden interpreteerden ze bijvoorbeeld vaker als typo's (en dus niet als echte fouten) wanneer de essay getypt was, maar als spelfouten wanneer de tekst geschreven was (terwijl ook in dit onderzoek de geschreven en getypte essays identieke kopieën waren van elkaar).

We kunnen dus concluderen dat – ook als de essays in feite identiek zijn – beoordelaars getypte en handgeschreven antwoorden niet altijd hetzelfde beoordelen. Welke kant dit effect op gaat lijkt sterk van de beoordelaars zelf af te hangen. Soms zijn beoordelaars bevooroordeeld ten gunste van getypte essays (e.g. Canz et al., 2020), andere keren ten gunste van geschreven essays (e.g. Mogey et al., 2010; L. Yu et al., 2004). Echter, de verschillen in scores tussen de beoordelaars in het onderzoek van Mogey et al. (2010) en van Rankin (2015) waren vele malen groter dan de verschillen in scores tussen de geschreven en getypte essays. Als het aankomt op een eerlijke beoordeling, is het dus vooral van groot belang dat de beoordelaars goed getraind worden zodat ze alle antwoorden hetzelfde behandelen – ook als deze alleen getypt of alleen geschreven mogen worden.

3.3.3 Is het proces van schrijven en typen hetzelfde?

Als we onze blik richten op het antwoordproces zelf, in plaats van het product en de beoordeling daarvan, blijkt ook dat typen en schrijven niet zonder meer vergelijkbaar zijn. Met behulp van vragenlijsten, interviews en hardop-denk-protocollen analyseerden Chan et al. (2018) en Li et al. (2006) het schrijfproces zelf. Chan et al. (2018) lieten bijvoorbeeld 153 studenten twee

verschillende IELTS *English for Academic Purposes essays* schrijven, één keer met de hand en één keer op een computer. Deze essays werden vervolgens beoordeeld door gecertificeerde IELTS-beoordelaars. Chan et al. (2018) vonden geen verschil tussen de handgeschreven en de getypte essays qua scores, maar wel qua schrijfproces. Studenten die met de hand schreven, rapporteerde vaker dat ze op voorhand meer en gedetailleerder planden en dat ze tijdens het schrijven zorgvuldiger nadachten over hun woordkeuze en zinsbouw en vaker op woordniveau herzieningen maakten. Studenten die typten gaven juist aan dat ze niet met een strikt plan begonnen met schrijven en tijdens het schrijven meer focusten op een goede opbouw van alinea's en de tekst in het geheel en vaker zowel tijdens als na het schrijven hun tekst reviseerden. Masterman (2018) komt op basis van een literatuurreview tot dezelfde conclusie: ook als er geen verschillen zijn tussen getypte en geschreven antwoorden wat betreft kwaliteit en hoe ze beoordeeld worden, wil dat nog niet zeggen dat schrijven en typen hetzelfde antwoord- of schrijfproces uitlokken bij leerlingen.

3.3.4 Perspectieven van leerlingen en beoordelaars

Wat we tot nu toe besproken hebben, gaat vooral over de wetenschappelijke of feitelijke vraag of typen beter is dan schrijven of andersom. In een recente *systematic review* heeft Chan (2023) vanuit een beleidsadviserend perspectief de voor- en nadelen van getypte versus geschreven *assessments* op een rijtje gezet. Als belangrijkste voordelen noemt zij dat (a) typen beter aansluit bij de verwachtingen van werkgevers wat betreft de digitale geletterdheid van hun werknemers, (b) mensen in de afgelopen jaren steeds meer gewend zijn geraakt aan typen en (c) de leesbaarheid van getypte examenantwoorden over het algemeen beter zijn dan van geschreven antwoorden. Vanuit het perspectief van de examenkandidaat noemt ze verder als groot voordeel dat mensen over het algemeen typen fijner vinden omdat dat minder fysiek belastend en vermoeiend is dan schrijven, zeker als een examen lang duurt (langer dan 70 minuten; Hillier, 2015). En kandidaten met dyslexie of een leerstoornis kunnen gebruik maken van de spellingscontrole, wat bij hen tot minder stress en kwalitatief betere antwoorden leidt. Voor docenten is een bijkomend belangrijk voordeel dat getypte antwoorden makkelijker en sneller na te kijken zijn en dat het in sommige gevallen mogelijk is om met behulp van een algoritme (semi) automatisch feedback te geven. Een onderzoek van de New Zealand Qualifications Authority (Cole & Edwards-Hill, 2021) toont daarnaast ook aan dat correctoren het ook fijner vinden om handgeschreven antwoorden digitaal te beoordelen, omdat dat minder administratie vereist en ze op minder goed leesbare antwoorden in kunnen zoomen.

Relevante nadelen die Chan (2023) noemt, zijn dat (a) het oneerlijk kan zijn om over te schakelen op het typen van examens als niet alle leerlingen even vaardig in typen zijn en (b) typen niet voor alle vakken (bijvoorbeeld natuur- en scheikunde, wiskunde) even geschikt is. Wat betreft dat eerste nadeel is het belangrijk om op te merken dat vasthouden aan schrijven even zeer oneerlijk kan zijn voor leerlingen die niet snel kunnen schrijven of een onleesbaar handschrift hebben. Bovendien is het waarschijnlijk dat in de wereld van de toekomst leerlingen alleen maar beter zullen leren typen en slechter zullen kunnen schrijven, want schrijven doen ze steeds minder. Wat betreft het tweede nadeel geeft onderzoek van Liu et al. (2016) aan dat typen (of op andere wijze digitaal antwoord geven) voor de exacte vakken inderdaad niet altijd de beste keus is. Liu en collega's hebben in hun analyse van de PARCC toets gekeken of er verschillen zaten in hoe de items op papier en op de computer gemaakt werden. Items uit het geometrie-onderdeel voor de hogere leerjaren vereisten van leerlingen dat ze hun uitwerking moesten tonen, hun antwoord moesten toelichten of een grafiek moesten tekenen. Deze items werden beduidend beter gemaakt op papier.

3.3.5 Samenvatting

Uit bovenstaand onderzoek kunnen we de conclusie trekken dat het erop lijkt dat getypte antwoorden van even goede of betere kwaliteit zijn als/dan handgeschreven antwoorden, behalve voor exacte vakken zoals wiskunde, natuurkunde en scheikunde. In dat geval is schrijven met de hand beter. Bovendien vinden studenten typen vaak fijner en beoordelen docenten liever getypte dan handgeschreven antwoorden (omdat die leesbaarder zijn), hoewel sommige beoordelaars ook juist bevooroordeeld zijn ten gunste van handgeschreven antwoorden.
Het zou daarom mogelijk niet eerlijk zijn om sommige leerlingen te laten typen en andere te laten schrijven, tenzij beoordelaars goed getraind worden zodat ze alle antwoorden hetzelfde behandelen, wat natuurlijk sowieso belangrijk is. Ook blijkt het schrijfproces bij handgeschreven en getypte antwoorden te verschillen, hoewel dit dus niet per se tot uiting komt in de kwaliteit van de antwoorden. Wel illustreert het onderzoek van Chan et al. (2018) één van de voordelen van typen: typen maakt het voor examenkandidaten makkelijker om zowel tijdens als na het schrijven hun tekst aan te passen.

Overstappen op digitale centrale eindexamens zou in de Nederlandse context voor sommige vakken ertoe kunnen leiden dat examenkandidaten betere antwoorden produceren en minder frustratie ervaren. Ook zou deze overstap leiden tot authentiekere toetsen, aangezien leerlingen typen meer en meer doen en schrijven juist minder. Dit geldt wel vooral voor examens waarvoor (veel) geschreven moet worden, zoals examens voor de maatschappijvakken, kunstvakken en mogelijk van de exacte vakken ook biologie. Voor de examens voor de talen is het effect mogelijk minder groot, aangezien die examens voor een groter deel uit gesloten vragen bestaan.
Voor vakken waar veel gewerkt wordt met speciale symbolen en notaties, zoals wiskunde, natuurkunde en scheikunde, zou overgaan op digitaal antwoord geven echter waarschijnlijk juist nadelig zijn.

3.3.6 Wat we nog niet weten

Wat nog niet duidelijk is, is of typen ook beter is wanneer antwoorden korter zijn. Vrijwel al het onderzoek dat we hierboven besproken hebben, heeft gekeken naar *essay-style prompts*, meestal ook uit taalvaardigheidstoetsen. Dit zijn schrijfopdrachten waarvoor leerlingen vaak eerst een korte tekst moeten lezen om die vervolgens samen te vatten of er een betoog over te schrijven. Dit zijn dus opdrachten die vergelijkbaar zijn met bijvoorbeeld de schrijfopdracht in een examen vmbo bb/kb Nederlands of de meerpuntsvragen in de vwo-examens maatschappijkunde/-wetenschappen, geschiedenis en filosofie. Het valt te verwachten dat het effect van het schrijfmedium groter is als er langere teksten geschreven moeten worden. Bij kortere antwoorden – en vooral voor examens die grotendeels uit kortere antwoorden bestaan – is het dus niet zeker dat typen een even grote meerwaarde heeft.

Ook weten we niet of typen nog steeds meerwaarde heeft als leerlingen af en toe speciale symbolen, notaties of letters uit niet-Latijnse alfabetten in hun antwoorden moeten gebruiken. In bijvoorbeeld de Facet-examens natuur- en scheikunde 1 voor vmbo bb/kb is dit echter al het geval. Voor sommige maar lang niet alle vragen moeten leerlingen af en toe werken met machten en symbolen. Dit zou zo ook kunnen werken voor bijvoorbeeld een vwo-examen Grieks waar af en toe een Grieks woord in het antwoord opgenomen moet worden. Er zou meer onderzoek gedaan moeten worden naar hoe het symbolentoetsenbord in Facet de vmbo-leerlingen bevalt.

4 Interviews: Hoe kan blended toetsen voor de centrale eindexamens ingezet worden?

Op basis van het literatuuronderzoek kwamen we tot vier eerste schetsen voor blended examens voor de exacte vakken, talen, maatschappijvakken en kunstvakken. Daarin schetsen we een beeld van hoe wij dachten dat een examen voor zo'n vak op een blended manier vormgegeven kon worden en lichtten we kort toe waarom. We stelden bijvoorbeeld voor om de examens voor de talen in principe volledig te digitaliseren, maar om de leesteksten op papier aan leerlingen te verschaffen, en legden uit dat dit gebaseerd was op de hierboven besproken onderzoeken die laten zien dat lezen van papier over het algemeen beter is om tot een goed begrip van de tekst te komen en getypte antwoorden over het algemeen van dezelfde of betere kwaliteit zijn dan handgeschreven antwoorden en veel leerlingen en docenten ook een voorkeur hebben voor typen.

