Nakijken met AI

Een verkenning van scenario's voor de centrale eindexamens vo

Anneke Dubbeld, Max van Haastrecht, Jasmijn Mulder, Feline de Wit, Joost Kruis (CitoLab, Stichting Cito)

Visualisaties: Tijmen Poell (CitoLab, Stichting Cito)

februari 2026

Inhoud

Samenvatting

Aanleiding

Het nakijken van de centrale eindexamens (CE's) vergt jaarlijks miljoenen beoordelingen, met hoge eisen aan de betrouwbaarheid van deze beoordelingen; fouten kunnen grote gevolgen hebben voor leerlingen en vertrouwen in het stelsel. Daarnaast is het nakijkwerk tijdsintensief en wordt er een hoge werkdruk ervaren door veel leraren. De uitdagingen op het gebied van consistentie en nakijktijd roepen de vraag op of en hoe Artificial Intelligence (AI) kan worden ingezet bij de beoordeling van de CE's zodat de kwaliteit van de beoordeling omhooggaat, zonder dat de autonomie van de beoordelaar wordt aangetast, en hoe een dergelijke inzet van AI het correctieproces ook efficiënter kan maken.

Methode

Voor drie vakken en schooltypen – wiskunde A (havo), Nederlands (vwo) en biologie (vmbo bb/kb) – zijn scenario's opgesteld met tien verschillende AI-functionaliteiten en niveaus van automatisering. De scenario's variëren van automatische spellings- en grammaticacontrole en clustering van korte open antwoorden tot een scorevoorspelling met uitleg bij langere open antwoorden. De scenario's zijn besproken in vijf focusgroepen met in totaal 26 stakeholders uit de examenketen, waaronder docenten, toetsdeskundigen en vertegenwoordigers van vakverenigingen, CvTE en Inspectie van het Onderwijs. De input is thematisch gecodeerd op vijf thema's, namelijk waardeoordelen of persoonlijke voorkeuren, kwaliteit van de beoordeling, efficiëntie van het correctieproces, autonomie van de beoordelaar en haalbaarheid. Het geheel aan opmerkingen is zowel kwantitatief als kwalitatief geanalyseerd.

Resultaten

De resultaten tonen dat de potentie van AI sterk afhankelijk is van vak en vraagtype. Laag-risico toepassingen scoren overwegend positief: automatische spelling- en grammaticacontrole (Nederlands) en clustering van korte open antwoorden (biologie) kunnen volgens deelnemers aan de focusgroepen bijdragen aan meer efficiëntie en consistentie in het nakijkproces. Voor wiskunde biedt AI mogelijkheden bij het markeren van denkstappen en het doen van scoresuggesties, maar bestaan er zorgen over de accuraatheid van de herkenning van denkstappen en afwijkende oplossingsstrategieën, en over het toepassen van vakspecifieke nakijkregels. Verdergaande toepassingen zoals scorevoorspellingen en zekerheidspercentages roepen verdeeldheid op en geven aanleiding tot vragen over transparantie, interpretatierichtlijnen, autonomie van beoordelaars, extra overlegbelasting en ongelijkheid tussen vroege en late nakijkmomenten.

Conclusies en aanbevelingen

Stakeholders zien mogelijkheden voor AI om de consistentie en efficiëntie in het nakijkproces te verbeteren. Implementatie vraagt echter wel om een gefaseerde aanpak, waarin per vak en vraagtype wordt bekeken wat de mogelijkheden en risico's zijn. Aanbevolen wordt om te starten met toepassingen met een beperkte procedurele rol voor AI (spellingscontrole, clustering, AI-aanvulling van antwoordmodellen), inclusief vakinhoudelijke validatie, alvorens ook onderzoek te doen naar toepassingen waarbij AI-systemen autonomer functioneren. Voor papieren examens zou dan gekeken moeten worden hoe leerlingantwoorden gestructureerd gedigitaliseerd beschikbaar kunnen worden gesteld. Essentieel in het proces zijn heldere governance (ontwikkelaar, beheer, aansprakelijkheid), transparantie en uitlegbaarheid van modellen en waarborging tegen bias. Pilots dienen technische validatie, onafhankelijke toetsing en draagvlakmetingen te combineren. Eerste verkennende stappen worden reeds gezet in het

programma *Digitalisering centrale examens*, waarin in de periode 2025-2029 onderzocht wordt welke mogelijkheden verdergaande digitalisering kan bieden voor de CE's in het voortgezet onderwijs. Samengevat kan AI, middels een voorzichtige, goed gecontroleerde en vakgerichte aanpak, een waardevolle ondersteuning bieden bij het nakijken van eindexamens, zonder dat de autonomie van de beoordelaar wordt aangetast of het vertrouwen in het examenstelsel wordt ondermijnd.

1. Aanleiding

Elk jaar kijken docenten in het voortgezet onderwijs miljoenen leerlingantwoorden na bij de centrale eindexamens (CE's). Er zijn hoge standaarden voor de betrouwbaarheid van beoordelingen van de CE's aangezien er voor leerlingen veel van afhangt. Fouten in de beoordeling kunnen grote consequenties hebben, niet alleen voor de leerling, maar ook voor het bredere vertrouwen in de kwaliteit van de examinering. Redenen voor inconsistenties in beoordelingen kunnen veelvuldig zijn en kunnen bijvoorbeeld te maken hebben met onduidelijkheden in het correctievoorschrift. Vanwege de vele papieren examens en het ontbreken van digitale scoringsdata, is het nog niet mogelijk om conclusies te trekken over de mate en aard van inconsistenties in het nakijken van eindexamens. Echter moeten we in de context van de CE's, waar kleine fouten grote consequenties kunnen hebben voor leerlingen, er altijd naar streven om fouten te voorkomen waar dat kan. Dit roept de vraag op of het middels automatisering mogelijk is om te zorgen voor consistentere beoordelingen.

Daarnaast kost het nakijken van de miljoenen leerlingantwoorden veel tijd. Idealiter zou het streven naar consistentere beoordelingen gepaard gaan met een efficiëntieslag. Tegelijkertijd moeten we er ook waakzaam voor zijn dat we met verregaande automatisering niet de autonomie van de docent ondermijnen. In dit onderzoek proberen we antwoorden te vinden op de vraag hoe we zorg kunnen dragen voor kwalitatief hoogwaardige beoordelingen, terwijl we de autonomie van de docent en efficiëntie in het proces van nakijken borgen en versterken. In deze aanleiding zullen we verder ingaan op de werkdruk en nakijktijd van leraren, die het belang van dit vraagstuk onderstrepen. Daarna gaan we verder in op de onderzoeksvragen en de opzet van het onderzoek.

1.1 Werkdruk en nakijktijd

Docenten in het voortgezet onderwijs zijn op bepaalde vlakken zeer positief over hun werk. Zo vindt ongeveer 90% dat de voordelen van leraar zijn duidelijk opwegen tegen de nadelen, zou meer dan 80% opnieuw voor deze baan kiezen en is meer dan 90% tevreden met hun baan (Buisman et al., 2025). Tegelijkertijd is deze tevredenheid eerder een signaal van de positieve grondhouding en de passie van docenten voor hun werk, dan dat het een reflectie is van een uitsluitend positief werkklimaat. Het aanhoudende lerarentekort en de hiermee gerelateerde problematiek rondom werkdruk zijn voorbeelden van de systemische uitdagingen die er wel degelijk zijn.

Het lerarentekort vormt een urgent probleem, niet alleen in het primair onderwijs, maar ook in het voortgezet onderwijs (Adriaens et al., 2024). Het ministerie van Onderwijs, Cultuur en Wetenschap (OCW) heeft daarbij becijferd dat de grootste tekorten qua aantallen FTE te zien zijn bij de vakken Nederlands en Wiskunde.[1] Deze ontwikkelingen vergroten de druk op scholen en docenten en maken het noodzakelijk om met innovatieve oplossingen te komen, onder andere om het werk van docenten te ondersteunen en efficiënter in te richten.

Van alle leraren in het vo ervaart 45% een hoge werkdruk. Dit ligt lager dan in het po (49%), maar hoger dan in het mbo (37%) en veel hoger dan het landelijk gemiddelde voor werkenden in Nederland (31%).[2] Kenmerkend aan het werk van docenten in het primair en voortgezet onderwijs is dat het vaak gepaard gaat met hoge taakeisen en een lage mate van autonomie (Van den Heuvel & De Vroome, 2023).

[1] https://www.ocwincijfers.nl/themas/arbeidsmarkt-en-leraren/trendrapportage-arbeidsmarkt/tekorten-vo/landelijke-tekorten-vo-naar-vak

[2] https://www.ocwincijfers.nl/indicatoren-funderend-onderwijs/de-mate-waarin-onderwijspersoneel-werkdruk-ervaart-in-het-po-vo-en-mbo

De combinatie van hoge taakeisen en beperkte autonomie, in de literatuur gekarakteriseerd als een 'high strain job', wordt gezien als een van de belangrijkste indicatoren van werkdruk. Het is dan ook niet verrassend dat het percentage burn-out klachten onder docenten in het voortgezet onderwijs substantieel hoger ligt dan bij andere banen in Nederland (Van den Heuvel & De Vroome, 2023). Dat werkdruk de belangrijkste vertrekreden is voor leraren (Van Casteren et al., 2023) geeft aan hoe belangrijk het is om oplossingen te vinden voor deze problematiek.

Grotere klassen zijn een logisch gevolg van de huidige lerarentekorten in het voortgezet onderwijs, en juist docenten van grotere klassen geven aan meer stress te ervaren van de hoeveelheid nakijkwerk en administratie (Adriaens et al., 2024). Uit het TALIS 2024 rapport blijkt dat 34% van de docenten aangeeft sterke werkstress te ervaren als gevolg van te veel nakijkwerk, en dat de grootste stressbron komt van te veel administratieve werkzaamheden (50%). Een voorname bron van nakijkwerk voor docenten in examenklassen van het voortgezet onderwijs komt van het beoordelen van CE's. Uit onderzoek naar de correctielast bij de CE's blijkt dat zowel het type examenvragen als het aantal leerlingen de nakijktijd beïnvloeden (Essen et al., 2023). Met name het vak Nederlands heeft voor alle schooltypen een hoge correctielast, doordat de nakijktijd per leerling relatief hoog ligt en er per docent veel leerlingen nagekeken moeten worden. Voor Nederlands op het vwo betekent dit concreet dat er met de 1.246 betrokken docenten in 2024[3] naar schatting 44.482 uur aan nakijktijd is besteed. Bijna 45.000 uur dus voor één vak in één stroming. Het moge duidelijk zijn dat een efficiënter en beter ondersteund nakijkproces veel potentie heeft om de werkdruk van docenten te verlagen en zo ruimte kan bieden voor taken die juist bijdragen aan werkplezier.

1.2 AI als oplossing?

De uitdagingen op het gebied van consistentie en nakijktijd roepen de vraag op of er oplossingen mogelijk zijn. Een veel genoemde oplossingsrichting in de laatste paar jaar is het integreren van Artificial Intelligence (AI) systemen bij het beoordelen van examens. AI zou in sommige gevallen een oplossing kunnen zijn, al zijn sommige oplossingsrichtingen alleen maar mogelijk bij gedigitaliseerde examens. Het programma digitalisering centrale examens,[4] waarbij de partners in de examenketen (OCW, CvTE, Cito, DUO) in de periode 2025-2029 onderzoeken waar er kansen liggen op het vlak van digitalisering, biedt in die zin een mooie opening om te kijken waar AI van meerwaarde kan zijn bij het nakijken. De actualisatie van de examenprogramma's in het kader van de curriculumvernieuwing[5] maakt het nog urgenter om docenten te ondersteunen bij het nakijken van examens, aangezien ze te maken gaan krijgen met nieuwe vraagvormen. Zo ligt er bij moderne vreemde talen het plan om vragen te stellen die opener zijn van aard en dus uitdagender om consistent na te kijken.

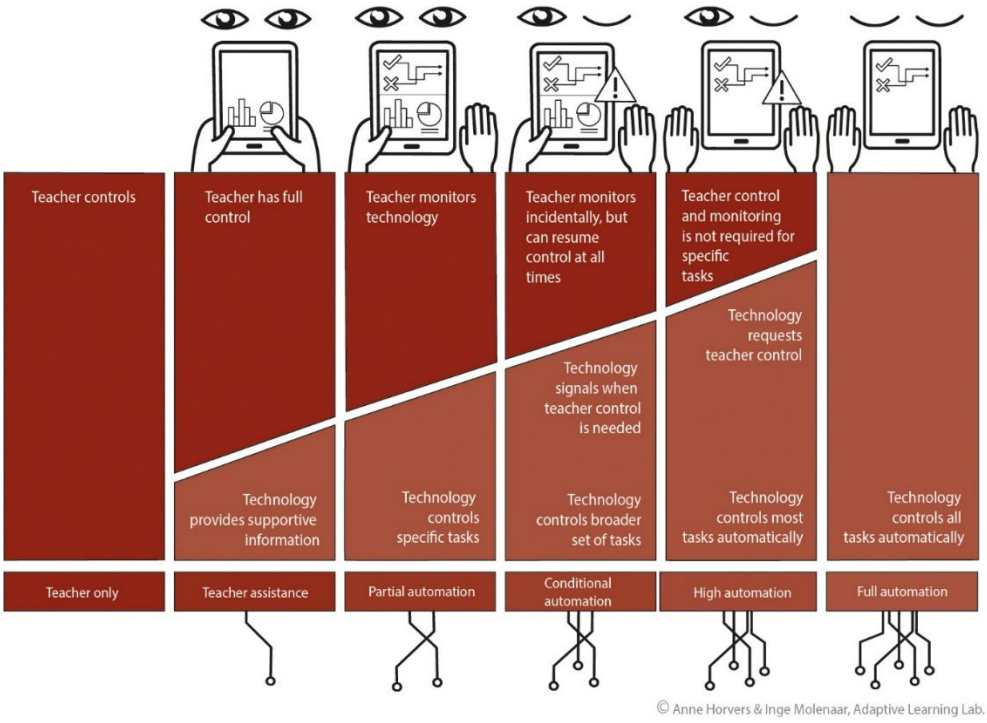
We moeten echter waakzaam zijn bij het toepassen van AI-automatisering om docenten te ontlasten en voor consistentere beoordelingen te zorgen. Hoe meer je automatiseert, hoe meer winst er potentieel te behalen is qua efficiëntie en consistentie, echter ook hoe minder autonomie er ligt bij de docent. Zoals bleek uit de eerdere bespreking van werkdruk, kan een gebrek aan autonomie weer een negatieve invloed hebben op de ervaren werkdruk. Daarnaast is het vertrouwen van leerlingen en ouders in een juiste beslissing bij de eindexamens gebaseerd op het oordeel van docenten, en (nog) niet op het oordeel van AI-algoritmes. We riskeren dus niet alleen het werkplezier van de docent door verregaand te automatiseren, maar ook het vertrouwen in toetsing, en in bredere zin ons onderwijs. Figuur 1 geeft met een schaal weer wat voor vormen er te kiezen zijn als het gaat om automatisering met technologie in het onderwijs, van de

[3] https://www.examenblad.nl/system/files/exam-document/2024-10/24101_quickscan_vwo-nederlands-2024-tijdvak-1.pdf
[4] https://www.cvte.nl/actueel/nieuws/2025/07/07/onderzoek-naar-verdere-digitalisering-centrale-examinering-in-het-voortgezet-onderwijs
[5] https://www.slo.nl/thema/meer/actualisatie-kerndoelen-examenprogramma/actualisatie-examenprogramma/

docent volledig in controle aan de linkerkant, tot volledig geautomatiseerd aan de rechterkant (Molenaar, 2022). Met de eerdere opmerking over vertrouwen in het achterhoofd is het logische pad om op zoek te gaan naar oplossingen waarmee we de kwaliteit van beoordelingen kunnen verbeteren, zonder daarbij de autonomie van docenten te veel aan te tasten. Dat vraagt om een verantwoorde manier van omgaan met AI, die in lijn is met de AI-verordening, de AVG, en in bredere zin de waarden die wij in Nederland hebben als het aankomt op integratie van algoritmes in high-stakes omgevingen.

Figuur 1

Schaal van verschillende niveaus van automatisering in het onderwijs (Molenaar, 2022)



2. Opzet en onderzoeksvragen

Uit het vorige hoofdstuk blijkt dat er aan de kant van efficiëntie winst te behalen valt in de context van de CE's. Ook is beschreven dat we moeten streven naar zo betrouwbaar mogelijke beoordelingen van de high-stakes CE's en inconsistenties in het nakijken zoveel mogelijk moeten voorkomen. Daarbij is het zaak om, in samenspraak met docenten, de balans te blijven zoeken tussen automatisering en autonomie. AI kan een middel zijn tot het doel, maar we hebben op dit moment nog niet een compleet beeld van waar AI waarde kan toevoegen in het nakijkproces. Daarom staat de volgende onderzoeksvraag centraal:

Hoofdvraag: Op welke wijze kan AI worden ingezet bij de beoordeling van centrale examens zodat de kwaliteit van de beoordeling omhooggaat, zonder dat de autonomie van de beoordelaar wordt aangetast? Hoe kan een dergelijke inzet van AI het correctieproces ook efficiënter te maken?

