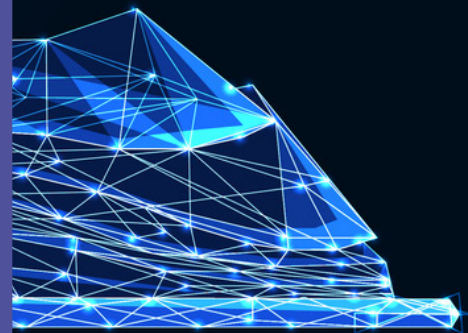




Er bestaat een brede
consensus dat we AI
alleen moeten gebruiken
in het onderwijs als het
processen van leren en
doceren ondersteunt.



AI voor onderwijskundig meten

Waar zit de waarde voor het onderwijs?

■ Max van Haastrecht

Dr. M. van Haastrecht is Onderwijskundig Onderzoeker bij CitoLab¹, Stichting Cito en docent bij de Master Toets- en Onderwijskwaliteit van Fontys Educatie in Tilburg.
E-mail: max.vanhaastrecht@cito.nl

■ Marjolein Jongbloed-Pereboom

Dr. M. Jongbloed-Pereboom is Onderwijskundig Onderzoeker bij CitoLab, Stichting Cito

■ Lotte de Groot

L. de Groot is Psychometrisch Onderzoeker bij CitoLab, Stichting Cito

■ Franka Buytenhuijs

Dr. F. Buytenhuijs is Psychometrisch Onderzoeker bij CitoLab, Stichting Cito

■ Joost Kruis

Dr. J. Kruis is Psychometrisch Onderzoeker bij CitoLab, Stichting Cito

Dit artikel is gebaseerd op het Engelstalige artikel “Artificial Intelligence for Educational Measurement: Where is the Value for Education?” dat eerder is gepubliceerd in *Frontiers in Education* (Van Haastrecht et al., 2025).

Artificiële intelligentie (AI) is niet intrinsiek waardevol voor ons onderwijs, maar ontleent haar waarde aan het bijdragen aan onderwijskundige doelen. Op het vlak van onderwijskundig meten is het nog vaak onduidelijk of en hoe AI precies kan helpen om die doelen te bereiken. In dit artikel bespreken we drie kernelementen om over na te denken bij de integratie van AI in toetsing: de onderwijskundige context waarin toetsing plaatsvindt, de prioritering van verschillende facetten van validiteit en het sociale contract tussen leerling en docent. We lichtten de kernelementen verder toe en laten zien hoe we ze hebben toegepast bij verschillende praktijkvoorbeelden, waaronder de CitoLab prototypes SchrijfBlik en CheckMate. We hopen dat deze kernelementen en de bijbehorende manier van denken aanzetten tot reflectie over wat wij in Nederland waardevol vinden in het onderwijs en hoe we, met of zonder hulp van AI, toetsen kunnen ontwerpen die recht doen aan deze waarden.

¹CitoLab is de onderzoeks- en innovatieafdeling van Stichting Cito.

Onderwijskundig meten en AI

Onderwijs is mensenwerk, waar de interacties tussen leerlingen en docenten centraal staan. Dit verklaart wellicht het oncomfortabele gevoel dat veel onderwijsprofessionals bekruipt als technologische innovaties de menselijke rol in het onderwijs dreigen te veranderen. Nergens is dit gevoel binnen het onderwijsveld sterker dan bij de ontwikkelingen omtrent AI. Er bestaat een brede consensus dat we AI alleen moeten gebruiken in het onderwijs als het processen van leren en doceren ondersteunt (Molenaar, 2022), maar er zijn verschillende ideeën over welke processen precies leiden tot goed onderwijs (Biesta, 2015). Voordat we het gesprek aan kunnen gaan over een mogelijke rol voor AI bij een bepaalde taak in een bepaalde context, moeten we dus een duidelijk beeld hebben van wat de betrokkenen in die context zien als waardevol en goed onderwijs.