Met die schetsen in de hand hebben we toetsdeskundigen op de vier vakgebieden geïnterviewd. Deze toetsdeskundigen stonden bij de vierde auteur van dit rapport bekend als betrokken bij projecten rondom de (verdere) digitalisering van de examens op hun vakgebied en waren dus goede gesprekspartners om onze ideeën over blended toetsen aan voor te leggen. In die interviews zijn we dieper ingegaan op de volgende de onderzoeksvraag:

Hoe kan blended toetsen ingezet worden voor de centrale examens voor jouw vakgebied?

In de interviews zochten we naar wat volgens de toetsdeskundigen de toegevoegde waarde van digitaal en blended toetsen is (op o.a. leerlingprestaties, de logistiek, tijd- en geldbesparing, etc.). Ook stonden we stil bij de vraag of de examens nog interactiever zouden moeten worden, bijvoorbeeld door leerlingen de mogelijkheid te geven om vragen en instructies in audio- of animatievorm te laten presenteren.

Naast interviews met toetsdeskudigen, hebben we ook drie interviews gehouden met (oud) docenten van het Stanislascollege in Delft, het Amadeus Lyceum in Vleuten en het Dominicus-college in Nijmegen die in hun onderwijspraktijk al geëxperimenteerd hebben met een vorm van blended toetsen. Op het Stanislascollege hebben we gesproken met een wiskundedocent en een muziek- en wiskundedocent en op het Amadeus Lyceum met een aardrijkskundedocent. Alle drie zijn ook betrokken bij de digitalisering van toetsen op hun school. Met hen zijn we in contact gekomen via een medewerker van Woots aan wie we gevraagd hadden of hij op de hoogte was van docenten die al op enige manier in hun onderwijspraktijk toetsen blended hadden vormgegeven. De oud-docent van het Dominicuscollege gaf vroeger filosofie en is nu vakdidacticus filosofie aan de Radboud Universiteit in Nijmegen. Met hem zijn we in contact gekomen via één van de toetsdeskundigen die we geïnterviewd hadden. In deze interviews met (oud)docenten stond de volgende vraag centraal:

Wat wordt er in de onderwijspraktijk nu al gedaan op het gebied van blended toetsen en wat kunnen we daarvan leren?

De leidraden die we voor de interviews gebruikt hebben, zijn opgenomen in de Bijlage. In dit onderdeel presenteren we de bevindingen die uit de interviews naar voren kwamen, waarin we per vakgebied de perspectieven van de toetsdeskundigen integreren met de perspectieven van de (oud)docenten. Ook combineren we de inzichten uit de interviews met die uit het literatuur-onderzoek om voor elk vak(gebied) tot suggesties te komen voor hoe blended toetsen voor dat vak(gebied) ingezet zou kunnen worden voor de centrale eindexamens.

4.1 Blended toetsen en de vakgebieden

4.1.1 De exacte vakken

Voor de exacte vakken hebben we een toetsdeskundige wiskunde en een toetsdeskundige scheikunde bevraagd. De toetsdeskundige wiskunde doet ook onderzoek onder andere naar het vervangen van de grafische rekenmachine door een digitale applicatie; de toetsdeskundige scheikunde heeft acht jaar ervaring met het creëren van de digitale Nask-I-examens voor vmbo bb/kb. Ter vertegenwoordiging van de onderwijspraktijk hebben we de perspectieven van de twee wiskundedocenten van het Stanislascollege toegevoegd.

Uit het interview met de toetsdeskundigen kwam sterk naar voren dat zij enthousiast zijn over de digitale examens voor vmbo bb/kb, maar dat ze digitale en zelfs deels digitale examens niet als optie zien voor de exacte vakken op havo- en vwo-niveau. De kwestie van digitalisering voor deze vakken hangt dus sterk samen met de onderwijsrichting. Dit komt volgens de toetsdeskundigen voort uit de inhoud: de lesstof voor vmbo is minder complex dan voor havo en vwo; digitale examens passen volgens de toetsdeskundigen wél goed bij de minder complexe onderwijs-inhoud van de vmbo-examens, maar niet bij de ingewikkeldere inhoud van de havo- en vwo-examens. Ze waren bijvoorbeeld positief gestemd over de examens wiskunde, Nask I en biologie voor vmbo bb/kb die nu al gedigitaliseerd zijn en gaven ze aan dat ze die juist niet terug naar papier zouden willen brengen. De functie in Facet die het mogelijk maakt om grafieken te tekenen werd bijvoorbeeld geprezen en genoemd als een reden om deze examens vooral digitaal af te blijven nemen. Doordat de antwoorden nu getypt worden, zijn die bovendien veel beter leesbaar, nog een duidelijk voordeel van digitaal toetsen. Ook de opslagfunctie van de rekenmachine voorkomt dat leerlingen fouten in hun berekeningen en het overnemen daarvan naar het antwoordveld.

Voor havo en vwo vergt de uitwerking van zo goed als elke vraag echter van leerlingen dat ze met speciale symbolen en notaties werken, zoals integralen, vectoren, breuken, wortels, verschillende soorten haken, super- en subscript en structuurformules voor scheikunde. De toetsdeskundigen zijn van mening dat het de leerlingen onevenredig veel tijd zou kosten om hun antwoorden – met al die symbolen en notaties – in Facet (of een andere digitale toetsomgeving) in te voeren. Ook de docenten waren het ermee eens dat het moeilijk is om een digitale toets te maken voor de exacte vakken die duidelijk meerwaarde heeft ten opzichte van een papieren toets. Zij lopen bijvoorbeeld tegen het probleem aan dat vragen voor wiskunde vaak versimpeld of omgebouwd moeten worden om ze geschikt te maken voor een digitale toets. Daarom is wiskunde één van de weinige vakken waarvoor op deze school ook nog volledig op papier getoetst wordt.

Bovendien betwijfelen de toetsdeskundigen of we als maatschappij leerlingen alleen omwille van de digitalisering van de examens willen leren omgaan met een digitale interface voor het invoegen van die speciale symbolen en notaties. Vanuit de vakvernieuwingscommissie voor wiskunde klinkt volgens deze toetsdeskundigen bijvoorbeeld de roep om juist zoveel mogelijk met de hand te blijven doen. Ons lijkt dat dat komt doordat wiskunde in de kern een vaardigheid is waar papier beter bij past. De vraag is dus of we die vaardigheid willen digitaliseren. Daar is volgens de toetsdeskundigen vanuit de onderwijspraktijk geen vraag naar; uit de interviews die we met docenten hebben gehad bleek dit inderdaad ook niet zo te zijn. Dus dan rijst ook de vraag: willen we desondanks scholen de opdracht geven om leerlingen hierin te trainen?
De toetsdeskundigen waren beiden van mening dat dit niet wenselijk is. Wel gaven ze aan dat er voor schei- en natuurkunde een wens is om meer te doen met simulaties. Echter zien ze geen noodzaak om dit summatief te toetsen (en dus in de examens op te nemen); daarom zien ze deze wens ook niet als reden om de schei- en natuurkunde-examens verder te digitaliseren. Leerlingen kunnen namelijk ook in de les ervaring opdoen met simulaties. Kortom, de tools voor verdere digitalisering van de examens zijn er wel, de wens om die in te zetten is echter niet sterk aanwezig.

Manieren om de examens voor de exacte vakken in een blended vorm te gieten, zagen deze toetsdeskundigen ook niet zitten. Ons initiële voorstel om leerlingen van een uitwerkbijlage op papier te voorzien voor antwoorden die speciale notaties of symbolen vereisen, werd niet met enthousiasme ontvangen. De belangrijkste reden waarom dit volgens de toetsdeskundigen niet zou werken, is dat de meeste antwoorden speciale notaties of symbolen vereisen en dat leerlingen dan dus slechts een aantal vragen digitaal zouden beantwoorden. De toegevoegde waarde daarvan is dan dus miniem. Bovendien werken de examens met contexten: elke vraag is onderdeel van een set vragen binnen een bepaalde context. Als sommige vragen digitaal beantwoord zouden worden en andere op papier, dan zouden die contexten óf onnatuurlijk opgeknipt worden óf zou bij het selecteren van de contexten rekening gehouden moeten worden met het antwoordmedium zodat één context geheel digitaal of geheel op papier getoetst zou worden. Dit zagen de toetsdeskundigen ook niet zitten.