Om deze hoofdvraag te kunnen beantwoorden, zullen we in dit rapport scenario's uitwerken om een breed palet aan mogelijkheden te schetsen voor hoe AI een rol kan spelen bij beoordelingen van de CE's. Deze scenario's zullen variëren over verschillende vakken en soorten vragen, maar zullen ook variëren in de verschillende niveaus van automatisering van Molenaar (2022). De scenario's worden in een volgend hoofdstuk uitgebreid beschreven, inclusief een onderbouwing voor de keuzes die gemaakt zijn omtrent vakken, soorten vragen en niveau van automatisering, en geven antwoord op de volgende subvragen:

Subvraag 1a: Welke scenario's zijn er bij verschillende vakken/vragen mogelijk voor het gebruik van AI bij het nakijken?

Subvraag 1b: Wat zijn de mogelijkheden die deze scenario's bieden om het nakijkproces efficiënter en consistenter te maken? Welke verschillen zien we hier over vakken en soorten vragen heen?

We zouden uiteraard zelf inschattingen kunnen maken over hoe bepaalde AI-interventies zouden uitpakken op het vlak van de consistentie en efficiëntie van beoordelingen en de gevoelde autonomie van docenten. Voor dit onderzoek kiezen we er echter bewust voor om de stem van betrokkenen bij het nakijkproces van de CE's expliciet mee te nemen middels focusgroepen. Aan de kant van docenten zijn de betrokkenen niet alleen de docenten die het nakijkwerk verrichten, maar ook de leden van Cito constructiegroepen, en de leden van CvTE vaststellingscommissies. Echter zijn er naast docenten nog veel meer mensen en organisaties betrokken bij het zorgen voor kwalitatief hoogwaardige examens. De partners in de examenketen (OCW, CvTE, Cito, DUO), de inspectie, SLO, LAKS, de VO-Raad en PLEXS zijn allen op hun eigen manier betrokken bij de examens. We pogen in dit onderzoek om een brede vertegenwoordiging uit al deze gelederen mee te laten doen aan de focusgroepen.

We zullen in onze focusgroepen deze stakeholders bevragen op hun inschatting van de impact van verschillende scenario's voor nakijken met AI. Dit doen we voor criteria die al eerder besproken zijn zoals consistentie, efficiëntie en autonomie. Daarnaast streven we er in dit onderzoek naar om scenario's te formuleren die reëel en beleidsmatig relevant zijn. Om te toetsen of we hierin zijn geslaagd nemen we ook de haalbaarheid van scenario's op als apart criterium. Met een haalbaar scenario waarmee we de autonomie van de docent waarborgen en zorgen voor een consistenter en efficiënter proces zijn we er echter nog niet. Met scenario's waar we volledige automatisering uitsluiten blijft er altijd een rol over voor de docent, en dus is het essentieel dat docenten in brede zin de AI-interventie als waardevol zin voor zichzelf en hun leerlingen. Het laatste criterium dat we toepassen op onze scenario's is dan ook de waarde van de

interventie vanuit het oogpunt van de focusgroep participanten. Al deze criteria tezamen resulteren in de volgende onderzoeksvraag:

Subvraag 2: Hoe kijken stakeholders in de examenketen naar het gebruik van AI bij het nakijken van centrale examens, specifiek op het gebied van kwaliteit (consistentie), efficiëntie, autonomie, waarde en haalbaarheid?

3. Achtergrond

3.1 AI in de (Nederlandse) onderwijscontext en het nakijkproces

Om de mogelijkheden van de inzet van AI in het nakijkproces te verkennen, is het allereerst nodig om de term te definiëren, aangezien de term vaak gebruikt wordt, maar niet altijd op dezelfde manier wordt begrepen. Het begrip is lastig precies te definiëren, omdat het sterk afhangt van de context waarin het wordt gebruikt (Krafft et al., 2020). Beleidsmatig en juridisch gezien is de meest relevante definitie van AI in de Europese context de definitie die in de AI-verordening wordt gehanteerd. Binnen de AI-verordening wordt een AI-systeem gekarakteriseerd als: “een op een machine gebaseerd systeem dat is ontworpen om met verschillende niveaus van autonomie te werken en dat na het inzetten ervan aanpassingsvermogen kan vertonen, en dat, voor expliciete of impliciete doelstellingen, uit de ontvangen input afleidt hoe output te genereren zoals voorspellingen, inhoud, aanbevelingen of beslissingen die van invloed kunnen zijn op fysieke of virtuele omgevingen” (Europees Parlement, 2024, Artikel 3 lid 1).

Met de snelle ontwikkeling van AI wordt ook in het onderwijs gekeken naar toepassingen van deze technologie, waarbij in dit rapport wordt gefocust op het automatisch nakijken van toetsen en examens. Verschillende onderzoeken laten zien dat AI op diverse onderdelen van het nakijkproces effectief kan worden ingezet. Zo kunnen klassieke algoritmen spellingfouten detecteren en corrigeren (Kukich, 1992). Bij korte open antwoorden maken technieken zoals het clusteren van vergelijkbare antwoorden het mogelijk meerdere antwoorden tegelijk te beoordelen (NOLAI, 2025). Dit verlaagt de werkdruk voor docenten (Basu et al., 2013; Horbach et al., 2014; Mieskes & Padó, 2018). AI kan bovendien scores voorspellen, al dan niet gecombineerd met korte verklaringen of gerichte feedback, wat aansluit bij inzichten uit onderzoek naar intelligent tutoring systems (Nye et al., 2014). Voor meer geavanceerde toepassingen, zoals wiskundeopgaven, laten recente studies zien dat multimodale AI-modellen zelfs handgeschreven antwoorden kunnen scoren, hoewel de nauwkeurigheid nog niet volledig vergelijkbaar is met menselijke beoordelaars (Caraeni et al., 2025). AI kan ook denkstappen herkennen of markeren, bijvoorbeeld via chain-of-thoughtbenaderingen, waardoor inzicht ontstaat in het redeneringsproces van studenten (Wei et al., 2023). Verder is het mogelijk om belangrijke woorden uit het correctievoorschrift te highlighten, wat docenten helpt te begrijpen waarom een bepaalde score is toegekend (Del Gobbo, 2023).

Onderzoek wijst uit dat AI hierdoor belangrijke voordelen kan bieden op het gebied van efficiëntie en consistentie. Zo toont een recente studie dat AI gemiddeld gesproken minder beoordelingsfouten maakt dan menselijke beoordelaars bij korte open antwoorden, met een afname van 44% in het aantal fouten (Gobrecht et al., 2024). Tegelijkertijd blijkt dat bij essays of langere open antwoorden het verschil in beoordelingskwaliteit kleiner wordt, en dat AI in sommige gevallen zelfs minder nauwkeurig scoort dan mensen (Flodén, 2024; Wetzler et al., 2024). Dit kan worden verklaard doordat langere antwoorden vaak complexere taal, redeneringen en nuances bevatten, waarbij menselijke beoordelaars beter in staat zijn context en subtiele details mee te nemen.

Hoewel er uit onderzoek dus nog geen eenduidige conclusies zijn te trekken over de betrouwbaarheid en validiteit van de inzet van het AI in het onderwijs- en nakijkproces, blijkt dat er in de praktijk al volop wordt geëxperimenteerd en gebruikgemaakt van AI. Uit vijfjaarlijks onderzoek van de Organisatie voor Economische Samenwerking en Ontwikkeling (OESO) blijkt dat een derde van de docenten in 55 onderzochte landen gebruikmaakt van AI, waarvan een kwart het gebruikt voor nakijkwerk

(OECD, 2025). Het percentage docenten dat AI gebruikt voor nakijkwerk verschilt per land, met 85% van de docenten in Oezbekistan als uitschieter naar boven. In Nederland gebruikt een derde van de docenten AI voor nakijkwerk.

Ook commerciële toetsplatforms en -applicaties hebben AI reeds in hun omgeving verwerkt, bijvoorbeeld middels een nakijkassistent die docenten helpt bij het beoordelen van open toetsvragen door de open antwoorden te analyseren en automatisch een score met een toelichting toe te kennen. Binnen Cito wordt ook in uiteenlopende projecten gewerkt aan kennis over AI en een doordachte en verantwoorde wijze van inzet binnen de toetscyclus. Binnen een aantal projecten wordt een prototype (digitale tool) ontwikkeld, in samenwerking met de onderzoeks- en innovatieafdeling CitoLab (Stichting Cito). Een voorbeeld van een prototype is Schrijfblik[6], een AI-gedreven schrijfomgeving, die helpt bij het beoordelen van zowel het eindproduct als het proces van schrijven (Poell et al., 2025). Een ander voorbeeld is Checkmate[7], een slimme nakijkoplossing die helpt bij het nakijken van (korte) open antwoorden, door vergelijkbare antwoorden te clusteren en een score voor te stellen.

3.2 Digitalisering eindexamens en het nakijken ervan met AI

In de context van het nakijken van eindexamens wordt AI momenteel niet ingezet, terwijl er juist kansen liggen op het gebied van efficiëntie en consistentie van het nakijkproces. Echter, zoals in de aanleiding reeds beschreven, hangt de inzet van AI bij het nakijken van CE's nauw samen met ontwikkelingen op het gebied van digitalisering en curriculumvernieuwing van verschillende schoolvakken, evenals ontwikkelingen op het gebied van AI zelf. In dit onderdeel worden enkele van deze voorwaarden en ontwikkelingen beschreven.

Om gebruik te kunnen maken van AI in het nakijkproces is het digitaal beschikbaar hebben van leerlingantwoorden vereist. Nu worden CE's voor vmbo bb en kb digitaal en flexibel[8] afgenomen in Facet (Examenblad, 2025). In het programma *Digitalisering centrale examens* wordt in de periode 2025-2029 onderzocht welke mogelijkheden verdergaande digitalisering kan bieden voor de CE's in het voortgezet onderwijs (CvTE, 2025). Het programma is een samenwerking tussen CvTE, Stichting Cito, DUO en is geïnitieerd door OCW. Digitalisering wordt in dit programma als *middel* gezien om tegemoet te komen aan behoeften uit het onderwijsveld. Naast dat digitalisering kansen biedt om examens flexibeler, toegankelijker en persoonlijker te maken voor iedereen, zou AI een oplossing kunnen zijn voor het verhogen van de efficiëntie van en consistentie binnen het nakijkproces van centrale digitale examens. In verschillende projecten binnen het programma wordt onderzocht waar AI van meerwaarde kan zijn voor de CE's. Hierbij wordt breder gekeken dan alleen het nakijken, en wordt ook onderzocht wat de rol van AI zou kunnen zijn bij bijvoorbeeld constructie en psychometrische analyse.

In het kader van constructieve afstemming is het belangrijk om voortdurend rekening te houden met de eigenschappen van de verschillende schoolvakken en de bijbehorende vak- en curriculumvernieuwingen (SLO, 2025). Veel huidige vakvernieuwingen richten zich op competenties zoals kritisch denken, bronnengebruik en probleemoplossend vermogen, waarbij examenvormen geleidelijk veranderen van meer reproductie naar meer open opdrachten en authentieke taken. Zo is er binnen taalvakken zoals Nederlands en moderne vreemde talen in het vernieuwde curriculum een toenemende aandacht voor schrijf- en spreekvaardigheden, waarbij creativiteit, argumentatiekracht, interpretatie en communicatieve

[6] https://cito.nl/onderzoek-innovatie/portfolio/innovatieve-prototypes/vaardigheden-voor-een-nieuwe-wereld/schrijfblik/
[7] https://cito.nl/onderzoek-innovatie/portfolio/innovatieve-prototypes/de-waarde-van-data/checkmate/
[8] Flexibel betekent dat scholen zelf kunnen bepalen wanneer ze de examens afnemen, ergens tussen begin april en half juli.

vaardigheden van groot belang zijn. Dit type open vragen maakt consistent nakijken voor menselijke beoordelaars lastiger. Tegelijkertijd wordt het geautomatiseerd nakijken ook complexer, al zou AI al een rol kunnen spelen in het automatisch beoordelen van spelling/grammatica en voor structurele aspecten zoals tekstlengte. Bij exacte vakken zoals wiskunde ligt de nadruk in de curriculumvernieuwing onder andere op de toepassing van wiskunde in realistische contexten, waarbij examens veel opgaven bevatten met uitwerkruimte en redeneringsvragen. Bij gesloten antwoorden en geformaliseerde oplossingspaden zou automatisch nakijken kunnen werken, terwijl voor opgaven met redeneringsstappen de automatische beoordeling complexer kan zijn. Curriculumvernieuwingen in vakken als geschiedenis, maatschappijleer en kunst leggen de nadruk op interpretatieve vaardigheden, kritische bronbeoordeling en creatieve expressie. Deze eigenschappen zijn lastig te kwantificeren in eenduidige beoordelingscriteria, hoewel AI zou kunnen helpen bij het signaleren van sleutelbegrippen, consistentie en formele aspecten zoals bronvermeldingen.

AI met al zorgt de verkenning van de mogelijkheden van digitalisering in de examenketen voor mogelijkheden op het gebied van nakijken met AI, maar de mogelijkheden en risico's verschillen per vak en soort opgave. Omdat dit een verkennend onderzoek betreft, is in het huidige onderzoek gekozen voor het concreet toespitsen op enkele schoolvakken en type opgaven voor het toepassen van nakijkoplossingen met AI, waarmee is geprobeerd een zo breed mogelijk palet aan vakken en type opgaven te dekken.

3.3 Technische, ethische en praktische randvoorwaarden

De verkenning van de mogelijkheden van de integratie van AI in het nakijkproces van CE's vraagt ook om rekening te houden met technische, ethische en praktische randvoorwaarden. Naast de reeds genoemde digitalisering van examens, is er behoefte aan voldoende representatieve en vakinhoudelijke getrainde AI-modellen, transparantie over beoordelingsalgoritmen, en waarborging van privacy. Dit hangt samen met de AI-verordening, die in artikel 56 het volgende vermeldt over nakijken:

> AI-systemen die in het onderwijs of voor beroepsopleidingen worden gebruikt, met name voor het bepalen van toegang of toelating, voor het toewijzen van personen aan instellingen voor onderwijs of beroepsopleidingen of programma's op alle niveaus, voor het evalueren van leerresultaten van personen, voor de beoordeling van het passende onderwijsniveau voor een persoon en het uitoefenen van wezenlijke invloed op het niveau van onderwijs en opleiding dat aan die persoon zal worden verstrekt of waartoe die persoon toegang zal kunnen krijgen, of voor het monitoren en opsporen van ongeoorloofd gedrag van studenten tijdens tests, moeten echter als **AI-systemen met een hoog risico** worden geclassificeerd, aangezien ze bepalend kunnen zijn voor de onderwijs- en beroepsloopbaan van die persoon en derhalve invloed kunnen hebben op zijn vermogen om in zijn levensonderhoud te voorzien. (Bureau voor publicaties van de Europese Unie, 2024)

De technische, ethische en praktische randvoorwaarden staan in direct verband met het algemene draagvlak voor het nakijken met AI, waar in het huidige onderzoek op wordt gefocust. Waar sommige docenten en stakeholders in de examenketen welwillend en positief zijn ten opzichte van mogelijkheden en ontwikkelingen, blijkt ook dat er bijvoorbeeld docenten zijn die AI niet willen inzetten voor nakijkwerk: "Nakijken kan ik zelf veel beter"[9]. Dit spanningsveld heeft alles te maken met de mate van autonomie die bij de docent versus AI ligt, zoals weergegeven in Figuur 1 met zes lagen van automatisering in een onderwijscontext. In het huidige onderzoek verkennen we welke taken AI kan overnemen en wat dat

[9] https://www.rtl.nl/nieuws/binnenland/artikel/5532212/eenderde-docenten-gebruikt-ai-onderwijs#:~:text=Een%20derde%20van%20de%20Nederlandse,kan%20ik%20zelf%20veel%20beter.%22

betekent voor de rol van de docent, maar nog belangrijker: in hoeverre is er draagvlak voor dergelijke automatisering?

Samenvattend wordt er in het huidige onderzoek, naast de focus op mogelijke verbeteringen in zowel consistentie als efficiëntie in het nakijkproces, zo veel als mogelijk ook rekening gehouden met de afstemming op vakinhoudelijke vernieuwingen, technische, praktische en ethische voorwaarden en het vertrouwen van alle betrokkenen.

4. Scenario's

4.1 Achtergrond en ontwikkeling van de scenario's

Zoals in de opzet beschreven, zijn er voor de focusgroepen scenario's ontwikkeld waarin AI-oplossingen worden ingezet. De scenario's zijn ontwikkeld om tijdens de focusgroepen een concreet beeld te kunnen schetsen aan de deelnemers hoe AI-toepassingen bij het nakijken van CE's eruit kunnen zien. Op deze manier kunnen de deelnemers hun meningen, verwachtingen en zorgen over dergelijke toepassingen gericht formuleren. Dergelijke visuele hulpmiddelen stimuleren namelijk rijkere en diepgaandere reflecties, omdat ze het gesprek richting geven, ideeën opwekken en het gemakkelijker maken om verschillende perspectieven te begrijpen (Page et al., 2022). Bovendien bevorderen deze scenario's de gezamenlijke kennisvorming over de mogelijkheden en grenzen van AI in het nakijkproces.