Onderwijskundig meten draait in de kern om een proces van bevraging, of in goed Engels “inquiry” (Delandshere, 2002). Dit heeft als doel om inzicht te krijgen in de vaardigheden van een leerling om bij te kunnen sturen in de richting van de onderwijskundige doelen die we nastreven. Onderwijskundig meten is daarmee een middel om het doel van goed onderwijs te bereiken. Op eenzelfde manier kan AI een middel zijn richting het doel van goede toetsing. Er is echter geen garantie dat AI-interventies een positieve impact hebben op toetsing. Met de huidige uitdagingen over bias, uitlegbaarheid en betrouwbaarheid van AI-systemen, is er een reële kans dat we het opgebouwde vertrouwen in toetsing verspelen door AI ondoordacht te implementeren (Aloisi, 2023). Deze uitdagingen spelen in verhoogde mate bij generatieve AI-systemen gebaseerd op grote taalmodellen, die van nature probabilistisch en complex zijn. We moeten ons daarom afvragen wanneer AI-systemen beter onderwijs faciliteren en wanneer niet.

In dit artikel introduceren we een manier van denken die helpt om te begrijpen in welke situaties AI-systemen waarde toevoegen via verbeterde toetsing, en in welke situaties AI-systemen het onderwijs eerder verarmen dan verrijken. Een dergelijke verarming kan bijvoorbeeld ontstaan door de verstoring van de constructieve afstemming tussen onderwijskundige doelen, onderwijsactiviteiten en toetsing (Biggs, 1996). We baseren deze manier van denken op twee onderliggende principes. Ten eerste, dat technologische innovaties zoals AI niet in zichzelf waardevol zijn voor het onderwijs, maar dat ze waarde ontleen aan hun bijdrage aan onderwijskundige doelen. Ten tweede, dat de manier waarop AI waardevol kan zijn in een bepaalde onderwijskundige context afhankelijk is van wat de mensen in die context waardevol vinden.

De waarde van AI traceren

De Europese Unie (EU) definieert in de AI-verordening een AI-systeem 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.”
[Europese Unie, 2024, p.46, Artikel 3 (1)]

Deze definitie geeft niet alleen een algemeen beeld van wat volgens de EU een AI-systeem is, maar benoemt ook een aantal specifieke kenmerken van AI-systemen, zoals de mate waarin het systeem leerlingdata vereist als input (Williamson et al., 2020) en het niveau van autonomie van het AI-systeem (Molenaar, 2022). Deze kenmerken kunnen we zien als knoppen van het systeem waar we aan kunnen draaien om zo de impact bij inzet in het onderwijs te veranderen. Om de waarde van een AI-systeem voor onderwijskundig meten

te kunnen bepalen is een compleet beeld nodig van de impact van het systeem op hoe we betekenis toekennen aan het bewijs dat voortkomt uit toetsing. Oftewel, we moeten precies weten hoe de verschillende knoppen van het systeem zijn ingesteld.

In de volgende paragraaf evalueren we de waarde van verschillende AI-systemen aan de hand van drie kernelementen van onderwijskundig meten. De eerste van deze drie kernelementen is de onderwijskundige context. Een voorwaarde voor het ontwikkelen van betekenisvolle toetsing is het kunnen specificeren van de onderwijskundige doelen die je met toetsing wil bereiken en hoe die doelen zich tot elkaar verhouden binnen de context. Om te bepalen of een AI-systeem waardevol is moet daarom duidelijk zijn waar de prioriteiten liggen in de balans tussen kwalificatie, socialisatie en persoonsvorming (Biesta, 2015), en of je het toetsingsvraagstuk wilt aanvliegen via toetsing van leren, toetsing voor leren of toetsing als leren (Earl, 2003).