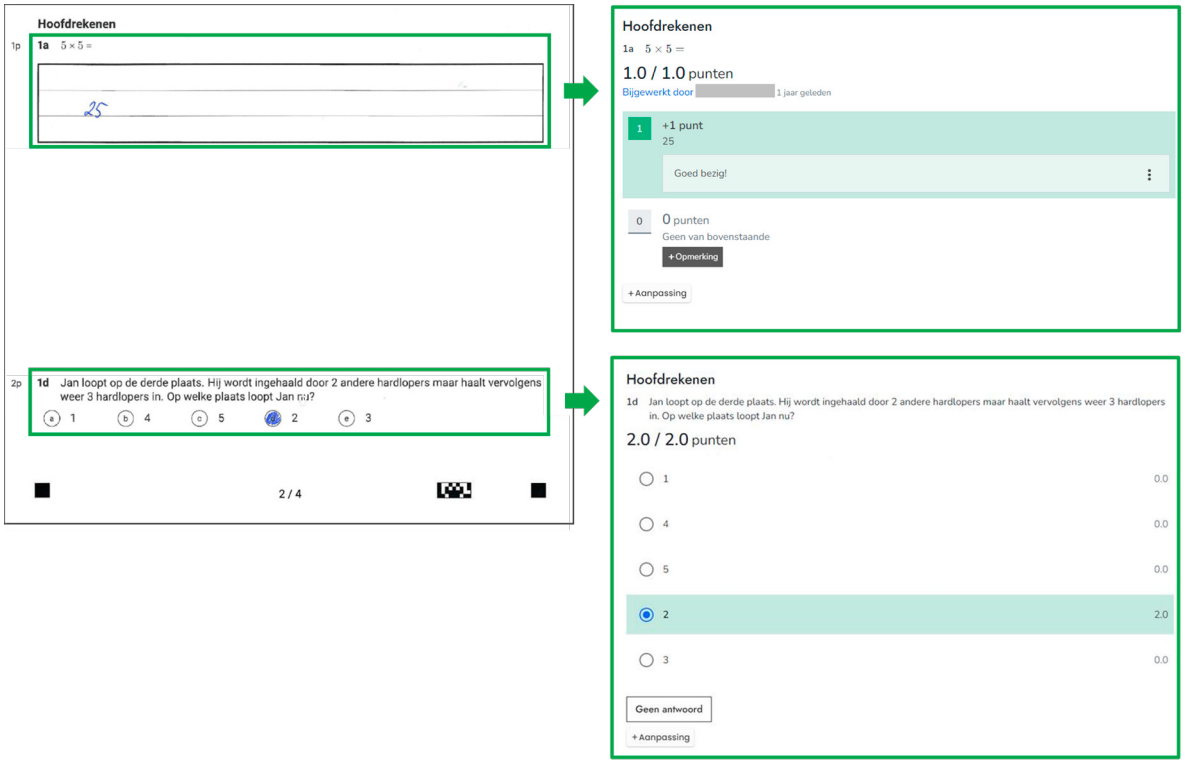
Sowieso waren die contexten voor hen nog een reden om de examens voor de exacte vakken niet verder te digitaliseren op havo- en vwo-niveau. Voor de eerste vraag binnen een context is de hoeveelheid informatie die bij die vraag hoort vaak nog op één scherm weer te geven. Die context wordt voor de daaropvolgende vragen echter steeds aangevuld met nieuwe informatie en past daardoor niet altijd meer goed op één scherm. In de digitale vmbo bb/kb-examens voor de exacte vakken zijn de contexten zijn korter en minder complex en vormen daardoor geen duidelijk obstakel voor digitalisering. Havo- en vwo-leerlingen zouden dan echter in de opgave moeten scrollen – iets wat gezien het onderzoek dat we hierboven besproken hebben mogelijk ook niet wenselijk zou zijn –, tussen bronnen of delen van de context moeten kunnen schakelen of beschikking moeten hebben tot twee schermen – wat voor scholen de logistieke puzzel nog ingewikkelder zou maken. En wanneer voor het beantwoorden van een vraag het antwoord op een vorige vraag nodig is, dan moeten leerlingen bovendien steeds terug-bladeren in de digitale omgeving om dat op te zoeken, iets waar volgens de toetsdeskundigen nu al klachten over binnenkomen. Hier geldt dus ook weer: er zijn omwegen te vinden om de huidige papieren examens te digitaliseren, maar die oplossingen zijn volgens de toetsdeskundigen niet aantrekkelijk. Ze doen een groter beroep op het schakelvermogen van de leerlingen en vergen meer tijd.

Er waren twee aspecten van de mogelijke digitalisering van de examens die wel als positief beschouwd werden. De eerste daarvan is het vervangen van de grafische rekenmachine door een digitale applicatie. Hier loopt al een onderzoek naar, dus dit idee zullen we verder niet behandelen. Het tweede aspect was de mogelijkheid om digitaal na te kijken. De examens kunnen digitaal namelijk makkelijker horizontaal (per vraag) nagekeken worden en ook makkelijker alleen door andere docenten dan die van de leerling zelf. Dit zou ervoor zorgen dat er objectiever, consistenter en betrouwbaarder nagekeken zou worden, ook volgens de docenten die we sproken zelf. Hierbij is het goed om op te merken dat deze twee dingen op papier in principe ook mogelijk zijn, maar digitaal makkelijker te implementeren zijn[4]. Bovendien zou vooral de combinatie van de twee – dat een willekeurig aangewezen docent voor een subset van leerlingen een bepaalde set vragen nakijkt –volgens de toetsdeskundigen veel meerwaarde bieden en alleen digitaal mogelijk zijn.

Via wat we het 'ANS/Woots-model' noemen zou een digitale nakijkomgeving ook te combineren zijn met papieren examens. ANS en dochteronderneming Woots zijn digitale toetsplatformen waarin het mogelijk is om een toets te creëren die op papier afgenomen kan worden maar nog wel digitaal nagekeken kan worden. De papieren toetsen kunnen namelijk na afname ingescand en geüpload worden. Het platform 'knipt' de ingescande toetsen vervolgens weer op in de losse

4 In de nakijkmodule van Facet is het mogelijk om horizontaal na te kijken. Ook kan de afnameplanner het nakijkwerk in principe enkel toewijzen aan correctoren die niet de eigen docent van de leerling zijn. In de praktijk blijkt echter dat docenten graag het werk van hun eigen leerlingen nakijken.

opgaven en beoordeeld automatisch alle gesloten vragen (tenzij het antwoord onduidelijk ingevuld is); open vragen kijkt de docent zelf digitaal na. Dat kan per vraag en naar wens ook door een onafhankelijke docent en anoniem (dan krijgt de corrector de naam van de leerling niet te zien). Voor een eenvoudig voorbeeld van hoe zo'n papieren toets en de digitale nakijkomgeving eruitzien, zie Figuur 1. Als we dit zouden toepassen op de examens voor de exacte vakken, dan zouden deze dus geheel op papier afgenomen kunnen worden en daarna ingescand en bijvoorbeeld geüpload worden naar de correctiemodule van Facet. Op het moment is dit technisch nog niet mogelijk, maar de toetsdeskundigen waren erg enthousiast over een toekomst waarin op dezelfde manier ingescande papieren eindexamens voor de exacte vakken ook digitaal nagekeken kunnen worden.



Figuur 1 | Een voorbeeld van het 'ANS/Woots-model'. Links zijn twee vragen te zien uit een op papier afgenomen Woots-toets en rechts wordt de nakijkomgeving voor die twee vragen getoond. De open vraag kijkt de docent zelf na; de gesloten vraag is automatisch nagekeken. © Woots

Als de afweging gemaakt wordt tussen digitaal toetsen, blended toetsen en toetsen op papier, is een papieren examen plus digitale nakijkomgeving volgens deze toetsdeskundigen de meest aantrekkelijke toekomst voor de havo-/vwo-examens voor de exacte vakken, in elk geval voor wiskunde, natuurkunde en scheikunde. Hoewel de tools en omwegen er dus wel zijn om verdere digitalisering mogelijk te maken, vergt leren omgaan ermee volgens hen onevenredig veel tijd en inspanning van de leerlingen en docenten. Bovendien leeft de wens om deze tools te gaan gebruiken ook niet in de dagelijkse onderwijspraktijk. Mogelijk zou een deels digitaal (dus blended) examen volgens de toetsdeskundigen wel geschikt zijn voor biologie, aangezien voor dat vak vaker gesloten vragen voorkomen en antwoorden op open vragen minder vaak speciale symbolen of notaties vereisen. Voor die vragen zou dan een papieren uitwerkbijlage verschaft kunnen worden, terwijl de rest van de vragen digitaal beantwoord wordt. Die papieren uitwerkbijlage zou dan idealiter ook ingescand worden en met de digitale antwoorden gecombineerd worden.

4.1.2 De talen

Voor de talen hebben we gesproken met een toetsdeskundige Nederlands die veel ervaring heeft met het digitaliseren van onder andere de NT2- en mbo-examens en een toetsdeskundige Spaans. Beide zijn nu ook verbonden aan het project Talenexamens van de Toekomst, waarvan het doel is om uit te proberen hoe nieuwe combinaties van (taal)vaardigheden in één (digitaal) examen getoetst kunnen worden. In de interviews met (oud)docenten kwamen de talen ook regelmatig ter sprake, al was geen van de (oud)docenten die we gesproken hebben zelf taaldocent.

Waar voor de exacte vakken de docenten het eens waren met de toetsdeskundigen, was dat niet altijd het geval voor de talen. Zo waren beide toetsdeskundigen voor de talen het erover eens dat het authentieker is om de talen zoveel mogelijk digitaal te toetsen, omdat leerlingen steeds vaker digitaal lezen en schrijven en dat ook steeds meer zullen gaan doen, zowel de leerlingen van nu in hun latere leven als de leerlingen van de toekomst tijdens hun schoolperiode. In een digitaal examen zouden leerlingen dan bijvoorbeeld vragen kunnen beantwoorden over een website of blog die er nagenoeg echt uitziet en die mogelijk ook interactief is, iets wat natuurlijk niet op papier kan. Ook de (oud)docenten merkten deze trend op en gaven aan dat digitaal toetsen daardoor authentieker kan zijn. Toch waren zij alle vier nog wat voorzichtiger en gaven ze ook aan dat leerlingen op het moment nog gewend zijn aan papier en daar ook vaak nog een voorkeur voor hebben. Zo vinden veel leerlingen het volgens de (oud)docenten nog makkelijker en fijner om op een papieren examen of in een papieren bronnenboekje aantekeningen te maken en de belangrijkste dingen te onderstrepen.

De toetsdeskundigen en (oud)docenten waren het wel eens over de kwestie van typen of schrijven. Hoewel volgens de docenten er onder leerlingen niet altijd een duidelijke voorkeur is voor typen, blijkt uit hun persoonlijke ervaringen met digitaal en blended toetsen dat dit de leesbaarheid en kwaliteit van de antwoorden (sterk) verbetert. Ook voor typen geldt bovendien dat dit als authentieker wordt beschouwd dan schrijven door zowel de (oud)docenten als de toetsdeskundigen. Daarnaast maakt digitaal toetsen het volgens de toetsdeskundigen makkelijker om de verschillende taalvaardigheden als lezen, luisteren, schrijven en spreken geïntegreerd te toetsen, een wens die momenteel leeft in het veld en waaraan met papieren examens niet voldaan kan worden.

Voor de talen komt het besluit om verder te digitaliseren of voor een vorm van blended toetsen te gaan dus neer op een afweging tussen wat authentieker is aan de ene kant en wat onderzoek over digitaal lezen laat zien aan de andere kant. Ook hangt de kwestie samen met hoe snel de samenleving verder digitaliseert en hoe de centrale eindexamens zich in de nabije toekomst ontwikkelen. De examens voor de talen richten zich op het moment nog voornamelijk op het toetsen van leesvaardigheid. Het beschikbare onderzoek toont aan dat mensen op het moment nog beter lezen op papier. Ook gaven de docenten aan dat leerlingen hier vaak zelf nog een voorkeur voor hebben. Zolang deze twee dingen zo blijven, zou het beter zijn om de leesteksten op papier aan te blijven bieden. Het examen zou dan dus een blended vorm aannemen: bronnen op papier en vraag en antwoord digitaal.