We hebben scenario's ontwikkeld waarvan het redelijkerwijs te verwachten is dat het op de lange termijn haalbaar zou zijn om ze in te voeren. Daarom werken we alleen met scenario's waar een mens, vaak de beoordelende docent, nog in controle is. Dit sluit niet alleen aan bij de visie op het belang van de rol van de docent in het nakijkproces, maar sluit ook aan bij eisen die vanuit de AI-verordening worden gesteld aan hoog risico AI-systemen die ingezet worden bij high-stakes toetsing. We werken dus niet met scenario's in de laatste stap van Molenaar (2022) van volledige automatisering. Verder variëren de scenario's over schooltype (vmbo, havo, vwo), vak en vraagtype, namelijk:

1. Wiskunde A - havo
2. Nederlands - vwo
3. Biologie - basis- en kaderberoepsgerichte leerweg (bb/kb)

Op dit moment gebruiken alleen vmbo bb en vmbo kb digitale afnames van examens op grote schaal, en worden de examens van vmbo gl/tl, havo en vwo nog op papier afgenomen. Aangezien de inzet van AI bij het nakijken vaak vereist dat er een digitale omgeving is waar de docent het nakijkproces doorloopt, schetsen we voor de scenario's van vakken waar nu nog geen digitale examens zijn een toekomstbeeld waarbij dit wel zo zou zijn. Gegeven de huidige verkenning naar de mogelijkheden voor verdere digitalisering in de examenketen binnen het programma Digitalisering Centrale Examens (2025-2029) zien we dit als een beleidsmatig relevant toekomstbeeld.

De variatie van scenario's over verschillende vakken is gekozen omdat er over vakken heen inhoudelijk andere mogelijkheden zijn voor de inzet van AI. Zo zijn er bij examens waar leerlingen denkstappen moeten zetten, zoals de examens wiskunde, mogelijkheden voor het herkennen van misvattingen in denkstappen. Bij het examen Nederlands is die mogelijkheid er weer niet, maar zijn er wel mogelijkheden om met AI de spelling- en grammaticacontrole efficiënter en consistenter te laten verlopen. De laatste belangrijke variabele in de scenario's is het vraagtype. Zo worden multiplechoicevragen nu al automatisch nagekeken voor de digitale examens vmbo bb en vmbo kb. Voor korte open antwoord vragen van een paar woorden zijn er hele andere mogelijkheden en uitdagingen bij het automatiseren van het nakijken dan bij langere open antwoord vragen van een paar zinnen of zelfs een hele brief. En voor open vragen bij de exacte vakken waar niet alleen taal, maar ook formules en grafieken een rol spelen, zijn er weer andere mogelijkheden. Deze diversiteit aan schooltypen, vakken en vraagtypen biedt dus uiteenlopende mogelijkheden voor het verkennen van AI-toepassingen bij het nakijken van CE's.

De ontwikkeling van de scenario's vond plaats via een iteratief proces. De eerste versies zijn opgesteld op basis van literatuuronderzoek en afgeleid van bestaande AI-nakijkprototypes. Deze concepten zijn vervolgens voorgelegd aan toetsdeskundigen van Stichting Cito van de betreffende

vakken. Vanuit hun expertise gaven zij feedback op zowel de voorgestelde toepassingen als op mogelijke aanvullende inzet van AI binnen het nakijkproces.

Op basis van de ontvangen input zijn de scenario's verder verfijnd, verduidelijkt en contextueel aangepast. Zo ontstonden schetsen die abstracte AI-mogelijkheden vertalen naar concrete, toepassingsgerichte voorbeelden voor de CE's. De schetsen bieden deelnemers een duidelijke basis om hun gedachten en reacties te delen en vormen het vertrekpunt voor de focusgroepgesprekken. In de volgende sectie worden de scenario's per vak toegelicht, inclusief de specifieke AI-toepassingen.

4.2 Beschrijving scenario's

4.2.1 Wiskunde A Havo

Beschrijving examen

Het examen wiskunde A voor havo is een schriftelijk examen waarbij leerlingen 3 uur de tijd krijgen om het examen te maken. Het examen bestaat uit open vragen, waarbij de leerlingen hun werkwijze en antwoorden duidelijk moeten presenteren. Bij de meeste vragen moeten er meerdere denkstappen gemaakt worden, welke duidelijk in het antwoord terug moet komen.

Het examen heeft betrekking op vaardigheden zoals het toepassen van wiskundige technieken, logisch redeneren en het zorgvuldig weergeven van berekeningen. Daarnaast wordt aandacht besteed aan algebra en rekenen, waaronder het werken met algebraïsche uitdrukkingen en het oplossen van vergelijkingen. Een belangrijk onderdeel van het examen is het beschrijven en analyseren van verbanden met behulp van grafieken en formules, zowel lineair als niet-lineair. Ook statistiek speelt een rol, waarbij leerlingen onder andere gegevens moeten interpreteren, verwerken en beoordelen, en grafieken en verdelingen analyseren. Het examen van 2024 bestond uit 21 open vragen, waarvoor er maximaal 76 punten gehaald konden worden.

Correctielast

De gemiddelde correctielast van het wiskunde A examen voor havo is ongeveer 20 uur. Voor vwo is de nakijktijd vrijwel gelijk (ook ongeveer 20 uur). Voor wiskunde vmbo gl/tl ligt de correctielast iets lager: een docent heeft gemiddeld 12 uur nodig om een wiskunde examen voor vmbo gl/tl na te kijken. Het wiskunde examen is gemiddeld intensief qua correctielast (Essen et al., 2023).

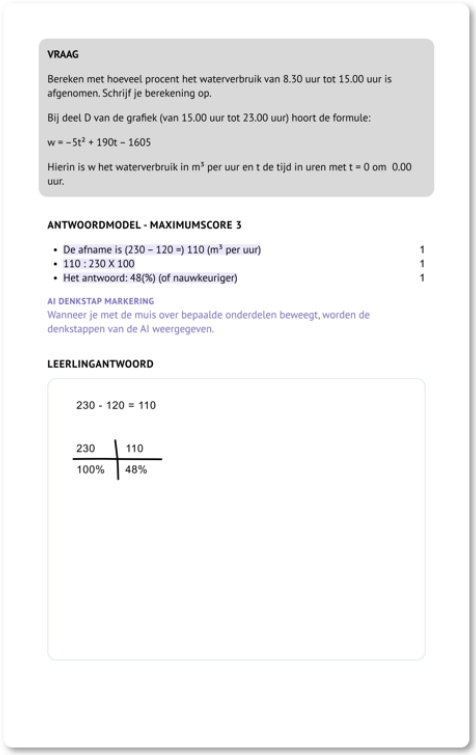
Uit de gesprekken die gevoerd zijn met toetsdeskundigen wiskunde bleek dat volledig juiste of volledig onjuiste antwoorden bij het wiskunde examen relatief gemakkelijk na te kijken en te scoren zijn. De uitdagingen in het nakijken van een wiskunde examen liggen voornamelijk in het bepalen van misvattingen in de verschillende denkstappen die een leerling moet maken. Een vraag kan hierdoor namelijk gedeeltelijk juist zijn: als er een reken- of denkfout gemaakt is in een eerdere stap in de berekening, maar de daaropvolgende denkstappen in de berekening wel correct zijn uitgevoerd, worden voor de laatste denkstappen (meestal) wel scorepunten toegekend.

Beschrijving scenario

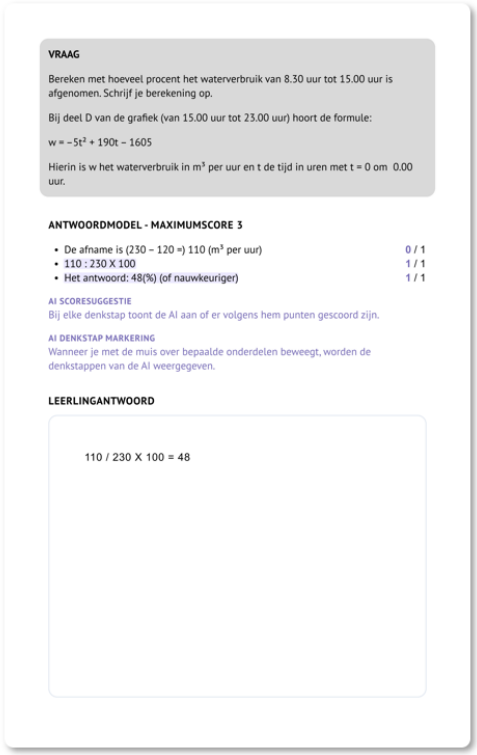
Het huidige wiskunde A examen wordt op papier afgenomen. Voor dit scenario doen we echter alsof het examen digitaal afgenomen wordt. AI kan in dit scenario twee rollen vervullen: het markeren van de verschillende denkstappen in het antwoord of het markeren van denkstappen en daarbij automatisch een scoresuggestie geven (zie Figuur 2) (Wei et al., 2023). De markering wordt weergegeven als de docent met de cursor beweegt over ofwel het onderdeel in het antwoordmodel, ofwel een onderdeel van het leerlingantwoord. De schetsen van deze opties, inclusief animaties, zijn ook te bekijken via https://BOO-

Nakijken-AI-figma.web.app, waarbij animatie 1.0 de kale examenopdracht weergeeft en de nummering 1.1 en 1.2 correspondeert met de bijbehorende optie.

Figuur 2
Visuele weergave opties scenario 1 – wiskunde A havo



Optie 1.1 – markeringen denkstappen



Optie 2.1 – markeringen denkstappen en automatische scoresuggestie

De mogelijkheden voor nakijken met AI zoals hierboven beschreven vallen onder de noemer van docent assistentie in het model van Molenaar (2022). Hierbij heeft de docent volledige controle en geeft AI alleen ondersteunende informatie om de docent bij keuzes te ondersteunen.

4.2.2 Nederlands vwo

Beschrijving examen

Het examen Nederlands voor vwo is een schriftelijk examen waar leerlingen 3 uur de tijd voor hebben. Leerlingen ontvangen het examenblad en een bijlage met teksten. Leerlingen worden geïnstrueerd om antwoorden in correct Nederlands te geven en niet meer antwoorden te geven dan worden gevraagd. Als uitleg hierbij staat: “Als er bijvoorbeeld één zin wordt gevraagd en je antwoordt met meer dan één zin, wordt alleen de eerste zin in de beoordeling meegeteld.”[10]

Over correctheid van taalgebruik bij beantwoording staat tevens onder andere: “De correctheid van het taalgebruik wordt getoetst bij alle antwoorden van de kandidaat op open vragen. Onder incorrect taalgebruik moet worden verstaan: spelfouten (inclusief fouten in de werkwoordspelling) en fouten in de zinsbouw (inclusief verkeerde woordvolgordes en congruentiefouten). Bij de beoordeling van de spelling dient uitgegaan te worden van de schrijfwijze volgens de Leidraad bij de Woordenlijst Nederlandse Taal (het Groene Boekje). Voor incorrect taalgebruik worden maximaal vier scorepunten in mindering gebracht.

[10] https://www.examenblad.nl/system/files/exam-document/2024-07/vw-1001-a-24-1-o-spr.pdf

Hierbij dient de volgende aftrekregeling te worden gehanteerd: 0 fouten 0, 1 fout of 2 fouten 1, 3 of 4 fouten 2, 5 of 6 fouten 3, 7 of meer fouten 4."[11]

Het centraal examen Nederlands heeft betrekking op domein A (leesvaardigheid) en domein D (argumentatieve vaardigheden). Er is voor de volgende subdomeinen aandacht: A1 (Analyseren en interpreteren), A2 (beoordelen), A3 (samenvatten) en domein D voor zover het analyseren en beoordelen betreft.[12]

Het examen van 2024 bestond uit 38 vragen waarvoor maximaal 65 punten gehaald konden worden. De gewenste responses zijn ruwweg te verdelen over multiple choice (9 vragen, 11 punten), sleepvragen (3 vragen, 7 punten), korte open antwoorden van maximaal 30/40 woorden (16 vragen, 27 punten) en korte open antwoorden van een aantal zinnen (10 vragen, 20 punten). Vragen waarbij lange open antwoorden van meerdere alinea's nodig zijn staan niet in het Nederlands examen van 2024.

Correctielast

Met een gemiddelde correctielast per docent van 27 uur is Nederlands het meest correctie-intensieve vak van het vwo (Essen et al., 2023). Bij vmbo gl/tl is Nederlands (19 uur) tevens het meest correctie-intensieve vak, en bij havo komt Nederlands (28 uur) na Geschiedenis (31 uur) op plek twee. De correctie-intensiteit komt voornamelijk door het hoge aantal examenkandidaten per docent van 36,75, het hoogste aantal van alle vakken. Uit Tabel BIII.5 van het onderzoek van Essen et al. (2023) blijkt dat naast het hoge aantal examenkandidaten per docent en de hoge gemiddelde correctielast, bij het vak Nederlands op het vwo ook meespeelt dat het na Kunst (11,06 uur) de hoogste gemiddelde voorbereidingstijd vergt (8,54 uur). De reden hiervoor ligt erin dat veel docenten aangeven het examen van tevoren zelf te maken (82,8%) en dat ze antwoorden van tevoren uitwisselen met vakgenoten (78,7%) om voor de daadwerkelijke beoordeling al goed voorbereid te zijn.

Bij het nakijken van het examen Nederlands zijn multiple-choice en sleepvragen relatief eenvoudig na te kijken: de beoordeling is eenduidig en kost niet al te veel tijd. Deze vragen zouden ook gemakkelijk geautomatiseerd nagekeken kunnen worden, maar dit levert naar verwachting weinig tijdswinst op. Open vragen kosten meer tijd om na te kijken. Een andere uitdaging in het nakijken van het examen Nederlands is de grammatica- en spellingscontrole. Voor incorrect taalgebruik worden over het gehele examen maximaal 4 scorepunten in mindering gebracht. Je kunt dus niet alleen per vraag bekijken hoeveel grammatica- of spellingsfouten de leerling gemaakt heeft. Dit kost nog extra tijd voor docenten.

Beschrijving scenario

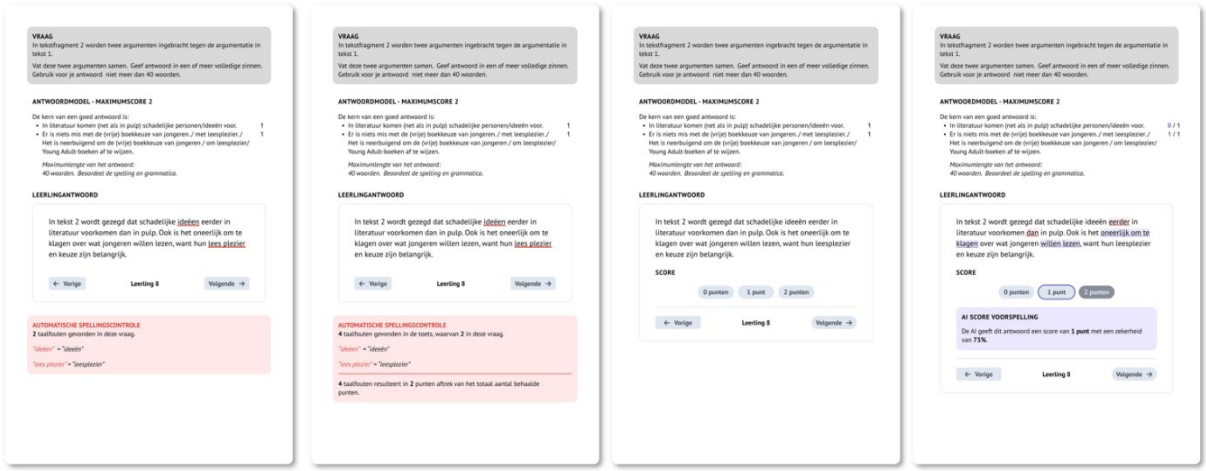
In het huidige scenario zijn vier opties beschreven hoe AI ingezet zou kunnen worden bij het nakijken (zie Figuur 3). De schetsen van deze opties, inclusief animaties, zijn ook te bekijken via https://BOO-Nakijken-AI-figma.web.app, waarbij animatie 2.0 de kale examenopdracht weergeeft en de nummering 2.1, 2.2, 2.3 en 2.4 correspondeert met de bijbehorende optie.

Bij de eerste optie worden grammatica- en spellingsfouten automatisch gedetecteerd per vraag (Kukich, 1992). Bij de tweede optie is dit ook zo, maar geeft AI aanvullend ook informatie over de grammatica- en spellingsfouten in de rest van het examen, en het totaalaantal puntenaftrek waarin dit resulteert. Als derde optie geeft AI een score suggestie voor een volledig open antwoord. De docent moet bij deze optie eerst zelf een score aanklikken en ziet hierna de scoresuggestie van AI. Hier is voor gekozen zodat de docent wel in de lead blijft, en AI als hulpmiddel en extra controle functioneert. Als laatste optie doet AI tevens deze scoresuggestie, maar staat hierbij ook een inschatting van de zekerheid die AI heeft over de score (Nye et al., 2014).

[11] https://www.examenblad.nl/system/files/exam-document/2024-05/vw-1001-a-24-1-c.pdf
[12] https://www.examenblad.nl/system/files/2022/syllabi/syllabus_nederlands_4f_vwo_2024_8_juli_2022_versie_2.pdf

Figuur 3

Visuele weergave opties scenario 2 – Nederlands vwo



Optie 2.1 – automatische spellingscontrole (opdracht) | **Optie 2.2 – automatische spellingscontrole (toets)** | **Optie 2.3 – AI voorspelling score** | **Optie 2.4 – AI voorspelling score + uitleg scoring**

De eerste twee opties voor nakijken met AI zoals hierboven beschreven vallen onder de noemer van gedeeltelijke automatisering in het model van Molenaar (2022), omdat de grammatica- en spellingscontrole geautomatiseerd kan worden uitgevoerd, met enkel een controle van de docent. De laatste twee mogelijkheden voor nakijken met AI zoals hierboven beschreven vallen onder de noemer van docent assistentie. Bij de docent assistentie heeft de docent volledige controle en geeft AI alleen ondersteunende informatie om de docent bij keuzes te ondersteunen.