Het tweede element is het beeld dat er binnen de context heerst over wat valide toetsing is. De verschillende eisen die aan bod komen bij de vraag wat kwalitatief hoogwaardige toetsing nou precies behelst, kunnen worden uitgesplitst in criteria zoals validiteit, betrouwbaarheid, transparantie, uitvoerbaarheid, enz. In dit artikel behandelen we validiteit echter als een overkoepelend concept (Messick, 1989), waaronder alle facetten vallen die relevant zijn voor de vraag of we meten wat we willen meten. Hierbij gaat het niet alleen om het meetinstrument zelf, maar ook om het gebruik van dat instrument en de, soms langere termijn, consequenties hiervan (Kane, 2013). Binnen dit element is het belangrijk om te weten welke vaardigheden je in kaart wil brengen, hoe het AI-systeem de meting van vaardigheden beïnvloedt en op welke facetten van validiteit je meer of minder nadruk legt. Hiermee bedoelen we niet dat bepaalde

facetten per definitie voorrang krijgen boven anderen, maar dat expliciet moet zijn welke facetten in deze context meer of minder nadruk krijgen (Van Haastrecht et al., 2024; Williamson et al., 2012).

Het derde element is het sociale contract tussen leerling en docent (UNESCO, 2021). Wanneer een AI-systeem zich mengt in de belofte van een leerling om eerlijk werk af te leveren, of in de belofte van een docent om dat werk eerlijk te beoordelen, dan roept dit fundamentele vragen op over wat we precies met onze toetsing beogen. Als we willen bepalen of een AI-interventie een waardevolle bijdrage levert aan toetsing, moeten we weten hoe de afstemming tussen leerling en docent over onderwijskundige doelen wordt beïnvloed, en hoe verschillen in interpretatie over de rol van het AI-systeem de perceptie van leerlingen en docenten kleurt (Williamson et al., 2012). Het sociale contract is een onderbelicht element binnen traditionele validiteitsraamwerken, zoals Messick's overkoepelende benadering en Kane's argument-gebaseerde aanpak. Daarom kiezen we ervoor om het sociale contract als apart kernelement van de voorgestelde denkwijze op te nemen.

Casussen van AI voor onderwijskundig meten

Om inzicht te geven in hoe je een discussie over de inzet van AI in het onderwijs vorm kan geven, bespreken we drie verschillende casussen en hoe daar vanuit de drie kernelementen gedacht kan worden over de waarde van AI.

Item constructie

Het ontwikkelen van toetsitems vergt veel tijd en expertise. AI-systemen zouden hier een ondersteunende rol kunnen spelen door het constructieproces aanzienlijk te versnellen. Uit eerder onderzoek is echter gebleken dat het proces van item constructie met behulp van AI nog vaak resulteert in items waarvan de kwaliteit te wensen overlaat (Kruis et al., 2024). Wie

items wil genereren met behulp van AI op een manier die waarde toevoegt, zal dus generatieve AI-modellen moeten inbedden in een proces met protocollen die kwaliteit waarborgen.

Onderwijskundige context

Hoe AI waarde kan toevoegen in het constructieproces zal niet alleen afhangen van het onderliggende onderwijskundige doel, maar ook de vorm en het belang van de toets. Bij toetsen die de reproductie van feitelijke kennis meten in het kader van kwalificatie zijn er andere mogelijkheden voor het gebruik van AI dan wanneer er ervaringsdoelen worden beoordeeld in het kader van persoonsvorming. Bij summatieve toetsen waar veel van afhangt, is de borging van kwaliteit essentieel en zou AI kunnen ondersteunen bij het uitvoeren van bestaande evaluaties van de kwaliteit van items. Bij formatieve toetsen kan het juist essentieel zijn dat het type opdracht aansluit bij de visie van de school en kan AI wellicht ondersteunen bij het creatieve proces van het maken van passende vragen. In beide gevallen blijft menselijke controle over de uiteindelijke itemselectie noodzakelijk.