In een toekomst waarin in de taalexamens het toetsen van digitale geletterdheid een grotere focus krijgt, zou het natuurlijk voor de hand liggen om juist de bronnen digitaal te presenteren zodat die de digitale realiteit zo goed mogelijk kunnen nabootsen. Standaard leesteksten zouden nog steeds op papier aangeboden kunnen worden, afhankelijk van wat onderzoek op dat moment zegt over het effect van leesmedium op tekstbegrip. In meerdere van de artikelen die we in het literatuuronderzoek hebben besproken, werd bovendien de vraag geopperd of het mogelijk is om mensen digitale leesstrategieën aan te leren. In de toekomst zou het effect van

leesmedium op tekstbegrip daardoor misschien helemaal kunnen verdwijnen. Zodra leerlingen gewend zijn aan digitaal lezen en leren, zou het dus authentieker zijn om over te gaan op volledig digitale examens waarin alle bronnen, vragen en antwoorden digitaal gegeven worden, ook om aan te sluiten op de wens om andere vaardigheden zoals luisteren, schrijven en spreken in het centrale eindexamen te toetsen,.

4.1.3 De maatschappijvakken

Voor de maatschappijvakken hebben we gesproken met een beleidsadviseur examenvernieuwing die vroeger toetsdeskundige aardrijkskunde was en een toetsdeskundige economie die ook nog in dat vak lesgeeft. Beiden zijn betrokken bij onderzoeken om bestaande examens verder of zelfs volledig te gaan digitaliseren. De perspectieven uit de praktijk komen van de aardrijkskunde-docent van het Amadeus Lyceum en een oud-docent filosofie van het Dominicuscollege (nu vakdidacticus filosofie aan de Radboud Universiteit).

De toetsdeskundigen en (oud)docenten waren het over één ding duidelijk eens: leerlingen kunnen volgens hen beter laten zien wat ze weten als ze hun antwoorden mogen typen. Hier gaven ze verschillende redenen voor. Leerlingen typen over het algemeen sneller dan ze met de hand schrijven. Typen kost dus minder tijd. Daardoor houden ze aan het einde meer tijd over om hun antwoorden nogmaals te lezen én om ze te verbeteren. Het is namelijk ook makkelijker om getypte antwoorden aan te passen (verbeteren, herstructureren, herordenen, etc.) dan handgeschreven antwoorden. Vooral bij een vak als filosofie leidt schrijven met de hand volgens de oud-docent filosofie vaak tot krassen, onafgemaakte of slechtlopende zinnen en pijlen van het ene deel van het antwoord naar het andere. Toen deze oud-docent zijn leerlingen het eindexamen filosofie liet typen, viel hem op dat die getypte antwoorden veel beter geformuleerd en gestructureerd waren; ook de leerlingen vonden het typen om deze redenen fijner dan schrijven. Voor correctoren vraagt het nakijken van geschreven antwoorden vaak bovendien meer werk, want zulke antwoorden zijn slechter leesbaar. Ook moeten correctoren soms gokken naar de bedoeling van de leerling. Typen kan zo dus ook de betrouwbaarheid van de beoordeling ten goede komen. Bijzonder is wel dat volgens de docenten leerlingen desondanks deze voordelen niet een duidelijke voorkeur voor typen hebben.

Uit de interviews kwam niet eenduidig naar voren of het beter is om de bronnen voor deze examens op papier aan te bieden of digitaal. De (oud)docenten gaven bijvoorbeeld aan dat leerlingen nog steeds graag aantekeningen maken op het examen en dat dat vooralsnog makkelijker is op papier, zeker als het gaat om afbeeldingen, grafieken, tabellen en dergelijke. In dat geval zou een vorm van blended toetsen waarbij in elk geval de bronnen op papier aangeboden worden dus handig kunnen zijn. Tegelijkertijd noemde de aardrijkskundedocent als voordeel van digitaal toetsen dat afbeeldingen dan in kleur getoond kunnen worden en dat leerlingen erop in kunnen zoomen. Bovendien zou het mogelijk zijn om een digitale (online) atlas te gebruiken, mits die gebruiksvriendelijk is. Ook noemden de toetsdeskundigen als voordeel dat de bronnen digitaal natuurlijk geen tafelruimte in beslag nemen. Schermruimte echter wel, en het viel ons op dat in sommige van de examens die in de Facet-oefenomgeving beschikbaar zijn (bijvoorbeeld vwo kunst algemeen) niet alle bronnen tegelijkertijd op het scherm passen. De toetsdeskundigen gaven zelf ook al aan dat in digitale examens mogelijk ingeboet moet worden op het toetsen van informatievaardigheden, omdat bijvoorbeeld tekstbronnen dan korter gemaakt zouden moeten worden. Het zou dus in elk geval in sommige situaties beter kunnen zijn om ze op papier aan te bieden. Voor de opgavecontexten zou dit ook beter kunnen zijn. Het is ook volgens de toetsdeskundigen voor de maatschappijvakken niet wenselijk (en bovendien niet conform de constructie-opdrachten) om de vragen binnen die contexten volledig los van elkaar te knippen. En hoewel het mogelijk is om de context steeds te herhalen en aan te vullen, is dit zoals eerder gezegd ook niet per se een aantrekkelijke oplossing.

Door de contexten op papier aan te bieden, kunnen die goed als geheel bij elkaar gehouden worden.

Voor de maatschappijvakken kunnen we ons dus ook goed voorstellen dat blended toetsen van toegevoegde waarde is. Duidelijk is ook dat de keuze om op papier, blended of digitaal te toetsen per vak gemaakt moet worden, omdat dezelfde vorm niet voor elk vak even goed werkt. Voor aardrijkskunde zou bijvoorbeeld besloten kunnen worden om het examen in principe geheel digitaal af te nemen maar om de contexten en/of bronnen op papier aan te bieden. Met uitzondering van het kaartenkatern: zowel de toetsdeskundigen als docenten hebben behoefte aan onderzoek om uit te zoeken of het papieren kaartenkatern door een gebruiksvriendelijke digitale (online) atlas vervangen zou kunnen worden. Voor maatschappijkunde voor vmbo gl/tl zou juist de keuze kunnen worden gemaakt om het examen geheel te digitaliseren, omdat de contexten en bronnen vaak alleen korte teksten zijn, terwijl voor maatschappijwetenschappen voor havo/vwo gekozen zou kunnen worden om de bronnen (langere teksten en grafieken) op papier te presenteren. Als we uitgaan van een toekomst waarin het 'ANS/Woots-model' de standaard voor nakijken wordt, kan ten slotte voor een vak als (bedrijfs)economie ook nog overwogen worden om een uitwerkbijlage op papier te verschaffen voor vragen met grafische componenten of (complexe) berekeningen, die naderhand ingescand wordt en gecombineerd wordt met de digitale antwoorden van de leerling.

4.1.4 De kunstvakken

Voor de kunstvakken hebben we gesproken met een toetsdeskundige die de examens maakt voor kunst algemeen en voor tekenen, handvaardigheid en textiele vormgeving en een muziekdocent van het Stanislascollege. Op het moment worden de kunstexamens al in een blended vorm afgenomen: de vragen en bronnen (afbeeldingen, teksten, film- en muziekfragmenten, etc.) worden digitaal gepresenteerd en leerlingen geven op papier antwoord. Voor muziek krijgen ze hiervoor een speciale uitwerkbijlage (met voor sommige vragen bijvoorbeeld notenbalken); voor de andere vakken gebruiken ze de standaard (gelinieerde) antwoordvellen van hun eigen school. De ervaringen van toetsdeskundigen, docenten en leerlingen met de blended examens voor de kunstvakken geven dus een inkijkje in hoe blended examens voor de andere vakken mogelijk ontvangen zouden worden.

Zowel de toetsdeskundige als de docent waren het erover eens dat het voor leerlingen veel fijner is om vragen met geluids- en videofragmenten digitaal aangeboden te krijgen, zoals nu dus het geval is, omdat ze die dan in hun eigen tempo kunnen maken en de fragmenten zo vaak kunnen afspelen als ze willen (in tegenstelling tot voor de gedeeltelijke digitalisering van de kunstexamens, toen de fragmenten nog centraal afgespeeld werden). Bovendien is het kunnen tonen van videofragmenten (welhaast) een voorwaarde voor de kunstvakken drama en dans. Daarnaast past een computerafname volgens de toetsdeskundige beter bij de beeldende kunstvakken omdat ook de hedendaagse beeldende kunst zelf steeds vaker audio, video en andere digitale componenten bevat. Om bij de tijd te blijven, moe(s)ten de examens voor de kunstvakken dus in elk geval deels gedigitaliseerd worden.

De toetsdeskundige gaf verder aan dat digitaal examineren ook beter past bij een aantal wensen die leven, zoals de wens om leerlingen ook digitaal antwoord te laten geven. Dit zou op basis van het literatuuronderzoek ook aan te raden zijn, aangezien de antwoorden meestal alleen tekst bevatten en geen grafische elementen of speciale symbolen of notaties. Typen is dan dus mogelijk én stelt leerlingen waarschijnlijk in staat meer van zichzelf te laten zien. Een uitzondering is wanneer leerlingen voor muziek antwoord moeten geven in de vorm van bladmuziek of voor dans in de vorm van een ontwerp van een vloerpatroon. Hiervoor zou ook weer – net zoals nu – een papieren uitwerkbijlage gemaakt kunnen worden. De muziekddocent

met wie we gesproken hebben doet voor zijn toetsen al: de leerlingen maken toetsen in principe geheel digitaal in Woots, maar voor een aantal vragen verschaft hij een papieren uitwerkbijlage met notenbalken[5].

Ook is er volgens de toetsdeskundige een wens om ontwerp- en praktijkvaardigheden meer digitaal te gaan toetsen. Dat zou kunnen met behulp van digitale ontwerpprogramma's voor bijvoorbeeld grafisch ontwerp en audio- en videobewerking, idealiter geïntegreerd in de toetsomgeving zodat zo weinig mogelijk geschakeld hoeft te worden. Daarnaast zou de toetsdeskundige graag zien dat het praktijkwerk van de leerlingen zelf in het examen opgenomen zou kunnen worden. Dat zou bijvoorbeeld kunnen met behulp van een upload-functie zodat leerlingen hun werk in gedigitaliseerde vorm (bijvoorbeeld foto's ervan) kunnen uploaden. Zo kunnen theoretische (kunsthistorische) opdrachten en reflectie-opdrachten gericht worden op het eigen praktijkwerk van de leerling. Deze tweede wens kan volgens de toets-deskundige ook alleen gerealiseerd worden als de examens (verder) gedigitaliseerd worden, dus biedt verdere digitalisering ook op dit vlak voordelen.