4.2.3 Biologie vmbo bb/kb

Beschrijving examen

Het examen biologie voor vmbo bb/kb is een digitaal examen, afgenomen in Facet, waar leerlingen anderhalf uur de tijd voor hebben. De kandidaten krijgen geen aanvullende algemene instructies over hoe de vragen beantwoord moeten worden.

Het examen heeft betrekking op de volgende domeinen en subdomeinen: Leervaardigheden in het vak biologie, Cellen staan aan de basis, Planten en dieren en hun samenhang: de eigen omgeving verkend, Het lichaam in stand houden: voeding en genotmiddelen, energie, transport en uitscheiding, Reageren op prikkels, Van generatie op generatie.

Het examen biologie vmbo bb van 2024 bestond uit 38 vragen waarvoor maximaal 44 punten gehaald konden worden. 1 vervallen vraag kreeg altijd de score 1. De vraagsoorten voor dit examen bestonden uit 24 multiple-choice (24 punten), 12 korte open antwoord vragen van verschillende vormen (17 punten) en 2 overige vragen (3 punten).

Correctielast

De gemiddelde correctielast van het biologie-examen voor vmbo gl/tl is 11 uur. Er is geen gemiddelde correctielast berekend voor het vmbo bb examen. Met de genoemde nakijktijd is biologie één van de minst nakijk intensieve vakken: alleen de moderne vreemde talen en nask kosten de docent minder nakijktijd. Voor havo en vwo heeft biologie een iets hogere correctielast: gemiddeld 15 uur. Ook voor havo en vwo is biologie één van de minder intensieve vakken qua correctielast (Essen et al., 2023).

Bij de digitale afname van het examen worden de multiple-choice vragen al automatisch nagekeken. Hierin valt dus geen tijds- of consistentiewinst meer te behalen. De korte open antwoorden worden niet

automatisch nagekeken, en daarmee ook niet altijd consistent (zie onderdeel ‘consistentie’ in aanleiding). Er bestaat een discussieforum voor docenten om met elkaar in gesprek te gaan over leerlingantwoorden of opgaven. Dit forum wordt nog niet veel gebruikt.

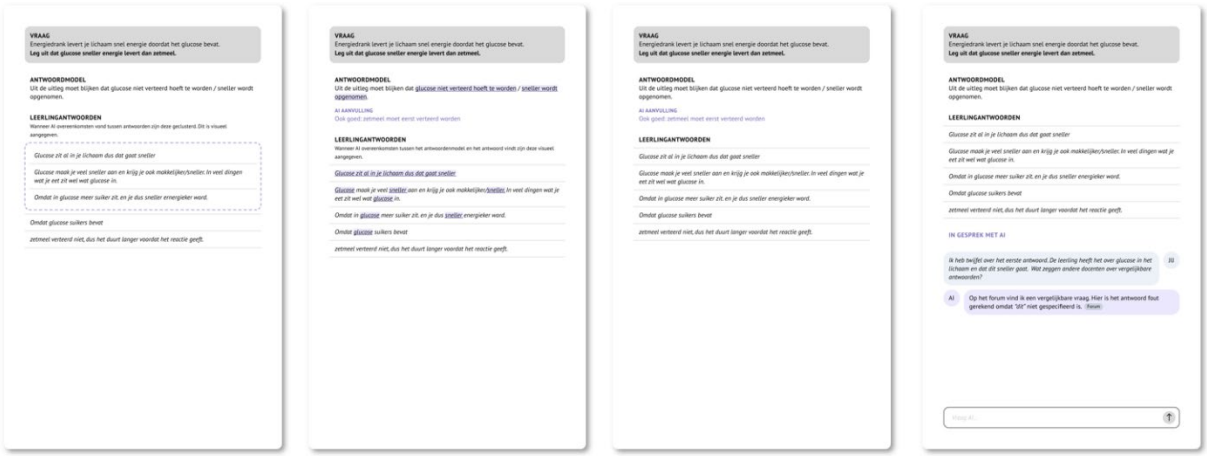
Beschrijving scenario

In dit scenario zijn wederom vier opties beschreven hoe AI ingezet zou kunnen worden bij het nakijken (zie Figuur 4). De schetsen van deze, inclusief animaties, zijn ook te bekijken via https://BOO-Nakijken-AI-figma.web.app, waarbij animatie 3.0 de kale examenopdracht weergeeft en de nummering 3.1, 3.2, 3.3 en 3.4 correspondeert met de bijbehorende optie.

Bij de eerste optie kan AI vergelijkbare antwoorden groeperen. Hierdoor kan een docent makkelijker vergelijken welke antwoorden wel en niet goed zijn (Basu et al., 2013; Horbach et al., 2014; Mieskes & Padó, 2018). Omdat de antwoorden in deze optie per vraag nagekeken worden, verkleint dit de kans dat een antwoord van een leerling aan het einde van het nakijkproces anders nagekeken wordt dan een antwoord aan het begin van het nakijkproces. Bij de tweede optie markeert AI belangrijke woorden uit het correctievoorschrift in het antwoord van de leerlingen, wat mogelijk kan dienen als ondersteuning voor docenten bij het bepalen van de correctheid van een antwoord (Del Gobbo, 2023). Bij de derde optie wordt er een uitgebreider correctievoorschrift door AI gegenereerd. Dit kan bijvoorbeeld op basis van analyses bij pre-tests, waarna de extra AI-opties worden vastgesteld door de CvTE vaststellingscommissie. Als vierde optie kan AI hulp bieden bij het nakijken via het discussieforum. Hierbij is het discussieforum geïntegreerd in de nakijkapplicatie. De docent kan, bij twijfel over een bepaald antwoord, aan AI vragen om op het forum op zoek te gaan naar relevante vragen/antwoorden van andere docenten en deze laten samenvatten. Deze laatste twee suggesties komen voort uit gesprekken met toetsdeskundigen.

Figuur 4

Visuele weergave opties scenario 3 – biologie vmbo bb/kb



Optie 3.1 – clustering antwoorden | **Optie 3.2 – markeringen van woorden uit antwoordmodel** | **Optie 3.3 – aanvulling AI op correctievoorschrift** | **Optie 3.4 – forum**

De mogelijkheden voor nakijken met AI zoals hierboven beschreven vallen grotendeels onder de noemer van docent assistentie in het model van Molenaar (2022). Hierbij heeft de docent volledige controle en geeft AI alleen ondersteunende informatie om de docent bij keuzes te ondersteunen.

5. Focusgroepen

5.1 Methode

Voor dit onderzoek zijn in totaal vijf focusgroepen georganiseerd met deelnemers afkomstig uit verschillende stakeholdergroepen. De werving van deelnemers vond plaats via meerdere kanalen. Enerzijds werden warme contacten benaderd via interne netwerken, toetsdeskundigen en andere relevante connecties. Anderzijds zijn er uitnodigingen verstuurd naar algemene e-mailadressen van organisaties en instellingen om een bredere groep geïnteresseerden te bereiken.

In totaal namen 26 personen deel aan de focusgroepen, verdeeld over 5 sessies met ieder 4 tot 7 deelnemers. De deelnemers zijn werkzaam binnen verschillende typen organisaties, waaronder middelbare scholen, de VO-raad, Inspectie van het Onderwijs, het College voor Toetsen en Examens (CvTE), Stichting Cito en de Federatie van Onderwijsvakorganisaties (FvOv). Deze samenstelling zorgde voor een divers perspectief op de besproken thema's. In Bijlage A is de exacte samenstelling te vinden per focusgroep, met daarin de verdeling over typen organisaties en functies.

De deelnemers konden kiezen tussen een online of een fysieke focusgroep. De opbouw van de sessies was in beide vormen gelijk. Elke focusgroep begon met een introductie, gevolgd door het bespreken van de drie scenario's en een afsluiting. De totale duur van elke sessie bedroeg circa twee uur. De sessies werden begeleid met behulp van een PowerPointpresentatie en stonden onder leiding van een moderator. Daarnaast waren er één tot twee notulisten aanwezig om de gesprekken vast te leggen en belangrijke inzichten te documenteren.

Bij de bespreking van de scenario's werd eerst individueel input verzameld via Padlet[13], zodat deelnemers hun eerste meningen en reacties konden delen zonder beïnvloeding door anderen. In de Padlet werd specifiek gevraagd: *Wat vind je van deze opties, gelet op (1) de kwaliteit van de beoordeling, (2) de autonomie van de beoordelaar en (3) de efficiëntie van het correctieproces?* Deze vragen hielpen om gerichte feedback te verzamelen en om inzicht te krijgen in de verschillende afwegingen die deelnemers maakten. Daarnaast was er ruimte om algemene opmerkingen te noteren, bijvoorbeeld over de haalbaarheid en wenselijkheid van het scenario. Bovendien hielp de Padlet de discussie te structureren en leverde het aanvullende input op die achteraf gebruikt kon worden voor analyse. Daarna werd in een groepsdiscussie dieper ingegaan op de verschillende scenario-opties, waarbij deelnemers hun perspectieven toelichtten en gezamenlijk reflecteerden op mogelijke implicaties en voorkeuren.

Alle gemaakte notities en antwoorden uit de Padlet zijn geordend per scenario, per toepassing en per focusgroep in een Excel-bestand. Om de data beter te kunnen analyseren, zijn alle opmerkingen voorzien van thematische labels. De thema's zijn als volgt:

1. **Waardeoordelen of persoonlijke voorkeuren**. Dit thema omvat uitspraken waarin deelnemers hun houding, gevoel of persoonlijke voorkeuren uitdrukten.
2. **Kwaliteit van de beoordeling**. Dit thema omvat opmerkingen over de betrouwbaarheid, nauwkeurigheid en rechtvaardigheid van de beoordeling met AI.
3. **Efficiëntie van het correctieproces**. Dit thema omvat alle opmerkingen over tijdsbesparing en werkdruk als gevolg van de AI-ondersteuning.
4. **Autonomie van de beoordelaar**. Dit thema omvat uitspraken over de mate van vrijheid en controle die de beoordelaars behouden bij het gebruik van AI tijdens het nakijken, en in hoeverre AI hun professionele beslissingen beïnvloedt.

[13] https://padlet.com/

5. **Haalbaarheid**. Dit thema omvat opmerkingen over de haalbaarheid van de AI-optie, waarin deelnemers hun enthousiasme of juist vraagtekens delen.

Omdat er twee notulisten aanwezig waren, zijn dubbele opmerkingen uit de Excel verwijderd om herhaling te voorkomen. Daarnaast is elke opmerking voorzien van een label over de toon: positief, negatief of neutraal, afhankelijk van de strekking van de reactie. Deze ordening maakt het mogelijk om patronen in de houding ten opzichte van AI-toepassingen te herkennen en analyseren.

Om de data te analyseren, is een score toegekend aan elke opmerking, namelijk positief = 1, neutraal = 0 en negatief = -1. Per label (waarde, kwaliteit, efficiëntie, autonomie en haalbaarheid) is voor elke AI-optie het aantal opmerkingen opgeteld en een gemiddelde score berekend. Per optie is eveneens een gewogen gemiddelde berekend, wat diende als algehele indicator voor het sentiment rond deze optie. Naast deze kwantitatieve gegevens, zijn de opmerkingen van de deelnemers per optie bekeken en zijn deze gebruikt om de kwantitatieve gegevens te duiden. Naast de opmerkingen over de scenario's en AI-opties zelf, kwamen er ook verschillende keren algemene opmerkingen naar boven die relateren aan het nakijken van eindexamens met AI. Een samenvatting van de belangrijkste strekking van deze opmerkingen is per scenario toegevoegd.

5.2 Resultaten

5.2.1 Beschrijving deelnemers

Zoals in de methode beschreven, bestonden de focusgroepen uit een mix van stakeholders uit de examenketen, waaronder docenten van diverse vakken, sectiehoofden, clustermanagers, leden van vaststellingscommissies en beleidsmedewerkers. Om een indruk te krijgen van het sentiment bij de deelnemers rondom AI, is in de voorstelronde gevraagd om kort te omschrijven in hoeverre deelnemers AI gebruiken in hun dagelijks leven en werk en om hun gevoel bij AI in één woord te omschrijven.

Van de 26 deelnemers was er geen één deelnemer die aangaf nooit gebruik te maken van AI. Wel varieerde de mate van gebruik. Sommige deelnemers waren erg enthousiast en gaven over hun gebruik aan: "Overal (zowel privé als werk)", "Veel privégebruik" en "Heel veel, adem er nog net niet mee". Bij anderen was het nog minder geïntegreerd in hun dagelijkse praktijk en gaf iemand bijvoorbeeld aan dat het binnen de organisatie waar de persoon werkzaam is, verboden is om AI te gebruiken. In privé wordt AI door de deelnemers bijvoorbeeld ingezet voor het genereren van recepten, persoonlijke ontwikkeling en het opdoen van nieuwe ideeën. Op werkgebied gaven deelnemers aan het onder andere te gebruiken voor het herschrijven of samenvatten van teksten, het genereren van opdrachten en lesideeën en als programmeerhulp.

De omschrijvingen in één woord waren uiteenlopend: termen als "enthousiast", "nieuwsgierig", "vernieuwend" en "nuttige tool" stonden naast woorden die onvermijdelijkheid of toekomstgerichtheid uitdrukten, zoals "onontkoombaar", "onvermijdelijk" en "toekomst". Ook kwamen ambivalente gevoelens naar boven, zoals "dubbel", "gemengde gevoelens", "spannend" en "complex", en meer kritische of bezorgde termen zoals "negatief", "problematisch", "creativiteit verminderen" en "duurzaamheidsissue". Uit andere reacties bleek vooral de fascinatie voor AI: "intrigerend", "fascinerend" en "razend interessant". Al met al laten de reacties zien dat er binnen de verschillende groepen een divers palet aan ervaringen en sentimenten aanwezig was, variërend van enthousiast en toekomstgericht tot kritisch en terughoudend, wat duidt op een breed uitgangspunt voor de gesprekken over de inzet van AI bij het nakijken van eindexamens.

5.2.2 Scenario 1. Wiskunde A havo

Tabel 1

Overzicht per AI-optie en per label van de gemiddelde score en het aantal opmerkingen waarop deze is gebaseerd voor Scenario 1

	waarde		kwaliteit		efficiëntie		autonomie		haalbaarheid			
scenario	gem. score	aantal	gem. score	aantal	gem. score	aantal	gem. score	aantal	gem. score	aantal	gewogen totaal	aantal totaal
1.1	0,67	6	0,22	18	0,05	20	0,28	21	0,00	0	0,23	65
1.2	0,69	13	0,07	42	0,78	28	0,00	27	0,00	1	0,30	111

In totaal werden 176 opmerkingen verzameld over de verschillende opties binnen *Scenario 1. Wiskunde A havo*. In Tabel 1 staat een overzicht per optie en per label van de gemiddelde score en het aantal opmerkingen waarop deze gebaseerd is. Zowel bij scenario 1.1 als bij 1.2 kwamen vooral de thema's kwaliteit, efficiëntie en autonomie veelvuldig aan bod. Opvallend is echter dat er bij scenario 1.2 aanzienlijk meer opmerkingen zijn verzameld dan bij scenario 1.1. De thema's waarde en haalbaarheid kwamen in beide scenario's weinig ter sprake.

Over het algemeen waren de reacties positief over de waarde voor de beoordelaar van zowel scenario 1.1 als 1.2. Deelnemers zagen daarnaast voordelen op het gebied van efficiëntie binnen scenario 1.2, maar dit scenario riep tegelijkertijd veel vragen op over de kwaliteit, autonomie en haalbaarheid. In de volgende secties worden de belangrijke bevindingen uitgewerkt.

Optie 1.1 Denkstap markeringen

Van de 176 opmerkingen waren er 65 gerelateerd aan scenario 1.1. De gemiddelde score van 0,23 wijst op een overwegend neutrale beoordeling. Deelnemers hadden geen écht sterke bezwaren, maar waren ook niet uitgesproken enthousiast. Zij zien het scenario vooral als een mogelijke optie, maar vinden het nog onvoldoende overtuigend om er volledig achter te gaan staan.

Opvallend is dat deelnemers vooral positief waren over de waarde van deze toepassing (M=0,67; n=13). Zo werd het scenario "handig" genoemd en sprak iemand van "een mooi voorbeeld, het is echt een hulpmiddel." Ook vonden deelnemers het waardevol dat zichtbaar wordt "welke tussenstap bij het leerlingantwoord hoort." Tegelijkertijd klonk er voorzichtigheid door in opmerkingen als: "interessant hulpmiddel, maar roept ook op in hoeverre je het hulpmiddel kunt vertrouwen."

De kwaliteit van de toepassing werd breder en gevarieerder besproken (M=0,22; n=18). Met name de manier waarop de AI omgaat met verschillende soorten leerlingantwoorden riep vragen op. Zo vroeg een deelnemer zich af: "Ik ben wel benieuwd wat deze tool toe kan voegen als een leerling een minder helder antwoord formuleert, bijvoorbeeld als de leerling onnodige tussenstappen gebruikt maar wel uiteindelijk het goede antwoord vindt." Een ander benadrukte dat leerlingen vaak creatief zijn: "Het cv geeft vaak 1 van de mogelijke oplossingen, leerlingen zijn in staat om nog 10 andere oplosstrategieën te bedenken, zeker bij complexere wiskundevragen. Kan AI daar mee omgaan?"

Daarnaast wezen verschillende deelnemers op voorwaarden die cruciaal zijn voor de ervaren kwaliteit. Zo benadrukten sommigen dat "je er een soort vertrouwen in moet hebben dat AI het goed doet, wanneer durf je dat?" en dat ze "wel 1000 voorbeelden nodig hebben om te zien dat het kan." Deze opmerkingen laten zien dat de beoordeling van kwaliteit sterk samenhangt met vertrouwen in het systeem en voldoende bewijs van consistent functioneren. Tegelijkertijd zagen anderen juist voordelen: "Door het oplichten van de passage in de uitwerking, wordt er minder over het hoofd gezien door een docent. Dat is goed voor de kwaliteit," en ook: "Beide opties verhogen kwaliteit van de correctie."

Wanneer het ging over efficiëntie waren deelnemers over het algemeen terughoudender (M=0,05; n=20). De meesten betwijfelden of het markeren van denkstappen echt tijdwinst oplevert. Zo merkte iemand op: “Optie 1 is helemaal niet efficiënt, markeringen maar je moet het zelf nog nakijken, als je ook nog een beeldscherm met AI moet nakijken duurt het alleen maar langer.” Anderen vroegen zich af: “Is dit wel tijdbesparend?” of gaven aan te twijfelen “of dit een grote efficiëntieslag gaat teweegbrengen.” Toch zagen enkele deelnemers juist voordelen, vooral bij complexere opdrachten. Het markeren van denkstappen kan volgens hen “bij complexe opdrachten wel helpen bij het corrigeren, met name met sneller nakijken,” en iemand stelde: “Ik denk dat het een voordeel is dat AI al snel op zoek gaat naar de denkstappen. Dit versnelt volgens mij het nakijkproces.”

Tot slot heerste er relatief veel overeenstemming over de mate van autonomie van de docent (M=0,28; n=21). Veel deelnemers vonden dat de docent duidelijk de regie houdt: “Het is sowieso ondersteuning van de docent, kan helpen, is optioneel dus autonomie blijft gewaarborgd,” en ook: “Je bent zelf nog steeds de corrector, AI is je hulpmiddel en niet je correctiemodel.” Een ander benadrukte: “De verantwoordelijkheid blijft dan echt bij de docent.” Een kleinere groep plaatste hier echter kanttekeningen bij en wees erop dat autonomie in de praktijk wel afhankelijk blijft van docentgedrag. Zo vond iemand dat het de vraag is “in hoeverre de corrector zelf de autonomie invult” en een ander gaf aan: “Docent heeft altijd laatste woord, zolang docent niet te automatisch op AI afgaat.”

Optie 1.2 Denkstap markeringen en scoresuggesties

Van de 176 opmerkingen waren er 111 gerelateerd aan scenario 1.2, waarin de denkstapmarkeringen en scoresuggesties verder waren uitgewerkt. De gemiddelde score van 0,3 duidt op een voorzichtig positieve beoordeling. Deelnemers waren met name positief over de waarde en de efficiëntie van dit scenario. Hoewel het scenario dus kansen biedt, is het enthousiasme nog niet uitgesproken. Men ziet duidelijke voordelen, maar deze zijn nog niet sterk genoeg om tot een overtuigend positief oordeel te komen.

De deelnemers waren overwegend positief over de waarde van de toepassing (M=0,69; n=13). Hun opmerkingen richtten zich vooral op het automatisch toekennen van scores in combinatie met de gemarkeerde denkstappen. Zo werd het scenario omschreven als “een mooie en heldere manier om de denkstappen te visualiseren, met als bijkomend voordeel dat er meteen een scorevoorstel wordt gedaan.” Anderen benadrukten dat het markeren van denkstappen overzicht geeft en dat het automatisch berekenen van de score “erg handig” is. Daarnaast gaf een deelnemer aan dat de toepassing helpt om met een frisse blik naar leerlingantwoorden te kijken: “Ik vind het heel helpend. Het biedt soms echt een andere manier om ernaar te kijken. Ik ben zelf vrij streng.”

Over de kwaliteit van de toepassing waren deelnemers een stuk terughoudender (M=0,07; n=42). Een belangrijke zorg betrof de vraag in hoeverre de AI in staat is om uiteenlopende leerlingantwoorden correct te herkennen, vooral wanneer leerlingen afwijkende of minder heldere formuleringen gebruiken. Zoals een deelnemer zei: “Ik vraag me wel af in hoeverre dit tot problemen kan leiden als leerlingen een andere manier van noteren gebruiken, maar ik ben geen vakexpert, dus kan hier vrij weinig over zeggen.” Een ander vroeg zich af hoe de tool omgaat met antwoorden waarin onnodige tussenstappen worden gebruikt, maar het eindantwoord wel klopt: “Ik ben wel benieuwd wat deze tool toe kan voegen als een leerling een minder helder antwoord formuleert.”

Daarnaast riep vooral de scoresuggestie vragen op. In het scenario werd bij een deelantwoord 0 punten toegekend, terwijl deelnemers zelf hiervoor wél een punt zouden geven. Hoewel dit een fout in het scenario bleek en niet in de AI, beïnvloedde het wel de beoordeling. Een deelnemer merkte hierbij op: “De scoresuggestie lijkt nog niet correct te werken, en zou alleen nuttig zijn indien het goed werkt.” Ook werden vakspecifieke vragen gesteld, bijvoorbeeld: “Bij wiskunde hoeft een antwoord wat tussen haakjes staat niet door de leerling opgeschreven te worden. Het totaal aantal punten moet dus 3 zijn en niet 2. Zijn de vakspecifieke regels bij AI bekend?” De bredere terugkerende zorg betrof de betrouwbaarheid: “En wat als

leerlingen het op een andere manier noteren, kan AI het er dan ook nog uithalen? En hoe detecteert AI een rekenfout, maar als de stappen wel juist zijn?"

Over de efficiëntie waren deelnemers juist opvallend positief (M=0,78; n=28). Veel deelnemers zagen een duidelijke tijdsbesparing in het automatisch markeren en scoren van denkstappen. Reacties als "zorgt voor meer efficiëntie, minder tijd kwijt", "veel kan tijd besparen", "het werk gaat sneller, er is minder zoekwerk en er ligt al een voorstel" en "alles wat alle punten heeft, hoef je niet meer na te kijken" illustreren die overtuiging. Enkele deelnemers waren minder overtuigd, vooral wanneer het ging om eenvoudige antwoorden of situaties met meerdere correctoren. Zoals iemand verwoordde: "Ik vraag me af of dit helpt, omdat de makkelijk vindbare antwoorden toch ook snel na te kijken zijn. Ik ben er niet van overtuigd dat het tijdswinst gaat opleveren."

Wanneer het gesprek op autonomie kwam, liepen de meningen sterk uiteen (M=0,00; n=27). Een deel van de deelnemers maakte zich zorgen over het risico dat docenten te afhankelijk zouden worden van de AI. Zij wezen erop dat het "lastig [is] om te zorgen dat je heel scherp blijft" en dat sommige docenten mogelijk "klakkeloos overnemen wat AI zegt", waardoor fouten van de AI extra problematisch kunnen worden. Andere deelnemers ervaarden juist dat de autonomie van de docent behouden blijft. Voor hen was duidelijk dat de AI slechts een hulpmiddel is en geen vervanger van het oordeel van de docent. Zij benadrukten dat "de docent de baas blijft" en dat de docent in alle gevallen "het recht houdt punten toe te kennen of niet."

Over de haalbaarheid van het scenario werd weinig gesproken (M=0,00; n=1). Slechts één deelnemer gaf een reactie. Deze persoon benadrukte dat het scenario in de praktijk goed getest moet worden voordat het kan worden ingezet: "Ik zou het wel gewoon echt uitgewerkt moeten zien met een echt examen. Het moet echt goed zijn, als je wil dat de kwaliteit omhooggaat en het tijd gaat schelen." Hiermee werd duidelijk dat voor een realistische beoordeling van de haalbaarheid een concrete toepassing en bewijs van de effectiviteit belangrijke voorwaarden zijn.

Algemene opmerkingen naar aanleiding van het scenario

Naast de inhoudelijke opmerkingen over de verschillende AI-toepassingen kwamen tijdens de gesprekken ook bredere zorgen naar voren over het digitaliseren van het wiskunde-examen en de inzet van AI bij het nakijken. Omdat het vak op dit moment nog volledig op papier wordt afgenomen, riep het idee van digitale examinering bij veel deelnemers praktische vragen op. Wat betekent dit in de praktijk? Wordt het hele examen digitaal afgenomen of worden papieren antwoorden simpelweg ingescand? En hoe zou zo'n overgang er concreet uitzien voor scholen, leerlingen en docenten?

Daarbij ging het gesprek regelmatig over de rol die het nakijken speelt in de professionele ontwikkeling van docenten. Een van de deelnemers verwoordde die zorg scherp: "Volgens mij leert een docent heel veel over zijn lesgeven bij het nakijken van examens. Door AI in te zetten mis je deze feedback voor de docenten." Anderen sloten zich daarbij aan en wezen op het risico dat docenten minder zicht krijgen op individuele leerlingen. Zoals iemand zei: "Met horizontaal nakijken kan je soms kwijt zijn hoe een leerling het over het hele examen heeft gedaan, hoe zou dat met AI gaan?" Een ander benadrukte het belang van het fysieke werk van de leerling: "Voor de correctie online heb je er wel minder gevoel bij, dat vind ik wel jammer. Het nakijken van een papieren examen kost meer tijd, maar dan heb je wel meer 'samen' het examen doorlopen, zie je het handschrift, herken je de leerling."

Een terugkerend thema was ook de vraag hoe beoordelingen in een AI-ondersteund systeem uitlegbaar en transparant blijven voor leerlingen. Verschillende deelnemers benoemden dat dit essentieel is voor de acceptatie en betrouwbaarheid van AI. Zoals een deelnemer formuleerde: "Vanuit de positie van de leerling, wat voor status heeft die AI-beoordeling precies? Is dat transparant voor de leerling?" Ook klonk de zorg: "Ook de uitleg richting de leerling. Snapt de leerling hoe je tot een beoordeling komt?"

Tot slot ontstond er een levendige discussie over de mogelijke zelflerende eigenschappen van het systeem. Enkele deelnemers vroegen zich af: "Leert AI van de nakijker? Kun je genegeerd worden door AI

als je te vaak een 1 geeft [in plaats van een 0]?" Een ander voegde toe: "Niet zo'n fan van gaandeweg leren op de docent zelf, dan krijg je een heel gekleurd systeem." Ook praktische vragen kwamen voorbij, zoals: "Hoe zit het dan met flexibele afname als een school in april iets doet en een andere school pas in mei?" Opvallend hierbij is dat deze zorgen niet direct voortkwamen uit de scenario's zelf: de technische werking van het systeem werd daarin niet benoemd. Toch geven de opmerkingen een helder beeld van de aannames en de onzekerheden die bij sommige deelnemers leven over de adaptiviteit van AI.

5.2.3 Scenario 2. Nederlands vwo

In totaal werden 200 opmerkingen verzameld over de verschillende opties binnen *Scenario 2. Nederlands vwo*. In Tabel 2 staat een overzicht per optie en per label van de gemiddelde score en het aantal opmerkingen waar deze op gebaseerd is. Er zijn veel opmerkingen gemaakt over de waarde van de opties, en kwaliteit en efficiëntie kregen ook veel aandacht. Het is opmerkelijk dat autonomie veel opriep bij optie 2.3 en 2.4, maar dat dit minder onderwerp van commentaar was bij optie 2.1 en 2.2. Over de haalbaarheid van de opties zijn nauwelijks opmerkingen gemaakt. Optie 2.1 en 2.2 werden overwegend positief beoordeeld en optie 2.3 en 2.4 ontvingen een gematigd positieve beoordeling. In de volgende secties worden de belangrijkste bevindingen uitgewerkt.

Tabel 2
Overzicht per AI-optie en per label van de gemiddelde score en het aantal opmerkingen waarop deze is gebaseerd voor scenario 2

	waarde		**kwaliteit**		**efficiëntie**		**autonomie**		**haalbaarheid**			
scenario	gem. score	aantal	gem. score	aantal	gem. score	aantal	gem. score	aantal	gem. score	aantal	gewogen totaal	aantal totaal
2.1	0,80	15	0,43	7	1,00	10	0,57	7	1,00	1	0,75	40
2.2	0,70	20	0,14	7	0,71	7	0,40	5	0,50	2	0,56	41
2.3	0,00	13	0,50	10	-0,22	9	0,58	19	0,00	1	0,27	52
2.4	0,65	26	0,60	10	0,08	12	0,47	19	0,00	0	0,49	67

Optie 2.1 Automatische spellingscontrole (opdracht) en optie 2.2 Automatische spellingscontrole (toets)

Van de 200 opmerkingen in totaal waren er 81 van toepassing op optie 2.1 (automatische spellings- en grammaticacontrole per opdracht) en optie 2.2 (dezelfde controle met aanvullende automatische berekening over de gehele toets). Beide AI-opties scoren hoog, waarbij optie 2.1 met een gewogen totaal van 0.75 (n=40) iets hoger uitvalt dan optie 2.2 met een gewogen totaal van 0.56 (n=41).

Deelnemers gaven een overwegend positief waardeoordeel voor beide opties (M=0.80; n=40 en M=0.70; n=41). Veel reacties benadrukten het praktisch nut: "Erg handig die spellingscontrolerol", "Buitengewoon handig" en "Ik vind de spellingscontrole mooi, dat kan AI vast ook heel goed". Deelnemers zagen de toepassing voor zich, maar benadrukten dat de docent de spellingscontrole goed moet nalopen en scherp moet blijven of AI het goed doet. Er werd een kanttekening geplaatst over de presentatie van de tussentotalen in optie 2.2: enkele deelnemers ervaarden dit als een informatieoverload ("Veel informatieoverload om de telling de gehele tijd onderin te zien"), wat bijdraagt aan de iets lagere score van optie 2.2.

De perceptie van kwaliteit verschilt tussen de opties. Optie 2.1 kreeg overwegend positieve beoordelingen (M=0.43; n=7), terwijl optie 2.2 richting neutraal scoorde (M=0.14; n=7). Voorstanders stelden dat "spelling en grammatica hoge kwaliteit kan hebben" en dat onnodige fouten hierdoor systematisch worden tegengegaan, wat "leidt tot consequent nakijkgedrag". Tegelijkertijd werd een belangrijk bezwaar

genoemd, omdat AI mogelijk ongelijk oordeelt over uiteenlopende taalregisters[14], met risico's voor kansengelijkheid: "(..) Leerlingen uit taalzwakkere gebieden hebben wellicht een ander taalregister dan leerlingen elders. Hier zou een bias in kunnen sluipen waar dit strenger wordt beoordeeld." Hoe verder het taalgebruik in een bepaald register afwijkt van de norm, hoe strenger leerlingen mogelijk worden beoordeeld.

Deelnemers verwachten een duidelijke winst in efficiëntie door automatisering in deze opties: "Efficiëntie bij de grammatica en spellingscontrole is héél hoog" en "Het automatisch aangeven van spellingsfouten lijkt mij erg efficiënt". Deelnemers gaven aan dat het routinewerk door AI kan worden overgenomen, waardoor er tijd vrijkomt voor de docent om zich te richten op de inhoudelijke beoordeling. Deelnemers waren het eens dat beide opties administratie- en correctietijd kunnen terugdringen, mits de gebruikersinterface en informatiepresentatie zorgvuldig wordt vormgegeven (bijvoorbeeld de totaalstand bij optie 2.2).

De minste opmerkingen werden gemaakt over autonomie (totaal n=12) en haalbaarheid (totaal n=3). Verschillende deelnemers benadrukten dat "de autonomie blijft", aangezien je als docent de doorslag geeft in de eindbeoordeling. Dit werd als fijn ervaren, zeker omdat er nog wat terughoudendheid bestond over hoe goed AI de controle kan uitvoeren en daarmee de haalbaarheid van de opties in het algemeen: "Ik twijfel of AI het er goed uithaalt. Wil wel controleren" en "Het is zeker haalbaar, wel moet er gekeken worden naar mogelijke opties. Soms zijn er meerdere woorden mogelijk voor dezelfde boodschap." Sommige deelnemers ervaarden deze terughoudendheid in haalbaarheid niet en vonden dat AI nog meer autonoom mag zijn: "Voor de grammatica- en spellingscontrole vind ik die autonomie niet per se nodig en mag hij wat mij betreft bij AI liggen (in eerste instantie)".

Optie 2.3 AI-voorspelling score en optie 2.4 AI-voorspelling score + uitleg

In totaal werden er 119 opmerkingen gemaakt over de opties 2.3 en 2.4. De beoordelingen van optie 2.3, waarbij AI een scorevoorspelling geeft nadat de docent een oordeel heeft gegeven, en optie 2.4, waarin deze voorspelling wordt uitgebreid met een uitleg door AI, zijn over het algemeen gematigd positief. Optie 2.4 (M=0.49; n=67) wordt door deelnemers positiever beoordeeld dan optie 2.3 (M=0.27; n=52), wat met name kan worden verklaard door verschillen in de toegekende waarde en kwaliteit van de beoordeling. Beide opties werden relatief negatief tot neutraal beoordeeld op efficiëntie.