Op de vraag of leerlingen het papieren examen missen, antwoordde de toetsdeskundige dat in de eerste jaren na de switch naar het digitaal aanbieden van zowel de bronnen als de vragen zowel docenten als leerlingen de papieren examenboekjes misten. Dit kwam volgens hem doordat leerlingen graag aantekeningen maken op het examen; dit wordt ze ook aangeleerd door docenten en tijdens examentrainingen. In Facet is het (inmiddels) mogelijk om tekst te markeren, maar leerlingen kunnen (nog) geen aantekeningen in de kantlijn of op bronnen maken of woorden onderstrepen of omcirkelen en dergelijke. Hoewel de wens om terug te keren naar papier niet meer zo luid klinkt, vermoedt de toetsdeskundige dat die wens er wel nog is. Zeker wanneer leerlingen voor één vraag meerdere bronnen moeten gebruiken, verwacht hij dat dit nog makkelijker is op papier. Zoals eerder gezegd passen in Facet niet altijd alle bronnen bij één vraag tegelijkertijd op het scherm. Een leerling kan in het ergste geval dan dus slechts een deel van de bronnen tegelijkertijd bekijken en moet de vraag en de andere bronnen in het werkgeheugen houden. Als de bronnen (ook) op papier aangeboden werden, zou het volgens de toetsdeskundige waarschijnlijk makkelijker zijn voor leerlingen om het overzicht te behouden, bronnen met elkaar te vergelijken en om alles tot één antwoord op de vraag te combineren.

Samengevat zouden we voor de kunstvakken dus aanraden om uit te gaan van het principe van "digitaal, tenzij". Dat wil zeggen, bronnen en vragen worden digitaal aangeboden en leerlingen geven ook digitaal antwoord. De tenzij omvat twee situaties. De eerste situatie is als bij een vraag meerdere afbeeldingen gegeven worden (zeker als die niet allemaal op het scherm passen). In dat geval zou het mogelijk beter zijn om die afbeeldingen (ook) op papier aan de leerlingen aan te bieden. De tweede situatie is als bij een vraag antwoord gegeven moet worden in de vorm van muzieknotatie of een tekening. Voor die vragen zou een uitwerkbijlage op papier beter zijn. Net als voor de exacte vakken zal papier hier waarschijnlijk altijd het beste medium voor zijn en is een blended examen dus waarschijnlijk een eindstation. Het tonen van meerdere afbeeldingen kan daarentegen ook prima digitaal, mits leerlingen tijdens het examen een groot genoeg scherm hebben of zelfs beschikken over twee schermen. In dit geval zou een blended examen mogelijk slechts een tijdelijke oplossing zijn.

5 Op het moment is het nog niet mogelijk in ANS/Woots om sommige van de vragen van één toets op papieren af te nemen en de rest digitaal. Voor de uitwerkbijlage maakt hij daarom een aparte papieren toets in Woots. De scans daarvan kan hij dus wel digitaal nakijken.

4.2 De voor- en nadelen van digitaal toetsen

In de interviews hebben we ook gevraagd naar wat de voor- en nadelen van digitaal toetsen zijn volgens onze gesprekspartner(s). Hoewel zij het dus niet altijd eens waren over welke (vakspecifieke) vorm de examens in de toekomst moeten aannemen – op papier, blended of volledig digitaal – waren ze het hier in grote lijnen wel over eens. Het grootste voordeel van de digitalisering van toetsen en examens is wat hen betreft het typen. De voordelen van typen zijn al vaker benoemd in dit rapport, maar voor de volledigheid herhalen we ze hier nog eens. Typen sluit beter aan op ontwikkelingen in de samenleving: mensen typen steeds meer en schrijven steeds minder met de hand. Typen is dus authentieker. Leerlingen kunnen bovendien vaak sneller typen dan ze kunnen schrijven. Daardoor houden ze tijd over en kunnen ze beter laten zien wat ze in hun mars hebben. Ook is het makkelijker voor leerlingen om hun antwoorden aan te passen, te herstructureren en herorganiseren. Getypte antwoorden zijn dus niet alleen beter leesbaar, maar vaak ook helderder geformuleerd. Nakijken vergt daarom minder moeite en is minder onderhevig aan de interpretatie van de corrector. Typen zou zo dus ook indirect kunnen bijdragen aan een betrouwbaardere beoordeling.

Wat betreft digitaal nakijken zagen zowel de toetsdeskundigen en (oud)docenten in het algemeen ook grote voordelen. Eén daarvan is natuurlijk dat gesloten vragen veelal volledig automatisch nagekeken kunnen worden. Maar ook getypte antwoorden – en mogelijk op termijn ook ingescande handgeschreven antwoorden – kunnen met behulp van AI (semi) automatisch nagekeken worden. In een digitale nakijkomgeving zou bovendien het makkelijker zijn om de examens alleen na te laten kijken door andere docenten dan die van de leerling zelf. Onze gesprekspartners verwachten allemaal dat dit zal leiden tot objectievere en betrouwbaardere scores. Verder is het digitaal ook makkelijker om horizontaal (vraag per vraag) na te kijken, wat ook bijdraagt aan de objectiviteit van de beoordeling, en om paarsgewijs na te kijken. Vooral voor productieve vaardigheden heeft dit meerwaarde, omdat mensen beter zijn in het vergelijken van twee producten (bijvoorbeeld korte essays) dan in het toekennen van een absolute score daaraan (Crompvoets et al., 2020; Heldsinger & Humphry, 2010). Paarsgewijs nakijken zou de betrouwbaarheid en validiteit van de scores dus ook kunnen vergroten. Bovendien is een groot voordeel van digitaal nakijken dat de examenantwoorden niet naar de tweede corrector opgestuurd hoeven te worden en dat beide correctoren nog toegang hebben tot de antwoorden voor hun overleg over de uiteindelijke scores.

Onze gesprekspartners gaven ook aan dat digitaal toetsen in het algemeen vaak authentieker is, onder andere vanwege de grote verscheidenheid aan (interactieve) vraagvormen. De leerlingen van vandaag (en morgen) gebruiken in de les ook steeds vaker een device. Digitaal toetsen sluit dus ook beter aan op het digitale leerproces. Daarnaast is het digitaal mogelijk om nieuwe constructen en vaardigheden te toetsen die op papier niet (makkelijk) te toetsen zijn. Een voorbeeld hiervan is de pilot om het toetsen van digitale geletterdheid op te nemen in de examens Nederlands. Digitale examens kunnen dus ook diverser zijn. Bovendien kunnen afbeeldingen in kleur getoond worden, kunnen leerlingen naar wens in- en uitzoomen en geluids- en filmfragmenten meerdere keren afspelen en is het met digitaal toetsen mogelijk om adaptief en flexibel te toetsen. Hier merkte de toetsdeskundige voor de kunstvakken wel bij op dat devices verschillen in kleurtonen en schermgrootte en dat afbeeldingen en filmfragmenten er dus niet voor elke leerling precies hetzelfde uit zullen zien. Dit zou ertoe kunnen leiden dat leerlingen antwoorden geven die onterecht fout gerekend worden, hoewel dit volgens de toetsdeskundige nog niet voorgekomen is.

Een ander voordeel van digitaal toetsen is dat de examens nog multimodaler ingericht zouden kunnen worden, door vragen en instructies in film- of animatievorm te laten presenteren. De toetsdeskundigen gaven aan dat animaties zeker voor wiskunde een positieve toevoeging

zou kunnen zijn, ter vervanging van de situatieschetsen die nu alleen uit tekst en afbeeldingen bestaan. Voor natuurkunde zouden animaties mogelijk ook van toegevoegde waarde zijn; voor scheikunde interactieve 3D-modellen van bijvoorbeeld moleculen. Ook de toetsdeskundige economie gaf aan dat filmpjes of animaties van meerwaarde zouden kunnen zijn voor dat vak, omdat de vragen dan beter gekoppeld kunnen worden aan de realiteit. Nu zijn de geschetste situaties in economie-examens vaak fictief. Bovendien zouden met filmpjes en animaties complexe contexten mogelijk makkelijker uit te leggen kunnen zijn. Twee docenten waren het hiermee eens; de andere docenten en toetsdeskundigen reageerden echter niet enthousiast. Voor sommige maar niet alle vakken biedt digitaal examineren om deze reden dus ook meerwaarde.

Een laatste voordeel van digitaal toetsen is de mogelijkheid om adaptief te toetsen, wat noch in de interviews noch in het literatuuronderzoek aan bod kwam maar wel het vermelden waard is. Bij een adaptieve toets wordt de toets (meestal in real-time) aangepast aan het niveau van de leerling (van Groen et al., 2021). Dit bevordert de motivatie van de leerling, omdat die zo min mogelijk vragen hoeft te beantwoorden die te makkelijk of te moeilijk zijn. Vaak is ook een kleiner aantal vragen nodig om het vaardigheidsniveau van de leerling goed in te schatten. Adaptief toetsen is ook mogelijk met papieren toetsen: een leerling maakt dan een set vragen met als resultaat een score of vaardigheidsinschatting op basis waarvan de volgende set vragen uitgekozen wordt. Of de leerling maakt de versie van een toets die het beste bij zijn of haar niveau past. Een echt adaptieve toets die zich in real-time aanpast aan het niveau van de leerling kan echter alleen digitaal afgenomen worden (van Groen et al., 2021). Op dit vlak biedt digitaal examineren dus weer een groot voordeel, al zal het waarschijnlijk lastig blijken om computergestuurd adaptief toetsen te combineren met blended toetsen.

Natuurlijk werden er ook zorgen geuit over de nadelen van digitaal examineren. In bijna alle interviews kwam de kwestie van middelen en veiligheid ter sprake. Kort samengevat hebben scholen niet genoeg geld om voor elke leerling een examendevice te kopen. Vaak is dit ook dubbelop, omdat leerlingen voor gebruik in de les ook al een (door de school uitgekozen) device moeten aanschaffen. Daarnaast vergt digitaal examineren ook extra investering in de infrastructuur, zoals de stabiliteit van het wifinetwerk (om uitval tijdens het afnemen van een examen te voorkomen) en basale dingen als stopcontacten (voor het opladen van al die examen-laptops). Voordat de overstap op (volledig) digitaal examineren gemaakt kan worden, is het dus zaak om deze logistieke puzzel op te lossen. Daarnaast is het natuurlijk zeer belangrijk dat maatregelen genomen worden om fraude tegen te gaan, wat zeker voor scholen die regulieren toetsen digitaal afnemen op de devices van de leerlingen zelf een heikel punt is, waardoor digitaal toetsen niet overal enthousiast ontvangen wordt.