De perceptie van waarde verschilt duidelijk tussen de twee opties, waar optie 2.3 gemiddeld een neutrale beoordeling kreeg (M=0.00, n=13) en optie 2.4 overwegend positief scoorde (M=0.65; n=26). Het toekennen van punten door AI met bijbehorend zekerheidspercentage werd door een heel aantal deelnemers niet als positief ervaren. Zij zagen het als een stap te ver, onder meer omdat de onderliggende berekening van AI onduidelijk is: "Ik zou wegblijven van puntentoekenning, ook achteraf, gewoon als hulpmiddel", "Bij optie 3 en 4 zou ik percentage niet geven, als je het niet goed kan uitleggen, dan zou ik het niet noemen dat het bijvoorbeeld 73% is. Iedereen heeft daar blijkbaar een eigen gevoel bij", en "De procentscore is misschien geen toegevoegde waarde". Optie 2.4 scoort dan ook hoger, omdat het bij deze optie duidelijker is waar de puntentoekenning vandaan komt: "Het percentage zegt mij niet zoveel, want waar is dat op gebaseerd. Dan vind ik het duidelijker waar ze uitleg bij hebben (optie 4)", en "Optie 3 is mooi dat het achteraf gebeurd, maar zonder uitleg niet duidelijk genoeg". Echter, men vond de uitleg van AI middels markering niet altijd voldoende: "Een expliciete uitleg van het totstandkomen van de score AI zou wenselijk zijn, boven op het markeren van de relevante tekst." Tegenover de kritische deelnemers, ervaarden andere deelnemers de beschreven twijfels niet: "Optie 4 is volgens mij een hele sterke optie waarbij AI ook aangeeft waar de score gehaald wordt. Je hebt dan als docent een goed beeld waar deze score vandaan komt", en "Percentage erbij is mooi, dat je meteen het verschil ziet ook met je eigen

[14] Een taalregister is de manier waarop iemand taal gebruikt in een bepaalde situatie, denk aan variatie in woordkeuze, zinsbouw, formaliteit en stijl. Leerlingen uit taalzwakkere gebieden gebruiken soms een ander register, dat afwijkt van de standaardformulering. Ook NT2-leerlingen kunnen door invloed van hun moedertaal en leerproces een ander register hanteren.

beoordeling. Ook bevestigend als AI op dezelfde score uitkomt. Als het 98% is, kijk je wel met een andere blik dan als het bijvoorbeeld 53% is."

De inschatting van de kwaliteit is voor beide opties vergelijkbaar en bevindt zich gemiddeld tussen neutraal en positief (M=0.50; n=10 en M=0.60; n=10). Een belangrijk aandachtspunt, waar de meeste negatieve opmerkingen over gingen, is het risico op inconsistenties in de beoordeling wanneer een docent en AI verschillen van mening: "Je wordt toch wel beïnvloed. Je moet als docent heel sterk in je schoenen staan als je je tweede beoordelaar wilt overtuigen dat jij 2 punten wilt toekennen aan een vraag waar AI zegt 'Nee 1 punt met …% zekerheid.' Ik ben bang voor meer inconsistenties, want waar ligt de grens? Wanneer is AI zeker genoeg?" Anderen stemden hiermee in: "Maar de ene docent vindt misschien bij 90% ga ik erin mee en de andere 95%" en "Ik zou geen percentage noemen omdat de interpretatie van dit percentage onduidelijk is." Daarom werd een landelijke richtlijn voor de interpretatie van het zekerheidspercentage gewenst om de uniformiteit te waarborgen. Tegelijkertijd werd door andere deelnemers de consistentie van AI juist als voordeel genoemd: "Ik zie ook hier AI weer als een soort VAR die de beoordelaar ondersteunt."

De verwachtingen ten aanzien van de efficiëntie zijn voor beide opties voornamelijk neutraal tot negatief (M=-0.22; n=9 en M=0.08; n=12). Hoewel sommige deelnemers de feedback van AI nuttig leek, onder meer omdat de opties 'terugbladeren' in papieren of digitale examens kan verminderen en het daarmee tijd bespaart, gaven anderen juist aan dat het meenemen van een AI-oordeel en zekerheidspercentage extra werk en overleg met zich meebrengt: "Als jij eerst nakijkt en dan AI eroverheen komt, vind ik niet dat het proces er veel efficiënter door wordt." En: "Een scorevoorspelling met een percentage over de zekerheid lijkt mij heel lastig voor de autonomie en efficiëntie, krijg je dan niet juist een discussie over een percentage?" Sommige deelnemers kwamen op het idee om AI als tweede beoordelaar te laten functioneren in plaats van een mens, wat "een boel overlegwerk zou schelen".

De opmerkingen ten aanzien van de autonomie zijn gematigd positief (M=0.58; n=19 en M=0.47; n=19). Veel deelnemers gaven aan dat ze het goed vinden dat de rol van AI in deze opties adviserend is, waarbij de docent de uiteindelijke knoop kan doorhakken: "Het geven van een scorevoorstel achteraf lijkt me de goede volgorde; de docent voelt zich dan ook meer autonoom." En: "Fan van optie 3. Controle achteraf is heel mooi, omdat je als docent de verantwoordelijkheid blijft houden maar je wordt wel getest op kritische houding." Wel ontstond regelmatig de vraag hoe AI zich verhoudt tot de eerste en tweede beoordelaar: "Je hebt nu eigenlijk drie mensen die met elkaar discussiëren: AI, eerste en tweede corrector." En: "Wat wordt de rol van beide correctoren?" Deelnemers waren het unaniem eens dat hier, in het geval van de implementatie van AI in het nakijkproces, een duidelijke visie en werkwijze voor zou moeten worden ontwikkeld. Tot slot werd er over de haalbaarheid slechts één opmerking gemaakt in relatie tot beide opties. De deelnemer had vraagtekens bij de haalbaarheid van de opties vanuit het oogpunt van de AI-act, zoals beschreven in de aanleiding.

Algemene opmerkingen naar aanleiding van het scenario

Naast de opmerkingen die daadwerkelijk over de AI-opties gingen, kwamen enkele algemene onderwerpen naar voren die relevant zijn voor de overwegingen rond de inzet van AI bij het nakijken van eindexamens. Het principe van de aanwezigheid van de spellings- en grammaticacontrole binnen een digitaal eindexamen Nederlands werd kritisch onder de loep gelegd en deelnemers vroegen zich af hoe zinvol dit is: "Als je toch digitaal gaat examineren, mogen leerlingen dan geen toegang krijgen tot spellingscontrole? Dit is toch de wereld waar we in zitten." Daarbij werd benadrukt dat digitale examens andere vaardigheden vereisen, zoals het kunnen typen van leestekens als een trema, wat ook nog ongelijkheid tussen leerlingen kan vergroten door verschillen in toegang tot technologie en oefening: "Niet iedere leerling heeft dezelfde kansen, niet elke school heeft laptops."

Daarnaast werd, net als bij het eerste scenario, de rol en de functie van de docent meermaals besproken, in relatie tot de acceptatie van AI in het nakijkproces. Meerdere deelnemers benadrukten het

belang van de rol van de docent in het beoordelingsproces, onafhankelijk van AI-ondersteuning: “Het nakijken van een examen is onderdeel van de beroepspraktijk. Het verschaft inzicht in de kennis en kunde van je leerlingen.” En: “Je leert er veel van, zien wat er goed en fout gaat.” Toch gaven meerderen ook aan dat er wellicht ook sprake is van gewenning: misschien wordt het door de tijd heen normaler om een deel van het nakijkwerk uit handen te geven. Net zoals het gebruik van dergelijke hulpmiddelen in de reguliere lessen, kan het vertrouwd raken met AI-ondersteuning op termijn leiden tot een bredere acceptatie: “Nakijkassistentie komt er ook al in in de reguliere lessen – ook een stukje gewenning.”

5.2.4 Scenario 3. Biologie vmbo bb/kb

In totaal werden er 230 opmerkingen verzameld over de verschillende opties binnen *Scenario 3. Biologie vmbo bb/kb*. In Tabel 3 staat een overzicht per optie en per label van de gemiddelde score en het aantal opmerkingen waar deze op gebaseerd is. Over waarde, kwaliteit en efficiëntie zijn de meeste opmerkingen gemaakt en lopen de meningen per optie sterk uiteen. Zo werden opties 3.2 (markeren van antwoorden) en 3.4 (discussieforum) zeer negatief beoordeeld op het criterium kwaliteit, terwijl opties 3.1 (clustering) en 3.3 (aanvulling antwoordmodel) als zeer positief werden gezien. Over autonomie werden slechts 13 opmerkingen gemaakt, maar de opties werden op dit vlak over het algemeen wel positief beoordeeld. Ook bij haalbaarheid geldt dat het totaal aantal opmerkingen (11) zeer beperkt is, maar hier zijn de opmerkingen uitsluitend negatief, voornamelijk over optie 3.4. In de volgende secties worden de belangrijkste bevindingen uitgewerkt, waarbij we ook dieper ingaan op een aantal voorwaarden die in de focusgroepen werden benoemd om de opties succesvol uit te kunnen voeren.

Tabel 3

Overzicht per AI-optie en per label van de gemiddelde score en het aantal opmerkingen waarop deze is gebaseerd voor scenario 3

	waarde		**kwaliteit**		**efficiëntie**		**autonomie**		**haalbaarheid**			
scenario	gem. score	aantal	gem. score	aantal	gem. score	aantal	gem. score	aantal	gem. score	aantal	gewogen totaal	aantal totaal
3.1	0,53	17	0,55	29	0,65	23	0,83	6	0,00	0	0,60	75
3.2	0,21	24	-0,50	10	-0,21	14	1,00	2	-1,00	1	-0,04	51
3.3	0,81	16	0,62	13	-0,10	10	1,00	3	-1,00	3	0,44	45
3.4	-0,19	26	-0,89	18	0,17	6	0,50	2	-1,00	7	-0,44	59

Optie 3.1 Clustering antwoorden

In totaal werden er 75 opmerkingen gemaakt over de clustering optie, daarmee was het de meest besproken optie binnen dit scenario. De waardering was gemiddeld ook het hoogst, vooral vanwege de verwachte kwaliteitswinst door meer consistentie en de efficiëntie van het tegelijk scoren van meerdere antwoorden.

Op het gebied van de waarde voor de beoordelaar scoort deze optie goed, maar niet uitzonderlijk (M=0,53; n=17). Deelnemers gaven onder andere aan dat ze werken met clustering “heel fijn” zouden vinden, dat ze het als “héél positief” zien en dat het bij korte open antwoorden “echt fantastisch” zou zijn. Die laatste opmerking raakt ook meteen aan een voorwaarde die vaak naar voren kwam om dit een waardevolle optie te laten zijn, namelijk dat het vooral gezien werd als een meerwaarde bij echt korte open antwoorden. Een deelnemer gaf hierover aan: “Clustering is prettig, maar alleen bij korte antwoord vragen. Bij lange antwoord vragen is dat minder overzichtelijk lijkt me.” Als je het echt toespitst op korte open antwoorden, zo voegde een andere deelnemer toe, dan “werkt het echt als een tiet.”

Ook de kwaliteit van de beoordeling kan met behulp van clustering worden verhoogd volgens de deelnemers (M=0.55; n=29). De “foutkans wordt veel kleiner” en je kunt “eerlijker over de leerlingen nakijken.” Daarnaast werd bij deze optie besproken dat de clustering niet alleen op klasniveau hoeft plaats te vinden, maar eventueel ook ondersteund kan worden door een landelijk beeld. Dit moet voorkomen dat leerlingen met een in de klas afwijkend, maar landelijk juist veelvoorkomend antwoord buiten de clustering vallen en benadeeld worden. Een deelnemer zei hierover het volgende: “Juist die verscheidenheid aan antwoorden die worden gegeven, maken een heel overzichtelijk correctiemodel. En je gaat daarna clusteren op basis van landelijke informatie, dan wordt het heel overzichtelijk en [kan het] leiden tot [een] eerlijkere beoordeling.” Los van de basisfunctionaliteit van clustering werd de anonimisering ook als een meerwaarde gezien in het kader van een objectieve beoordeling: “anonimiseren vergroot de objectiviteit.”

Qua efficiëntie komt deze optie zeer goed uit de bus (M=0,65; n=23). De deelnemers geven aan dat de clustering tijd scheelt, vooral wanneer je meerdere antwoorden tegelijk kan scoren. Eén van de deelnemers stipte ook een andere manier aan hoe clustering kan schelen om op een efficiënte manier voor consistentie in beoordelingen te zorgen: “Het clusteren van antwoorden werkt wel sneller nu ga ik vaak terug naar eerdere antwoorden om zeker te zijn dat ik dezelfde score voor een vergelijkbaar antwoord geef.”

Op het vlak van autonomie en haalbaarheid waren er relatief weinig opmerkingen. Het algemene beeld was dat de autonomie van de docent niet werd aangetast en zelfs verder verbeterd kan worden door de mogelijkheid te bieden om de functionaliteit in- en uit te schakelen, en de optie te bieden voor docenten om zelf antwoorden in of uit een cluster te kunnen slepen. Een van de deelnemers zei daarover: “Ben ik het ermee eens hoe ze geclusterd zijn? Daar moet je nog wel in kunnen schuiven.” Op het gebied van haalbaarheid was men vooral sceptisch over de toepassing bij iets langere antwoorden. In de visualisatie bij het scenario waren de antwoorden vrij lang, wat wellicht deze reacties heeft uitgelokt. Er werd onder andere gezegd: “bij lange antwoord vragen is dat minder overzichtelijk lijkt me” en “onoverzichtelijk wanneer langere antwoorden en ook wel complex want hoe zorg je ervoor dat men ook goed blijft kijken naar de andere antwoorden (buiten het cluster)”. Een duidelijke voorwaarde voor de deelnemers is dan ook dat deze optie enkel wordt ingezet bij vragen met korte open antwoorden van hooguit een aantal woorden.

Optie 3.2 Markeringen van woorden uit antwoordmodel

In totaal werden er 51 opmerkingen gemaakt over de optie om woorden uit het antwoordmodel te markeren in de leerlingantwoorden. Het is hiermee op optie 3.3 na de minst besproken optie en uit de opmerkingen blijkt ook dat men over het algemeen weinig nadelen ziet aan deze optie, maar ook weinig echte voordelen.

Er werden voor deze optie relatief veel algemene opmerkingen gemaakt over welke waarde het zou toevoegen, die gemiddeld gezien net iets vaker positief dan negatief waren (M=0,21; n=24). Sommige deelnemers keken er positief naar en zeiden “als je puur op de match tussen antwoordmodel en leerlingantwoord checkt is het denk ik waardevol”. Anderen waren licht positief, maar stelden voorwaarden als dat het een kritische docent vergt: “Markeren goede woorden lijkt mooi, maar dat vraagt wel een kritische docent die leest in wat voor zin die woorden gebruikt worden. Zeker niet zaligmakend." Net als bij eerdere opties werd ook hier de mogelijkheid genoemd om de functionaliteit uit en aan te zetten: “Denk dat het wel fijn is als je hem aan kan zetten en daarna weer uitzetten.”

Over de impact op de kwaliteit van de beoordeling was men minder positief (M=-0,50; n=10). Zo gaf een deelnemer aan: “Ik schrik er al van hoe er nu al nagekeken wordt, en als het voorgekauwd wordt, wordt het misschien nog erger.” Een andere deelnemer wees op het risico dat het markeren van woorden kan leiden tot minder zorgvuldig nakijken: “Foutkans vergroot, woordje staat er in – goed!” Ook het verlies aan overzicht op het scherm werd genoemd als reden waarom de kwaliteit van beoordelingen mogelijk afneemt: “Markeren is verwarrend. Voegt niets toe.” Het belang van een eerlijke beoordeling kwam vaak naar voren in de focusgroep discussies, en zag men bij deze optie als een risico bij leerlingen die afwijkende of beschrijvende antwoorden geven: “Je benadeelt leerlingen die een omschrijving geven van het begrip.”

Over efficiëntie waren deelnemers bij deze optie gemiddeld gezien negatief (M=-0,21; n=14), en ook het negatiefst van alle opties binnen Scenario 3. Zo werd er gezegd: “bij markeren zou het het niet sneller maken” en “gaat heel erg om signaalwoorden, […], weet niet of ik sneller zou nakijken”. Aangezien een docent het leerlingantwoord alsnog volledig moet lezen en begrijpen, werd deze aanpak niet beschouwd als een optie die daadwerkelijk veel tijd bespaart.

Voor deze optie werden de criteria autonomie (2 keer) en haalbaarheid (1 keer) te weinig besproken om er conclusies over te kunnen trekken. Wel werd, zoals al eerder benoemd, de optie geopperd om de functionaliteit aan en uit te kunnen zetten.

Optie 3.3 Aanvulling AI op correctievoorschrift

In totaal werden er 45 opmerkingen gemaakt over de optie om het antwoordmodel of correctievoorschrift aan te vullen met behulp van AI. Het was hiermee de minst besproken optie van Scenario 3, hoewel de opmerkingen over dit scenario vaak wel uitgebreider en inhoudelijker waren dan bij de andere opties. Deze optie werd over het algemeen positief beoordeeld, en men vond het, meer dan bij andere opties, belangrijk om zorgvuldig te bepalen hoe ze precies geïmplementeerd zou worden. Het ging vooral om het moment in het proces, waarbij men vond dat de AI-aanvulling eerst aan de vaststellingscommissie moest worden voorgelegd, dus vóór de definitieve vaststelling en examinering.

Van alle opties over alle scenario’s heen werd deze optie als het meest waardevol gezien voor het beoordelingsproces (M=0,81; n=16). Een deelnemer gaf letterlijk aan dat het zijn “favoriet” was. Een andere deelnemer koppelde de meerwaarde duidelijk aan het criterium autonomie: “Als ik als docent nog steeds de vrijheid krijg om verder te kijken dan deze aanvullingen, zie ik zeker meerwaarde.” Daarnaast gaf een andere deelnemer aan de rol van AI in deze optie te waarderen: “Ja! Soort pre+ correctie, zeker nuttig, dan is AI een tool.”

Ook wat betreft kwaliteit scoort deze optie gemiddeld het hoogst van alle opties (M=0,62; n=13). Eén van de participanten ziet hierbij zelfs de mogelijkheid om “de kwaliteit van de beoordeling te verhogen”. Een andere deelnemer schetst een combinatie met de eerste optie, namelijk clustering, om zowel efficiëntie als consistentie in beoordelingen te bereiken: “Op deze manier wordt er efficiënter gewerkt, omdat elke beoordelaar elke leerling op dezelfde manier beoordeelt door het clusteren van de vragen en het correctiemodel completer is. Het scheelt tijd in het overleggen met een tweede beoordelaar.” Een laatste deelnemer legt nog uit: “Optie 3 lijkt me zeer wenselijk omdat antwoordmodellen niet volledig zijn. Wellicht zou een antwoord model beknopt kunnen zijn (werkbaar) en bij twijfel de optie kunnen geven voor meer mogelijkheden en verduidelijking.” Hierbij kan gedacht worden aan een beknoptere en uitgebreidere versie van het antwoordmodel, waar de docent in een digitale omgeving standaard de beknopte versie ziet en de optie heeft om de uitgebreidere versie te openen.

Over efficiëntie zijn de meningen verdeeld bij deze optie (M=-0,10; n=10). De één ziet een kans om efficiënter te werken, onder andere doordat het tijd scheelt “in het overleggen met een tweede beoordelaar." De ander concludeert dat het niet per se sneller zou zijn omdat een vaststellingscommissie “natuurlijk ook bezig [is] met de juiste antwoorden verzinnen, dan komt AI met nog iets anders.” Vanuit dit oogpunt gezien zou deze optie dus vooral zorgen voor een kwalitatief hoogwaardigere beoordeling, maar niet zozeer een sneller proces.

Over autonomie en haalbaarheid is bij deze optie wederom niet veel gezegd. Net als bij andere opties ziet men echter wel dat, mits een docent zijn vrijheden behoudt, deze optie in principe niet tornt aan de autonomie van de beoordelaar. Een van de deelnemers gaf aan: “De autonomie ligt nog steeds bij de docent, want die krijgt ondersteuning in het benoemen van de juiste hoeveelheid punten. Docent blijft leidend.” Wat betreft haalbaarheid worden een aantal kritische opmerkingen gemaakt, waarbij vooral geldt dat men het als een reële optie ziet mits het ingebed is in het formele proces van de vaststellingscommissie, en anders niet. Een participant vatte dit samen als: “Indien akkoord bevonden voor vc, vooral doen. Indien niet akkoord, vind ik dit een gevaarlijke weg.” Een vergelijkbare opmerking werd ook elders gemaakt: “Vooraf

aan de vaststellingscommissie vragen of dit wel of niet een goede correctie is." En een laatste relevante toevoeging relateerde de waarde die deze optie kan hebben aan het moment in het proces waarin het wordt ingeschoten: "Het lijkt mij heel waardevol als ik een beetje de kaders kan aangeven van welke nuances zijn ook nog goed. Niet achteraf, maar wel echt vooraf."

Het is ook interessant om te vermelden dat bij deze optie enkele deelnemers hebben gesuggereerd om AI nog eerder in de toetscyclus in te zetten, namelijk al bij de constructie van de toetsitems. Daarbij zou AI dan plausibele leerlingantwoorden genereren op basis van een voorlopige versie van een item, om de constructeur inzicht te geven in de mate waarin de vraagstelling en het bijbehorende antwoordmodel op elkaar aansluiten. Eén deelnemer benoemde dat een "andere interpretatie van [een] toetsvraag kan helpen bij het maken van de toetsvraag." Het belang van deze bredere blik en het openstaan voor alternatieve interpretaties werd ook door een andere deelnemer als toegevoegde waarde gezien van een dergelijke inrichting van deze optie: "AI zou ook gebruikt mogen worden bij het opstellen van de toetsvragen, omdat hierin nu nog vaak meerdere interpretaties mogelijk zijn."

Optie 3.4 Forum

In totaal werden er 59 opmerkingen gemaakt over de optie om het discussieforum te integreren in de nakijk omgeving en hierbij een chatbot samenvattingen vanuit het forum te laten presenteren bij vragen vanuit de beoordelaar. Deze optie kreeg van alle opties over alle scenario's heen het meest negatieve commentaar. Het algemene gevoel was dat dit eerder zou leiden tot kwalitatief slechtere beoordelingen dan dat er iets mee gewonnen zou worden.

Onder de deelnemers heerste een licht negatief beeld over de waarde voor de beoordelaar (M=-0,19; n=26). Het kwam er uiteindelijk vaak op neer dat men dit in theorie nog wel zag zitten, maar dat met het beperkte gebruik van discussiefora niet veel waarde zou toevoegen, of zelfs negatief zou uitpakken. Deze opmerking van een deelnemer vat dit beeld goed samen: "Optie 4 forum, lijkt me fantastisch, maar er gebeurt te weinig op het forum, dus AI zou dan dus zeggen ik kan niks vinden."

Wat betreft het effect op de kwaliteit van de beoordeling werd deze optie zeer negatief beoordeeld (M=-0,89; n=18), het meest negatief van alle onderzochte opties. Dit kwam vooral door de meningen die deelnemers hadden over de huidige hoeveelheid en kwaliteit van de inhoud van de fora. Zo werd er gezegd: "Communiceren met een leeg forum heeft geen zin." Ook het waarheidsgehalte van de opmerkingen werd in twijfel getrokken: "Op forum reageren veel mensen die meer ruimte zouden willen zien in CV of een bepaalde sterke mening heeft, maar dat is niet altijd goed of waar." En ook: "Bij Facebook groep wiskunde papieren examens is het gevaarlijk dat er soms onzin beweerd wordt, dan zou het met die AI-samenvatting opeens lijken alsof het het goede antwoord is." Een ander effect van deze optie die aan bod kwam was dat het een oneerlijke situatie zou veroorzaken doordat docenten van scholen met een vroeg afnamemoment dan minder ondersteuning krijgen dan docenten die later in de afnameperiode moeten nakijken. Er werd onder andere aangegeven: "Optie 4: Gezien de spreiding van de examens lijkt me dat niet fair voor de eerst nagekeken leerlingen." En: "Tijdlijn flexibele afname maakt het ook problematisch, docenten die in april een vraag hebben, kunnen eigenlijk nog geen antwoord krijgen."

Er werden weinig opmerkingen gemaakt over efficiëntie, autonomie en haalbaarheid, en deze bevestigden vooral de algemene perceptie dat de optie weinig waardevol was. Ook de praktische haalbaarheid qua invoering werd in twijfel getrokken, waarbij een paar keer werd benoemd dat men de verwachting had dat CvTE deze optie niet zou toelaten. Zo werd er onder andere gezegd: "Als AI op basis van het forum iets bedenkt dan geeft dat andere informatie aan forumgebruikers dan aan de rest van het veld, dit zal CvTE niet snel goedkeuren"

Algemene opmerkingen naar aanleiding van het scenario

Een aantal thema's kwamen meerdere keren terug in de bespreking van het scenario. Zo werd het aan- of uitzetten van een functionaliteit, afhankelijk van de voorkeur van de docent, vaak genoemd als middel om

autonomie te waarborgen. Daarnaast kan dit voorkomen dat docenten die geen meerwaarde zien in een functionaliteit, gefrustreerd raken doordat ze deze toch moeten gebruiken. Ook de mogelijkheid om opties 3.1 (clustering) en 3.3 (aanvulling correctievoorschrift/antwoordmodel) te combineren werd genoemd. Dit zou logisch zijn vanuit het oogpunt dat beide opties gemiddeld positief werden beoordeeld, maar heeft ook een inhoudelijke logica aangezien de verbreding van het antwoordmodel kan helpen bij een betere clustering en bijbehorende keuzes van de docent.

Daarnaast werden ook enkele bredere onderwerpen naar voren, waaronder anonimisering. Dit hoorde formeel bij optie 3.1, maar werd door meerdere docenten een aantal keer genoemd als een waardevolle toevoeging om de objectiviteit van beoordelingen te bevorderen. Hoewel deze functionaliteit niet de kern van de optie vormde, lijkt het de moeite waard om anonimisering verder te verkennen: wordt deze wens breed gedragen door docenten en andere relevante stakeholders, en hoe zou dit praktisch geïmplementeerd kunnen worden?

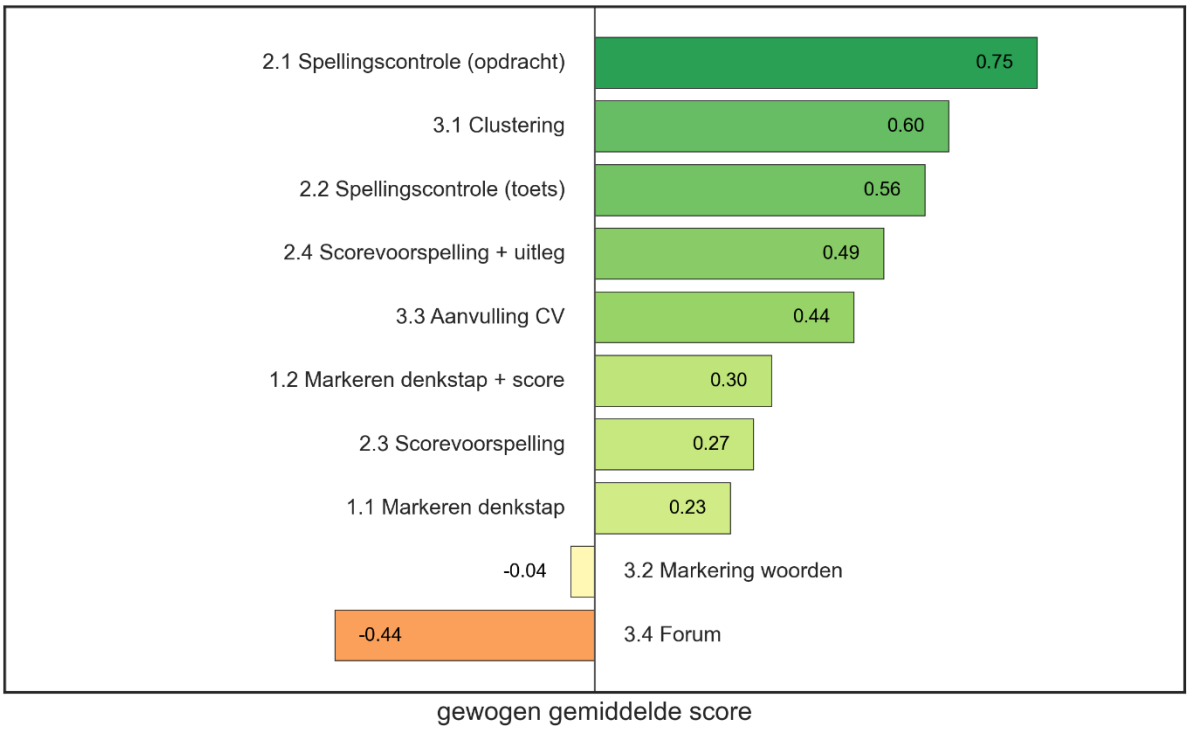
Het belang van een eerlijke beoordeling en gelijke kansen voor alle leerlingen werd herhaaldelijk benadrukt. De garantie dat de beoordeling eerlijk blijft, werd daarbij vaak genoemd als een belangrijke voorwaarde voor het willen invoeren van een AI-optie. Als laatste werden ook een paar opmerkingen gemaakt over de relatie tussen de autonomie van de docent en de inzet van AI. Een deelnemer gaf aan: “Autonomie en veel AI bijt elkaar. Tenminste als het om tijdswinst gaat.” Over het algemeen was het beeld van deelnemers wel dat er in de opties binnen dit scenario een mooie balans was gevonden: “Autonomie blijft in alle gevallen mooi in balans en bij de docent. Het zijn tools die kunnen helpen maar die niet de taak van de docent overnemen.”

6. Conclusie en discussie

6.1 Samenvatting bevindingen

In dit rapport hebben we onderzocht op welke wijze AI kan worden ingezet bij de beoordeling van CE's, zodat de kwaliteit van de beoordeling omhooggaat, zonder dat de autonomie van de beoordelaar wordt aangetast, en hoe een dergelijke inzet van AI het correctieproces ook efficiënter kan maken. Hiertoe zijn drie scenario's ontworpen met in totaal tien onderliggende opties voor het gebruik van AI bij het nakijken van verschillende vakken (wiskunde, Nederlands, biologie) en vraagtypen, met verschillende niveaus van automatisering van Molenaar (2022). Er is gekeken welke mogelijkheden deze scenario's bieden om het nakijkproces efficiënter en consistenter te maken. Deze scenario's zijn in vijf focusgroepen voorgelegd aan 26 stakeholders uit de examenketen, waarbij is gevraagd hoe ze kijken naar het gebruik van AI in deze scenario's op het gebied van kwaliteit (consistentie), efficiëntie, autonomie, waarde en haalbaarheid. In deze sectie worden de belangrijkste bevindingen samengevat en implicaties besproken voor beleid en praktijk.

Figuur 5
Gewogen gemiddelde scores van de verschillende opties van de drie scenario's



De scenario's tonen dat de potentie van AI sterk afhankelijk is van het vak en vraagtype. In Figuur 5 staan de gewogen gemiddelde scores van de afzonderlijke opties per scenario. Bij wiskunde A op havo-niveau (scenario 1) lag de focus op het herkennen en markeren van denkstappen (optie 1.1) en het doen van scoresuggesties voor deelstappen (optie 1.2): deelnemers zien hier mogelijkheden voor meer overzicht en tijdwinst bij complexe opgaven, maar stellen kritische vragen over bijvoorbeeld de accuraatheid van de herkenning van de denkstappen, vooral bij afwijkende oplossingsstrategieën, en over de correcte toepassing van vakspecifieke nakijkregels.

Bij Nederlands op vwo-niveau (scenario 2) zien stakeholders directe winst in het automatisch detecteren van spellings- en grammaticafouten (optie 2.1 en 2.2): dit wordt gezien als een relatief laagdrempelige toepassing met duidelijke winst in efficiëntie. Verdergaande opties – zoals puntentoekenning door AI en bijbehorende zekerheidsinschattingen (optie 2.3 en 2.4) bij open antwoorden – worden meer verdeeld ontvangen: ze bieden potentieel steun, maar vragen om transparantie, uitlegbaarheid en eenduidige interpretatierichtlijnen om ongewenste beïnvloeding of extra overleglast en verwarring te voorkomen.

Voor het derde scenario biologie vmbo bb/kb, waar digitale afname al gangbaar is, werd clustering van identieke of vergelijkbare korte antwoorden (optie 3.1) door veel deelnemers gezien als een optie met duidelijke meerwaarde. Clustering kan volgens stakeholders zowel de consistentie verhogen als tijd besparen, doordat meerdere antwoorden gelijktijdig beoordeeld kunnen worden; anonimiseren binnen zo'n clustering werd daarbij als een nuttige waarborg voor objectiviteit genoemd. Ook de optie om AI-aanvullingen op het correctievoorschrift te genereren (optie 3.3) kreeg brede waardering, mits die aanvullingen vooraf door vaststellingscommissies worden geaccordeerd. De integratie van een AI-samenvattende chatbot op het discussieforum (optie 3.4) werd duidelijk minder positief ontvangen; deelnemers wezen op de beperkte en soms onbetrouwbare inhoud van het forum, op ongelijkheid tussen vroege en late nakijkmomenten en op het risico dat AI-forum-samenvattingen een onterechte autoriteit krijgen. Tot slot werd de meerwaarde van het automatisch markeren van sleutelwoorden uit het antwoordmodel in leerlingantwoorden (optie 3.2) niet direct gezien: sommige deelnemers zagen dit als nuttige ondersteuning, maar anderen vonden het juist verwarrend en vreesden dat markeringen gemakzuchtig nakijken stimuleren of dat antwoorden ongelijkwaardig worden behandeld.

6.2 Terugkerende thema's ter overweging

In bovenstaande sectie zijn de scenario-specifieke bevindingen samengevat. In dit onderdeel reflecteren we op een aantal overkoepelende thema's die terugkwamen in de focusgroepen. Allereerst raakt de inzet van AI in het nakijkproces de taak- en verantwoordingsverdeling in het correctieproces. De algemene tendens vanuit de focusgroepen was dat de docent als eerste beoordelaar de primaire verantwoordelijkheid moet behouden. AI-functionaliteiten moeten daarom in eerste instantie ondersteunend en optioneel zijn: markeringen, clustering en scoresuggesties kunnen de beoordelaar helpen om sneller en consistenter te beslissen. In het huidige systeem wordt het examenwerk nog een keer nagekeken door een tweede beoordelaar van een andere school, als een kwaliteitsborg. Als AI ook een bepaalde rol inneemt in het correctieproces, bijvoorbeeld door het geven van scoresuggesties, kunnen we ons afvragen hoe deze eerste, tweede en zogenaamde 'derde beoordelaar' zich tot elkaar verhouden en wat dit doet met de transparantie. Mogelijke rollen voor AI zijn bijvoorbeeld: het geven van scoresuggesties vóór menselijke correctie, een controle uitvoeren ná menselijke correctie, parallel beoordelen naast de menselijke beoordelaar, of het vervangen van de tweede beoordelaar. Dit riep in de focusgroepen vragen op over de status van een dergelijke AI-beoordeling: in hoeverre weegt een AI-advies even zwaar als een menselijke beoordeling, wie neemt verantwoordelijkheid bij fouten en hoe worden verschillen tussen mens en machine opgelost? Maar ook: in hoeverre is de tweede beoordelaar van een andere school nog nodig, als AI als een 'algemene en onafhankelijke beoordelaar' fungeert in heel Nederland? Deze en soortgelijke vraagstukken vragen om duidelijke beleids- en procesafspraken alvorens AI geïmplementeerd kan worden in het officiële correctieproces.