Ook uitten veel van de toetsdeskundigen zorgen over de gevolgen van digitaal examineren voor kansengelijkheid. Leerlingen verschillen in hoe digitaal vaardig ze zijn, wat onder andere komt doordat niet elke leerling even veel toegang heeft tot een device op school én thuis. Ook hebben ze niet allemaal toegang tot dezelfde soorten devices en maken ze het examen (op het moment in elk geval) niet op hetzelfde (standaard)device. Al deze factoren kunnen van invloed zijn op hoe makkelijk het is voor een leerling om het examen digitaal te maken. Een groot aantal onder-zoeken heeft bijvoorbeeld al laten zien dat (kenmerken van) het device waarop het examen gemaakt wordt – zoals schermgrootte en of het device een fysiek toetsenbord heeft of niet – samenhangen met toetsscores (zie Way & Strain-Seymour, 2021, voor een overzicht). Bovendien zijn de verschillen in hoe digitaal vaardig leerlingen zijn mogelijk op het moment nog te groot om over te stappen op volledig digitaal examineren. Het is dus ook nodig om de ontwikkeling van de digitale vaardigheid van leerlingen te stimuleren en te monitoren in deze overgangs-periode.

Ten slotte bleek uit de interviews dat leerlingen – en docenten – veel dingen nog gewend zijn te doen op papier. Veel docenten hebben bijvoorbeeld nog een voorkeur voor nakijken op papier. Leerlingen vinden het vaak nog fijner om op papier te lezen en van papier te leren, en kunnen op papier beter het overzicht bewaren. Bovendien kwam vaak ter sprake dat leerlingen het fijn vinden om tijdens het examen aantekeningen te maken op het examen of de bronnen zelf.
De meeste van onze gesprekspartners verwachten dat dit nog een tijd zo zal blijven. Het is daarom wellicht geen gek idee om leerlingen tijdens een examen altijd iets op papier aan te bieden, al is het maar kladpapier, en om gebruik hiervan te stimuleren. Volgens de oud-docent filosofie zou zo'n stuk papier ook kunnen functioneren als een soort van podium voor denkwerk, terwijl het digitale device functioneert als een podium voor het produceren van de antwoorden.

5 Conclusie en discussie

5.1 De mogelijkheden van blended toetsen

Voor de meeste vakken van de vier vakgebieden de we behandeld hebben in dit rapport, is een vorm van een blended examen in te denken en zelfs aan te raden op basis van de (wetenschappelijke) onderzoeken die we besproken hebben en de interviews die we gehouden hebben. Hier zit tussen de vakken wel veel variatie in. Eén en dezelfde oplossing zou dus niet voor alle vakken even goed werken. Voor de kunstvakken zouden de examens bijvoorbeeld bijna geheel gedigitaliseerd kunnen worden. Papier is voor deze vakken op het moment vooral nog van toegevoegde waarde voor vakken zoals muziek als leerlingen bijvoorbeeld noten op een notenbalk moeten noteren. Zelfs wat dat laatste betreft is het niet uit te sluiten dat in de toekomst goede software geïntegreerd kan worden in Facet waarmee leerlingen digitaal muzieknoten op notenbalken kunnen plaatsen. Blended toetsen kan voor de kunstvakken dus ook slechts een tussenstation zijn op weg naar volledige digitalisering. Voor de talen en maatschappijvakken zou verdere digitalisering ook van toegevoegde waarde zijn: zowel docenten als leerlingen zouden baat hebben bij getypte examens. Bovendien draagt digitalisering bij aan de authenticiteit van de examens. Voor deze examens zou, afhankelijk van het vak, papier alleen nog het beste medium zijn voor bronnen en uitwerkbijlages, in elk geval totdat leerlingen gewend zijn aan digitaal lezen. Voor de taal- en maatschappijvakken is het blended examen dus ook mogelijk alleen een tussenstation. Alleen de exacte vakken wiskunde, scheikunde en natuurkunde vormen hier een uitzondering op. Voor deze vakken werkt papier nu – en waarschijnlijk in de toekomst – nog het beste volgens de toetsdeskundigen én (oud)docenten. Het ligt daarom voor de hand om deze examens nu en in de toekomst nog geheel op papier af te nemen.

Dit onderzoek geeft mogelijk ook aanleiding om te reflecteren op de digitale examens voor vmbo bb/kb. Zoals eerder gezegd is voor zover wij weten in de Nederlandse context geen onderzoek gedaan naar de verschillen in leerlingprestaties op de digitale en papieren examens. Ook hebben we geen tijd gehad om bij toetsdeskundigen voor alle vakken na te gaan hoe deze bij de leerlingen en docenten zijn bevallen. Het is dus onduidelijk wat de gevolgen zijn geweest voor de leerlingen van de volledige digitalisering van de vmbo bb/kb examens. Op basis van ons literatuuronderzoek is het echter niet uit te sluiten dat dit nadelig uitgepakt heeft voor sommige examens. Voor de taalexamens worden de leesteksten nu digitaal aangeboden. Uit de literatuur weten we dat dit het tekstbegrip van de lezer kan beïnvloeden. Zeker voor lezers met een lagere leesvaardigheid kan dit grote nadelige gevolgen hebben (Stiegler-Balfour et al., 2023). Tegelijkertijd waren beide toetsdeskundige voor de exacte vakken enthousiast over hoe de wiskunde-, biologie- en Nask I-examens voor vmbo bb/kb gedigitaliseerd waren. Ook hier geldt dus dat dezelfde aanpak niet voor alle vakken het beste is. Het is daarom belangrijk om te onderzoeken of de overstap op digitaal examineren ook in de Nederlandse context consequenties heeft gehad voor de prestaties van de leerlingen, of de gevolgen voor alle vakken hetzelfde waren en om dat in toekomstig beleid mee te laten wegen.

5.2 Personaliseerbare toetsen

Het idee van personaliseerbare toetsen vindt in de huidige maatschappij steeds meer steun, gevoed door het streven naar eerlijke kansen. Dit idee heeft raakvlakken met de kwestie van blended toetsen: zouden leerlingen de keuze moeten hebben of ze de vragen en bronnen op

papier of digitaal gepresenteerd willen krijgen en of ze schriftelijk of digitaal antwoord willen geven? We hebben dit niet expliciet in ons onderzoek opgenomen, maar kunnen op basis van alle kennis die we vergaard hebben hier wel iets over zeggen.

In een onderzoek met bachelorstudenten naar de effecten van leesmedium op tekstbegrip en kalibratie vonden Singer en Alexander (2017) bijvoorbeeld dat deze studenten een duidelijke voorkeur hadden voor digitaal lezen. De studenten die de teksten tijdens het onderzoek digitaal mochten lezen, dachten ook dat ze die beter begrepen hadden dan de studenten die de teksten op papier lazen. Dit bleek echter niet zo te zijn. De studenten die de teksten op papier hadden gelezen, hadden de belangrijkste informatie beter onthouden en ook hun eigen tekstbegrip beter ingeschat. Halamish en Elbaz (2020) kwamen tot dezelfde conclusie in hun onderzoek met leerlingen uit groep 7. Deze leerlingen begrepen de teksten die ze op papier hadden gelezen beter dan de teksten die ze op een computer hadden gelezen, ongeacht of ze zelf een voorkeur hadden voor lezen op papier of digitaal. Ook veranderden de leerlingen niet van voorkeur nadat ze het experiment gedaan wat, wat suggereert dat ze niet goed konden inschatten welk medium het beste voor ze is. Met andere woorden, als we leerlingen de keus geven om de leesteksten digitaal dan wel op papier te lezen, zal dit er niet per se voor zorgen dat leerlingen het examen onder eerlijke(re) omstandigheden maken.

Onderzoek lijkt ook aan te wijzen dat het geen goed idee zou zijn om leerlingen de keus te geven of ze het examen willen schrijven of typen. Volgens de docenten die we spraken heerst er onder leerlingen geen duidelijke voorkeur voor typen dan wel schrijven. Het klinkt dus verleidelijk om leerlingen die keus te geven. We hebben echter gezien dat onderzoek aantoont dat beoordelaars geschreven en getypte antwoorden niet altijd hetzelfde behandelen (Canz et al., 2020; Mogey et al., 2010; Rankin, 2015; L. Yu et al., 2004). Bovendien blijkt ook het proces van schrijven en typen niet hetzelfde te zijn (S. Chan et al., 2018; J. Li, 2006; Masterman, 2018). Dit zou kunnen betekenen dat leerlingen die hun antwoorden met de hand schrijven antwoorden geven van een net andere strekking dan leerlingen die typen. Dit zou het nakijken ingewikkelder kunnen maken. Ook op dit vlak is het dus waarschijnlijk beter om leerlingen geen keuze te geven.

5.3 Wat weten we nog steeds niet?

In dit rapport hebben we verkend hoe het concept van blended toetsen ingezet kan worden voor de centrale eindexamens in de Nederlandse context. Uiteraard hebben we niet aan alle aspecten en factoren die hierin een rol spelen even veel aandacht kunnen geven. In een vervolgonderzoek zou het daarom belangrijk zijn om bij de volgende dingen ook stil te staan.

De eerste daarvan is wat de gevolgen zijn van blended en digitaal toetsen voor de validiteit en betrouwbaarheid van de examens. Omdat digitaal toetsen mogelijk authentieker is, kan dit de validiteit van de examens vergroten. Tegelijkertijd kunnen de validiteit en betrouwbaarheid ook juist in het geding komen als de examens (ongewenst) ook de digitale vaardigheden van de leerlingen meten. Ook zou het goed zijn om in vervolgonderzoek kansengelijkheid expliciet mee te nemen. Zijn er leerlingen voor wie digitaal of blended toetsen uiteindelijk vooral nadelig uit zou pakken en leerlingen voor wie het juist voordelig zou zijn? In sommige onderzoeken is al gekeken naar factoren als sociaal-economische status, en leesvaardigheid, maar ook factoren zoals geslacht en digitale vaardigheid kunnen belangrijk zijn. Een vervolgonderzoek zou ook dieper in moeten gaan op de vraag of er inderdaad sprake is van een periode van gewenning, zoals de resultaten van Backes en Cowan (2019) lieten zien. Belangrijk hierbij zijn vragen als: Waar moeten leerlingen precies aan wennen? Zijn er groepen leerlingen die meer aan bepaalde dingen zullen moeten wennen dan andere leerlingen? Hoe lang duurt het wennen en welke vaardigheden komen daar bij kijken? Hierbij is het ook goed om op te merken dat volgens de

docenten die we spraken de vorm van het centrale eindexamen leidend is voor de toetspraktijk op hun scholen. Met andere woorden, als centrale eindexamens verder gedigitaliseerd worden, dan zullen scholen hun reguliere toetsen en de schoolexamens ook vaker in gedeeltelijk of geheel digitale vorm afnemen. Dit zal ongetwijfeld leerlingen helpen om te wennen aan digitale (en blended) toetsen.