Hieraan verwant is het thema eerlijkheid. Dit was geen expliciet criterium dat is meegenomen, maar kwam wel meerdere keren terug in de focusgroepen als belangrijk onderwerp. Dit gaat allereerst om bias en representativiteit: AI-modellen presteren het best voor groepen en taalgebruik die in de data

terugkomen waarop ze zijn getraind. Als trainingsdata vooral Standaardnederlands en/of antwoorden uit bepaalde regio's of scholen bevatten, lopen bijvoorbeeld NT2-leerlingen of leerlingen met een ander taalregister het risico structureel strenger beoordeeld te worden. Er werd in de focusgroepen opgemerkt dat dit niet alleen van toepassing is op talig niveau, maar ook bij bijvoorbeeld een vak als wiskunde: een creatieve leerling kan met andere stappen dan voorgeschreven op het cv komen tot het correcte eindantwoord; is AI ook in staat om deze correcte, afwijkende uitwerkingsvorm goed te interpreteren? Daarnaast speelt temporele eerlijkheid: opties zoals een AI-samenvattende chatbot kunnen ongelijkheden versterken als docenten die vroeg in de afnameperiode nakijken minder toegang hebben tot opgeloste discussies dan docenten die later nakijken. Tot slot betreft eerlijkheid ook de toegankelijkheid en gelijke condities bij digitale afname: digitale examens vereisen andere vaardigheden, wat ongelijkheid tussen leerlingen kan vergroten, onder andere door verschillen in leerlingkenmerken en de bredere sociale en nationale context (Dodds et al., 2025).

De context van de inzet van AI bij het nakijken – high stakes CE's versus lower stakes toetsen zoals huiswerkopdrachten – maakt een wezenlijk verschil in wat acceptabel en/of wenselijk wordt bevonden. Voor formatieve en schooleigen toepassingen is er meer speelruimte: verschillende docenten in de focusgroepen gaven aan reeds gebruik te maken van nakijktoepassingen met AI, en feedback kan in deze applicaties worden gebruikt om leren te ondersteunen; gebreken van AI hebben op dit gebied minder verstrekkende gevolgen. Daarom zagen deelnemers meer mogelijkheden om daarin sneller te experimenteren met meer autonome AI-functionaliteiten. Voor CE's ligt dit anders: omdat de beslissingen zwaarwegende consequenties hebben en betrouwbaarheid en validiteit cruciaal zijn, is een voorzichtige, gefaseerde aanpak vereist. AI-toepassingen waar een beperkte procedurele taak wordt uitgevoerd, zoals automatische spellings- en grammaticacontrole of clustering van open antwoorden, komen dan als eerste stap in aanmerking.

Overwegingen over de rechtmatigheid van de inzet van AI hangen nauw samen met de AI-verordening. AI-systemen die beslissingen of evaluaties ondersteunen in high-stakes onderwijscontexten vallen (uitzonderingen daargelaten) onder de hoog risico classificatie en moeten daarmee aan strenge eisen voldoen. Dit betekent dat elk scenario voor CE's niet alleen technisch moet worden gevalideerd, maar ook juridisch en organisatorisch getoetst. Daarnaast moet duidelijk zijn wie het algoritme voor dergelijke AI-beoordelingen levert en beheert, en moet het model goed gedocumenteerd en verifieerbaar zijn. Dit is cruciaal om verantwoordelijkheid en aansprakelijkheid te kunnen vaststellen, de uitkomsten en processen te kunnen controleren, data op een verantwoorde manier te beheren en de betrouwbaarheid van het systeem te waarborgen.

6.3 Suggesties voor vervolgonderzoek

We sluiten af met enkele suggesties voor vervolgonderzoek, waarin de beperkingen van het huidige onderzoek eveneens worden aangestipt. Allereerst dient onderzoek gedaan te worden naar de mate van inconsistenties in het nakijken van eindexamens en wat de aard is van dergelijke inconsistenties. Analyses zijn voor nu makkelijker te faciliteren voor digitale examens met korte open antwoord vragen dan voor papieren examens of examens met lange open antwoord vragen die zich niet makkelijk lenen voor het identificeren van inconsistenties. Mogelijk voor sommige delen van bijvoorbeeld biologie vmbo bb/kb examens, omdat deze al digitaal worden afgenomen, maar voor papieren examens en veel andere vakken ontbreekt de data, waardoor er nog geen vergelijkend onderzoek kan worden gedaan. Zodra meer leerlingantwoorden en bijbehorende scoringsdata digitaal beschikbaar komen, kan ook worden onderzocht welke factoren een belangrijke rol spelen bij het ontstaan van inconsistenties, zoals summiere antwoordmodellen, vraagonduidelijkheid of beoordelaarsvariatie. Daarbij is het van belang om per bron te

bepalen of en hoe AI kan bijdragen aan het verminderen van deze inconsistenties, bijvoorbeeld door het signaleren van beoordelaarsafwijkingen of het structureren van antwoordvarianten. Dit helpt om gericht te bepalen voor welke onderdelen van het correctieproces AI daadwerkelijk meerwaarde kan hebben.

Wat betreft de werkwijze in het huidige onderzoek, kwam in de rondvraag aan het einde van de focusgroepen meerdere keren naar voren dat men de concrete werkwijze met scenario's en bijbehorende schetsen goed en fijn vond. Stakeholders gaven bijvoorbeeld aan: "Hebben jullie nog meer scenario's? Op deze manier zou ik er nog wel twintig willen bespreken." Reeds bij het opstellen van de scenario's, maar ook bij het uitvragen van de meningen, bleek dat de uitwerking van meer scenario's per vak, schooltype en vraagtype cruciaal is, omdat de voor- en nadelen van AI sterk contextafhankelijk blijken. De drie scenario's in het huidige onderzoek maken dit al duidelijk: clustering werkt bijvoorbeeld goed bij korte open antwoorden van biologie vmbo bb/kb examens, maar kan bij lange, meer beschrijvende antwoorden snel onoverzichtelijk en eerder verwarrend werken. Het vak wiskunde vraagt eerder om functies die denkstappen kunnen herkennen en partiële scores kunnen voorstellen, terwijl bij een vak als Nederlands de spellings- en grammaticacontrole juist een directe winst in efficiëntie biedt. Dit pleit voor een werkwijze waarin per vak, schooltype en vraagtype concreet wordt uitgesplitst en uitgewerkt welke functionaliteiten mogelijk en behulpzaam zijn, en dat deze afzonderlijk worden onderzocht, getest en gevalideerd – en waar nodig gecombineerd. Het is belangrijk om de schaal van automatisering (Molenaar, 2022) expliciet mee te nemen in elk scenario en hierin bewuste keuzes te maken. In het huidige onderzoek lag de focus met name op de linkerkant van de schaal, dus met nog veel autonomie bij de docent. Wel werd meerdere keren gesuggereerd door deelnemers om juist scenario's op te stellen die zich meer aan de rechterkant van de schaal bevinden, dus opties waarbij AI meer autonoom is en de docent een meer controlerende rol vervult. Hier is wel voorzichtigheid geboden. Uit de bevindingen blijkt namelijk dat de meest positief beoordeelde toepassingen juist het minst ingrijpend zijn en het meest aansluiten bij bestaande juridische en ethische kaders. Scenario's waarbij AI-systemen autonoom handelen zijn vaak niet alleen ingrijpender, maar sluiten ook minder goed aan bij richtlijnen zoals die vanuit de AI-verordening. In sommige gevallen kan het interessant zijn om te filosoferen over toepassingen die buiten de huidige kaders vallen, omdat dit nieuwe inzichten kan opleveren over de mogelijkheden van AI en de grenzen van bestaande richtlijnen kan verkennen. Tegelijkertijd laten de resultaten zien dat men ook met realistische, haalbare scenario's al een vergaande discussie kan voeren.

Het verkennen van de implementatie van de AI-opties vraagt om een gefaseerde aanpak. Zoals reeds eerder vermeld, komen de laag-risicotoepassingen zoals een automatische spellings- en grammaticacontrole als eerste in aanmerking voor pilots. In het Digitaliseringsprogramma wordt momenteel bijvoorbeeld gewerkt aan het uitwerken en verder onderzoeken van *optie 3.3 Aanvulling AI op correctievoorschrift*. Parallel hieraan kunnen gecontroleerde experimenten worden opgezet voor opties zoals denkstapmarkeringen en scoresuggesties bij wiskunde. Hiervoor is meer onderzoek nodig, waarin bijvoorbeeld een daadwerkelijk algoritme wordt opgezet. Dit vereist onderzoek naar hoe goed een dergelijk algoritme moet presteren, voordat de inzet ervan acceptabel is. Hierbij is het belangrijk om te beseffen dat er geen eenduidige consensus zal bestaan over deze grens. Bovendien zal die grens in de loop van tijd verschuiven. Onderzoek levert daarmee slechts een momentopname. Daarnaast hangt de acceptabele prestatieniveau sterk samen met de mate van autonomie en invloed van het systeem. Een autonoom systeem met grote impact moet aantoonbaar zeer betrouwbaar functioneren, terwijl voor een systeem met beperkte autonomie en invloed mogelijk een lager prestatieniveau volstaat.

Door bewijzen aan te leveren van een betrouwbare en valide werking van AI-toepassingen, kan breder draagvlak en vertrouwen worden gecreëerd. Bij alle toepassingen – zowel met een hoog als laag risico – gelden de volgende cruciale elementen: vakinhoudelijke validatie, heldere governance (wie ontwikkelt, wie beheert, wie is aansprakelijk), transparantie naar gebruikers en leerlingen, en gerichte scholing van beoordelaars om het gebruik van AI verantwoord in te bedden. Alleen op deze manier kan AI bijdragen

aan meer consistente en efficiënte beoordelingen, zonder de professionele autonomie en het vertrouwen in het examenstelsel te ondermijnen.

Bibliografie

Adriaens, H., Feher, B., Elshout, S., & Elshout, M. (2024). Personeelstekorten voortgezet onderwijs: peildatum 1 oktober 2024. https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2024/12/17/rapporten-bij-trendrapportage-arbeidsmarkt-leraren-po-vo-en-mbo-2024/Personeelstekorten+voortgezet+onderwijs+2024.pdf

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics, 1*, 391–402. https://doi.org/10.1162/tacl_a_00236

Buisman, M., Krepel, A., Mariën, H., Lans, R. van der, Diepstraten, I., Stronkhorst, E., & Farzan, K. (2025). TALIS 2024: De werk- en leeromgeving op school in beeld. Nederlandse resultaten van de internationale TALIS-survey. https://www.talis2024.nl/wp-content/uploads/2025/10/1143-TALIS-2024.pdf

Bureau voor publicaties van de Europese Unie (2024). *Verordening (EU) 2024/1689 tot vaststelling van geharmoniseerde regels betreffende artificiële intelligentie (verordening artificiële intelligentie)*. Geraadpleegd op 21 oktober 2025 via Verordening - EU - 2024/1689 - EN - EUR-Lex

Caraeni, A., Scarlatos, A., & Lan, A. (2025). Evaluating GPT-4 at grading handwrittten solutions in math exams. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 2025. https://doi.org/10.48550/ARXIV.2411.05231

Van Casteren, W., Lodewick, J., Lommertzen, J., Luyten, E., & Van Mensvoort, C. (2023). Vertrekredenen leraren en docenten in het po, vo, en mbo. ResearchNed. https://www.voion.nl/publicaties/vertrekredenen-leraren-en-docenten-in-het-po-vo-en-mbo

CvTE (2025). *Onderzoek naar verdere digitalisering centrale examinering in het voortgezet onderwijs.* Geraadpleegd op 21 oktober 2025 via Onderzoek naar verdere digitalisering centrale examinering in het voortgezet onderwijs | CvTE

Del Gobbo, E., Guarino, A., Cafarelli, B., & Grilli, L. (2023). Gradeaid: A framework for automatic short answers grading in educational contexts–design, implementation and evaluation. *Knowledge and Information Systems, 65*(10), 4295–4334. https://doi.org/10.1007/s10115-023-01892-9

Dodds, H. J., El Masri, Y., Stratton, T. (2025). On-screen Assessment and Mode Effects: A review of the effects of on-screen assessment on levels of engagement, cognitive demands and performance. *Ofqual research report 25*/7280. Ofqual. https://www.gov.uk/government/publications/on-screen-assessment-and-mode-effects

Essen, A., Van der Ploeg, S., & Vermeulen, J. (2023). Correctielast centrale examens vo. CAOP, Oberon. https://www.examenblad.nl/media/1019

Examenblad (2025). *Bb en kb-flex, flexibele en digitale centrale examens bb en kb*. Geraadpleegd op 21 oktober 2025 via Flexibele en digitale centrale examens bb en kb (BB- en KB-flex) | 2025 | Examenblad.nl

Europees Parlement. (2024). *AI-verordening: eerste regels voor artificiële intelligentie*. Geraadpleegd op 18 juli 2024 via https://www.europarl.europa.eu/topics/nl/article/20230601STO93804/ai-verordening-eerste-regels-voor-artificiele-intelligentie

Flodén, J. (2024). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal, 51*(1), 201-224. https://doi.org/10.1002/berj.4069

Gobrecht, A., Tuma, F., Möller, M., Zöller, T., Zakhvatkin, M., & Wuttig, A. (2024). Beyond human subjectivity and error: a novel AI grading system. *ArXiv, abs/2405.04323*. https://doi.org/10.48550/arXiv.2405.04323

Van den Heuvel, S., & De Vroome, E. (2023). Werkdruk in het onderwijs. TNO Public. https://monitorarbeid.tno.nl/publicaties/werkdruk-in-het-onderwijs/

Horbach, A., Palmer, A., & Wolska, M. (2014). *Finding a tradeoff between accuracy and rater's workload in grading clustered short answers*. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 588–595). European Language Resources Association (ELRA).

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020). Defining AI in policy versus practice. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. https://doi.org/10.1145/3375627.3375835

Kukich, K. (1992). Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys*, *24*(4), 377–439. https://doi.org/10.1145/146370.146380

Mieskes, M., & Padó, U. (2018). *Work Smart – Reducing Effort in Short-Answer Grading*. In *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018)* (pp. 57–68). LiU Electronic Press.

Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *European Journal of Education*, 57(4), 632-645. https://doi.org/10.1111/ejed.12527

NOLAI. (2025). *AI in onderwijs* (pp. 36-37). https://www.ru.nl/sites/default/files/2025-09/nolai-magazine-2025-nederlands.pdf

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural-language tutoring. *International Journal of Artificial Intelligence in Education, 24*(4), 427–469. https://doi.org/10.1007/s40593-014-0029-5

OECD (2025), Education at a Glance 2025: OECD Indicators, OECD Publishing, Paris, https://doi.org/10.1787/1c0d9c79-en

Page, J., Broady, T., Kumar, S., & de Leeuw, E. (2022). Exploratory visuals and text in qualitative research interviews: How do we respond? *International Journal of Qualitative Methods, 21*. https://doi.org/10.1177/16094069221110302

Poell, T., Maas, L., Balk, M., & Van Haastrecht, M. (2025). "SchrijfBlik: safeguarding the validity of writing assessment in the age of AI," in *Proceedings of the 26th International Conference on Artificial Intelligence in Education*, eds. A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, and S. Isotani (Cham: Springer Nature Switzerland), 71–80.

Rijksoverheid (2025). NOLAI | Nationaal Groeifonds

SLO (2025). *Actualisatie kerndoelen en examenprogramma's*. Geraadpleegd op 21 oktober 2025 via Actualisatie kerndoelen en examenprogramma's - SLO

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv. https://arxiv.org/abs/2201.11903

Wetzler, E. L., Cassidy, K. S., & Wood, M. (2024). Grading the graders: Comparing generative AI and human assessment in essay evaluation. *Teaching of Psychology, 52*(3), 298-304. https://doi.org/10.1177/00986283241282696

Bijlage A Samenstelling focusgroepen

Tabel A1
Verdeling van het aantal en type deelnemers over de verschillende focusgroepen

Focusgroep	Vertegenwoordigde organisatie(s) en/of functie(s)
1	- Nederlandse Vereniging van Wiskundeleraren (NVvW)
	- Federatie van Onderwijsvakorganisaties (FvOv)
	- Docent biologie
	- Docent en sectiehoofd biologie en wiskunde
	- Lid Vaststellingscommissie (VC) en docent wiskunde
2	- VO-raad (2 deelnemers)
	- Inspectie van het Onderwijs
	- Toetsdeskundige wiskunde (3 deelnemers)
	- Docent wiskunde
3	- Docent biologie en LO
	- Docent biologie
	- College voor Toetsen en Examens
	- Toetsdeskundige mbo
4	- College voor Toetsen en Examens
	- VC-lid en docent aardrijkskunde
	- VC-lid en docent Arabisch
	- VC-lid en docent wiskunde
5	- Docent Engels en constructiegroeplid
	- Docent en vaksectievoorzitter biologie
	- College voor Toetsen en Examens
	- VC-lid en docent Media, Vormgeving en ICT
	- VC-lid en docent wiskunde
	- VC-lid en docent Informatie, Media Vormgeving, ICT en Game Design