Ten slotte is er één belangrijke vraag die dit onderzoek niet kan beantwoorden: werkt blended toetsen echt in de praktijk? Tijdens de interviews werd meerdere malen de zorg uitgesproken of blended toetsen niet een te groot beroep zou doen op het organisatorisch vermogen en de executieve functies van de leerlingen. Voor sommige examens is het volgens een aantal van onze gesprekspartners nu al het geval dat leerlingen te veel op hun tafel hebben. Voor het vwo-examen aardrijkskunde hebben leerlingen bijvoorbeeld een opgavenboekje, bijlage met bronnen, kaartenkatern op A3-formaat en een antwoordvel. Daarnaast mogen ze ook gebruik maken van extra hulpmiddelen zoals kladpapier, een geodriehoek, passer, rekenmachine en een woordenboek en hebben ze natuurlijk schrijfmateriaal nodig. In een blended vorm van dit examen zouden leerlingen een device hebben (met in de digitale toetsomgeving de opgaven en ruimte voor de antwoorden), een papieren bronnenboekje, kaartenkatern en mogelijk ook een uitwerkbijlage, plus de extra hulpmiddelen en eventueel nog een losse muis (ervan uitgaande dat dit toegestaan is). In totaal zouden leerlingen dus niet per definitie minder op hun tafel hebben liggen tijdens een blended examen. Bovendien zou het schakelen tussen het device en de papieren onderdelen een extra belasting op hun executieve functies kunnen betekenen.
De kunstvakken tonen al aan dat blended toetsen in principe kan werken, zolang de precieze vorm goed aansluit op de vakinhoud en behoeften van leerlingen. In een vervolgonderzoek zou het echter goed zijn om ook vanuit de principes van de *cognitive load theory* (Sweller et al., 1998, 2019) en *cognitive theory of multimedia learning* (Mayer & Moreno, 1998) naar de kwestie van blended examens te kijken.

6 Bibliografie

Annisette, L. E., & Lafreniere, K. D. (2017). Social media, texting, and personality: A test of the shallowing hypothesis. *Personality and Individual Differences, 115*, 154-158. https://doi.org/10.1016/j.paid.2016.02.043

Backes, B., & Cowan, J. (2019). Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Economics of Education Review, 68*, 89-103. https://doi.org/10.1016/j.econedurev.2018.12.007

Barkaoui, K., & Knouzi, I. (2018). The effects of writing mode and computer ability on L2 test-takers' essay characteristics and scores. *Assessing Writing, 36*, 19-31. https://doi.org/10.1016/j.asw.2018.02.005

Brunfaut, T., Harding, L., & Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing*, *36*, 3-18. https://doi.org/10.1016/j.asw.2018.02.003

Canz, T., Hoffmann, L., & Kania, R. (2020). Presentation-mode effects in large-scale writing assessments. *Assessing Writing*, *45*, 100470. https://doi.org/10.1016/j.asw.2020.100470

Carr, N. (2010). *The shallows: What the internet is doing to our brains*. WW Norton & Company.

Ceka, B., & O'Geen, A. J. (2019). Evaluating Student Performance on Computer-Based versus Handwritten Exams: Evidence from a Field Experiment in the Classroom. *PS: Political Science & Politics*, 52(4), 757-762. https://doi.org/10.1017/S104909651900091X

Chan, C. K. Y. (2023). A systematic review – handwritten examinations are becoming outdated, is it time to change to typed examinations in our assessment policy? *Assessment & Evaluation in Higher Education, 0*(0), 1-17. https://doi.org/10.1080/02602938.2023.2219422

Chan, S., Bax, S., & Weir, C. (2018). Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test. *Assessing Writing, 36*, 32-48. https://doi.org/10.1016/j.asw.2018.03.008

Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing, 16*(1), 49-71. https://doi.org/10.1016/j.asw.2010.11.001

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd dr.). Lawrence Erlbaum Associates.

Cole, R., & Edwards-Hill, A. (2021). *Research Report – How markers respond to digital (typed) and scanned (handwritten) exam responses*. New Zealand Qualification Authority. https://www.nzqa.govt.nz/assets/About-us/Future-State/NCEA-Online/research-and-innovation/OSM-2021-Research-Report.pdf

Crompvoets, E. A. V., Béguin, A. A., & Sijtsma, K. (2020). Adaptive Pairwise Comparison for Educational Measurement. *Journal of Educational and Behavioral Statistics, 45*(3), 316-338. https://doi.org/10.3102/1076998619890589

Delgado, P., & Salmerón, L. (2021). The inattentive on-screen reading: Reading medium affects attention and reading comprehension under time pressure. *Learning and Instruction, 71*, 101396. https://doi.org/10.1016/j.learninstruc.2020.101396

Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review, 25*, 23-38. https://doi.org/10.1016/j.edurev.2018.09.003

Eklundh, K. S. (1992). Problems in achieving a global perspective of the text in computer-based writing. *Instructional Science, 21*(1), 73-84. https://doi.org/10.1007/BF00119656

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. https://doi.org/10.3102/0034654321998074

Goldberg, A., Russell, M., & Cook, A. (2003). The Effect of Computers on Student Writing: A Meta-analysis of Studies from 1992 to 2002. *The Journal of Technology, Learning and Assessment, 2*(1), Article 1. https://ejournals.bc.edu/index.php/jtla/article/view/1661

Gordanier, J., Ozturk, O., & Zhan, C. (2023). Pencils Down? Computerized Testing and Student Achievement. *Education Finance and Policy, 18*(2), 232-252. https://doi.org/10.1162/edfp_a_00373

Gu, L., Ling, G., Liu, O. L., Yang, Z., Li, G., Kardanova, E., & Loyalka, P. (2021). Examining mode effects for an adapted Chinese critical thinking assessment. *Assessment & Evaluation in Higher Education, 46*(6), 879-893. https://doi.org/10.1080/02602938.2020.1836121

Halamish, V., & Elbaz, E. (2020). Children's reading comprehension and metacomprehension on screen versus on paper. *Computers & Education, 145*, 103737. https://doi.org/10.1016/j.compedu.2019.103737

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher, 37*(2), 1-19. https://doi.org/10.1007/BF03216919

Hillier, M. (2015). To type or handwrite: Student's experience across six e- Exam trials. *Globally Connected, Digitally Enabled,* 463-470. http://transformingexams.com/files/Hillier_2015_ascilite_fp.pdf

Hillier, M., & Lyon, N. (2019). Writing e-Exams in Pre-University College. In D. Passey, R. Bottino, C. Lewin, & E. Sanchez (Red.), *Empowering Learners for Life in the Digital Age* (pp. 264-274). Springer International Publishing. https://doi.org/10.1007/978-3-030-23513-0_26

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does It Matter if I Take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP. *Journal of Technology, Learning, and Assessment, 5*(2). https://eric.ed.gov/?id=EJ843858

Hou, J., Rashid, J., & Lee, K. M. (2017). Cognitive map or medium materiality? Reading on paper and screen. *Computers in Human Behavior, 67*, 84-94. https://doi.org/10.1016/j.chb.2016.10.014

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology, 33*(4), 410-422. https://doi.org/10.1080/0144929X.2012.710647

Jin, Y., & Yan, M. (2017). Computer Literacy and the Construct Validity of a High-Stakes Computer-Based Writing Assessment. *Language Assessment Quarterly, 14*(2), 101-119. https://doi.org/10.1080/15434303.2016.1261293

Kim, D.-H., & Huynh, H. (2008). Computer-Based and Paper-and-Pencil Administration Mode Effects on a Statewide End-of-Course English Test. *Educational and Psychological Measurement, 68*(4), 554-570. https://doi.org/10.1177/0013164407310132

Kim, H. R., Bowles, M., Yan, X., & Chung, S. J. (2018). Examining the comparability between paper- and computer-based versions of an integrated writing placement test. *Assessing Writing, 36*, 49-62. https://doi.org/10.1016/j.asw.2018.03.006

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. https://doi.org/10.1016/j.compedu.2018.05.005

Kuhlemeier, H., Straat, H., & Van der Molen, P. (2015). *Prestaties op papieren en digitale examens: Wat is het verschil?* Stichting Cito.

Latini, N., Bråten, I., Anmarkrud, Ø., & Salmerón, L. (2019). Investigating effects of reading medium and reading purpose on behavioral engagement and textual integration in a multiple text context. *Contemporary Educational Psychology, 59*, 101797. https://doi.org/10.1016/j.cedpsych.2019.101797

Li, D., Yi, Q., & Harris, D. (2017). *Evidence for Paper and Online ACT Comparability: Spring 2014 and 2015 mode comparability studies* (ACT Research Reports). American College Testing. https://www.act.org/content/dam/act/unsecured/documents/Working-Paper-2016-02-Evidence-for-Paper-and-Online-ACT-Comparability.pdf

Li, J. (2006). The mediation of technology in ESL writing and its implications for writing assessment. *Assessing Writing, 11*(1), 5-21. https://doi.org/10.1016/j.asw.2005.09.001

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms* (NCSER 2013-3000). National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. https://eric.ed.gov/?id=ED537446

Liu, J., Brown, T., Chen, J., Ali, U., Hou, L., & Costanzo, K. (2016). *Mode comparability study based on Spring 2015 operational PARCC test data.* Educational Testing Service. https://eric.ed.gov/?id=ED599049

Luyten, H., Merrell, C., & Tymms, P. (2017). The contribution of schooling to learning gains of pupils in Years 1 to 6. *School Effectiveness and School Improvement, 28*(3), 374-405. https://doi.org/10.1080/09243453.2017.1297312

Lynch, S. (2022). Adapting Paper-Based Tests for Computer Administration: Lessons Learned from 30 Years of Mode Effects Studies in Education. *Practical Assessment, Research & Evaluation, 27*. https://eric.ed.gov/?id=EJ1359345

MacCann, R., Eastment, B., & Pickering, S. (2002). Responding to free response examination questions: Computer versus pen and paper. *British Journal of Educational Technology, 33*(2), 173-188. https://doi.org/10.1111/1467-8535.00251

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research, 58*, 61-68. https://doi.org/10.1016/j.ijer.2012.12.002

Marshall, J. C., & Powers, J. M. (1969). Writing Neatness, Composition Errors, and Essay Grades. *Journal of Educational Measurement, 6*(2), 97-101. https://doi.org/10.1111/j.1745-3984.1969.tb00665.x

Masterman, E. (2018). *Typed versus handwritten essay exams: Is there a need to recalibrate the gauges for digital assessment?* 35th International Conference on Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education, Geelong, Victoria, Australia.

Mayer, R. E., & Moreno, R. (1998). A cognitive theory of multimedia learning: Implications for design principles. *Journal of educational psychology, 91*(2), 358-368.

Mazzeo, J., & Harvey, A. L. (1988). The Equivalence of Scores from Automated and Conventional Educational and Psychological Tests. *ETS Research Report Series, 1988*(1), i-27. https://doi.org/10.1002/j.2330-8516.1988.tb00277.x

Mogey, N., & Hartley, J. (2013). To write or to type? The effects of handwriting and word-processing on the written style of examination essays. *Innovations in Education and Teaching International, 50*(1), 85-93. https://doi.org/10.1080/14703297.2012.748334

Mogey, N., Paterson, J., Burk, J., & Purcell, M. (2010). Typing compared with handwriting for essay examinations at university: Letting the students choose. *ALT-J, 18*(1), 29-47. https://doi.org/10.1080/09687761003657580

Piolat, A., Roussey, J.-Y., & Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies, 47*(4), 565-589. https://doi.org/10.1006/ijhc.1997.0145

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). They Think Less of My Handwritten Essay If Others Word Process Theirs? *Journal of Educational Measurement, 31*(3), 220-233. https://doi.org/10.1111/j.1745-3984.1994.tb00444.x

Proctor, T. P., Chuah, S. C., Montgomery, M., & Way, W. D. (2019). *Comparability of Performance on the SAT® Suite of Assessments Across Pencil-and-Paper and Computer-Based Modes of Administration.* College Board. https://satsuite.collegeboard.org/media/pdf/comparing-performance-paper-digital-tests-sat-suite-assessments.pdf

Rankin, A. D. (2015). *A comparability study on differences between scores of handwritten and typed responses on a large-scale writing assessment* [Doctor of Philosophy, University of Iowa]. https://doi.org/10.17077/etd.2j82ewmk

Ronconi, A., Veronesi, V., Mason, L., Manzione, L., Florit, E., Anmarkrud, Ø., & Bråten, I. (2022). Effects of reading medium on the processing, comprehension, and calibration of adolescent readers. *Computers & Education, 185*, 104520. https://doi.org/10.1016/j.compedu.2022.104520

Russell, M., & Haney, W. (1997). Testing Writing on Computers. *Education Policy Analysis Archives, 5*, 3-3. https://doi.org/10.14507/epaa.v5n3.1997

Russell, M., & Plati, T. (2001). *Effects of computer versus paper administrations of a state-mandated writing assessment* (inTASC Publications). Council of Chief State School Officers.

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A Meta-Analysis of Interventions for Struggling Readers in Grades 4–12: 1980–2011. *Journal of Learning Disabilities, 48*(4), 369-390. https://doi.org/10.1177/0022219413504995

Schwabe, A., Lind, F., Kosch, L., & Boomgaarden, H. G. (2022). No Negative Effects of Reading on Screen on Comprehension of Narrative Texts Compared to Print: A Meta-analysis. *Media Psychology, 25*(6), 779-796. https://doi.org/10.1080/15213269.2022.2070216

Shin, S.-Y. S., Senyung Lee, Yena, Lee, S., & Park, Y. (2023). Exploring rater behaviors on handwritten and typed reading-to-write essays using FACETS. *In Fundamental Considerations in Technology Mediated Language Assessment.* Routledge.

Singer, L. M., & Alexander, P. A. (2017). Reading Across Mediums: Effects of Reading Digital and Print Texts on Comprehension and Calibration. *The Journal of Experimental Education, 85*(1), 155-172. https://doi.org/10.1080/00220973.2016.1143794

Steedle, J., Pashley, P., & Cho, Y. (2020). *Three Studies of Comparability between Paper-Based and Computer-Based Testing for the ACT* (ACT Research & Policy). American College Testing.

Stiegler-Balfour, J. J., Roberts, Z. S., LaChance, A. S., Sahouria, A. M., & Newborough, E. D. (2023). Is reading under print and digital conditions really equivalent? Differences in reading and recall of expository text for higher and lower ability comprehenders. *International Journal of Human-Computer Studies, 176*, 103036. https://doi.org/10.1016/j.ijhcs.2023.103036

Støle, H., Mangen, A., & Schwippert, K. (2020). Assessing children's reading comprehension on paper and screen: A mode-effect study. *Computers & Education, 151*, 103861. https://doi.org/10.1016/j.compedu.2020.103861

Stoukou, I., Papargyris, Y., & Coniam, D. (2023). A comparability study of handwritten versus typed responses in high-stakes English language writing tests. *ELT Forum: Journal of English Language Teaching, 12*(1), Article 1. https://doi.org/10.15294/elt.v12i1.66354

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review, 31*(2), 261-292. https://doi.org/10.1007/s10648-019-09465-5

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review, 10*(3), 251-296. https://doi.org/10.1023/A:1022193728205

Toroujeni, S. M. H. (2022). Computerized testing in reading comprehension skill: Investigating score interchangeability, item review, age and gender stereotypes, ICT literacy and computer attitudes. *Education and Information Technologies, 27*(2), 1771-1810. https://doi.org/10.1007/s10639-021-10584-2

van Groen, M. M., van der Scheer, E., & Keuning, J. (2021). *Wat is er adaptief aan een adaptieve toets?* CitoLab (Stichting Cito). https://cito.nl/media/wjsjiomk/2021_cito_whitepaper_vormen_van_een_adaptieve_toets.pdf

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of Computer-Based and Paper-and-Pencil Testing in K–12 Reading Assessments: A Meta-Analysis of Testing Mode Effects. *Educational and Psychological Measurement, 68*(1), 5-24. https://doi.org/10.1177/0013164407305592

Wästlund, E. (2007). *Experimental studies of human-computer interaction: Working memory and mental workload in complex cognition* [Göteborg University]. https://doi.org/10.1016/j.chb.2004.02.007

Wästlund, E., Reinikka, H., Norlander, T., & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: Psychological and physiological factors. *Computers in Human Behavior, 21*(2), 377-394. https://doi.org/10.1016/j.chb.2004.02.007

Way, D., & Strain-Seymour, E. (2021). *A Framework for Considering Device and Interface Features That May Affect Student Performance on the National Assessment of Educational Progress.* National Assessment of Educational Progress Validity Studies Panel. https://www.air.org/sites/default/files/Framework-for-Considering-Device-and-Interface-Features-NAEP-NVS-Panel-March-2021.pdf

Yu, L., Livingston, S. A., Larkin, K. C., & Bonett, J. (2004). Investigating Differences in Examinee Performance Between Computer-Based and Handwritten Essays. *ETS Research Report Series, 2004*(1), i-21. https://doi.org/10.1002/j.2333-8504.2004.tb01945.x

Yu, W., & Iwashita, N. (2021). Comparison of test performance on paper-based testing (PBT) and computer-based testing (CBT) by English-majored undergraduate students in China. *Language Testing in Asia, 11*(1), 32. https://doi.org/10.1186/s40468-021-00147-0

7 Bijlage: Interviewleidraden

7.1 Toetsdeskundigen

1 Optioneel: toestemmingsformulier laten tekenen als we het interview op willen nemen.
2 Voorstelrondje: wie zijn wij en wat doen we; wie interviewen we en wat doet diegene.
3 Welke ontwikkelingen spelen er al op het gebied van digitalisering/blended toetsen binnen jouw (soort) vak?
4 Wat zijn binnen jouw vakgebied (of breder het soort vak) de voor- en nadelen van de (mogelijke) digitalisering van de examens? Welke problemen en kansen zie je? Welke zijn daarvan het belangrijkst voor jouw vak/soort vak? Wat is de uiteindelijke meerwaarde van digitalisering?
5 Welke ontwikkelingen spelen er juist buiten de digitalisering om die de digitalisering mogelijk zouden bemoeilijken of vergemakkelijken?
6 Wat vind je van de geschetste uitwerking van blended toetsen voor jouw vak/soortgelijke vakken? Zou dat in de praktijk werken? Welke voor- en nadelen zie je?
7 Wat zou je veranderen aan onze schets? Wat zou je behouden?
8 In onze schets van blended examens worden de vragen nog steeds in (digitale) tekstuele vorm aangeboden. Zou het voor jouw (soort) vak ook mogelijk en wenselijk zijn om vragen in film-, animatie- of audiovorm aan te bieden?

7.2 (Oud)docenten

1 Optioneel: toestemmingsformulier laten tekenen als we het interview op willen nemen.
2 Voorstelrondje: wie zijn wij en wat doen we; wie interviewen we en wat doet diegene.
3 Definitie: wat verstaan wij onder blended toetsen?
4 Hoe geef jij blended toetsen vorm in jouw praktijk?
5 Wat vinden je collega's en leerlingen daarvan?
6 Wat vind jij de voor- en nadelen van hoe jij blended toetsen vorm geeft (ook qua leerlingprestaties, leerlingmotivatie, logistiek en tijdbesparing)? Welke problemen lost blended toetsen op voor jou?
7 Wat zijn jouw wensen wat betreft blended toetsen? Welke problemen loop je nog tegen aan? Wat mis jij nu nog om blended toetsen echt (op grote schaal) in te gaan zetten?
8 Hoe zie jij de toekomst van toetsen? Hoe wil jij dat toetsen over 10 jaar werkt? (Op papier, blended, volledig digitaal? Flexibele momenten of vaste? etc.)
9 Wat vind je van het idee om vragen in film-, animatie- of audiovorm aan te bieden?
10 Optioneel: onze voorlopige schets toelichten en vragen wat de geïnterviewde daarvan vindt.

CITO CTE ontwikkelt en adviseert bij wettelijke toetsen

Cito
Amsterdamseweg 13
6814 CM Arnhem
Postbus 1034
6801 MG Arnhem
T (026) 352 11 11
www.cito.nl

Fotografie: Gijs Versteeg