Validiteit

Verschillende facetten van validiteit zullen op verschillende manieren beïnvloed worden door de inzet van AI bij constructie. Het is dus essentieel om te weten aan welke facetten in een bepaalde context prioriteit wordt gegeven. Door informatie te delen met een AI-systeem over de gewenste inhoud, taal en moeilijkheid van items, kunnen we sturen op een efficiënter itemconstructieproces waarbij beter rekening wordt gehouden met de belevingswereld en het niveau van leerlingen. Tegelijkertijd bestaan er twijfels over de kwaliteit van door AI gegenereerde items, maar ook over de inclusiviteit en daarmee de eerlijkheid van de taal van AI-systemen (Kruis et al., 2024). Het zal afhangen van de context welke van deze facetten het zwaarst wegen bij de keuze om AI in te zetten.

Sociaal contract

Kijkend naar een context zoals de eindexamens in het voortgezet onderwijs, hebben docenten momenteel een belangrijke rol in het constructieproces. Als we een AI-systeem zouden integreren in dit proces, verandert de positie van de docent, en daarmee ook de basis waarmee het vertrouwen van leerlingen en ouders in de kwaliteit van examens is opgebouwd. Juist omdat vertrouwen in deze context zwaar weegt, vraagt de inzet van AI om een zorgvuldige afweging, inclusief de vraag welke niveaus van automatisering wenselijk en onwenselijk zijn.

In Tekstbox 1 lichten we per kernelement een vraag uit die bij deze casus van belang was om te stellen en te behandelen. Dergelijke tekstboxen zijn ook terug te vinden aan het eind van de twee casussen die nog volgen.

Tekstbox 1.

Vragen om te stellen bij de casus Item constructie

Onderwijskundige context

Vanuit welke onderwijskundige doelen zou de inzet van het AI-systeem in deze context verantwoord kunnen worden?

Validiteit

Welke facetten van validiteit hebben prioriteit bij de afweging om het AI-systeem in te zetten in deze context en waarom?

Sociaal contract

Wat is de mogelijke invloed op het vertrouwen van leerlingen en ouders in de kwaliteit van toetsing, nu de rol van de docent verandert door de inzet van het AI-systeem in deze context?

2. Beoordelen van schriftwerk

Leerlingen kunnen tegenwoordig generatieve AI gebruiken voor hun schriftopdrachten, en docenten hebben vaak geen zicht op welk deel van het geschreven werk afkomstig is van de leerling zelf en welk deel van AI. SchrijfBlik is een prototype dat is ontworpen om leerlingen in het voortgezet onderwijs korte schriftopdrachten (bijv. een brief, betoog of essay) te laten schrijven met de hulp van een ingebouwde chatbot. Docenten kunnen, als de leerling de opdracht heeft ingeleverd, het schrijfproces inclusief de interactie met AI inzien (Poell et al., 2025).

Onderwijskundige context

Docenten geven aan vooral een rol te zien voor SchrijfBlik in formatieve toetsing, waarin het leerproces centraal staat. In de huidige situatie, waarin leerlingen AI gebruiken bij hun schriftwerk en docenten hier geen inzicht in hebben, is het immers moeilijk om gericht feedback te geven op het vlak van schrijfvaardigheid. Omdat summatieve toetsen vooral plaatsvinden in een gecontroleerde omgeving waarin AI niet gebruikt mag worden, zien docenten hierin nu niet een rol voor SchrijfBlik. Als docenten een idee willen krijgen van de AI-geletterdheid van hun leerlingen of de vaardigheid in het samenwerken met AI, kan SchrijfBlik ook bij summatieve toetsing een waardevolle bijdrage leveren.

Validiteit

SchrijfBlik zorgt met het creëren van transparantie over het schrijfproces voor beter zicht op het construct schrijfvaardigheid. Daarnaast biedt SchrijfBlik een authentieke toetscontext, omdat het gebruik van AI op een natuurlijke manier is geïntegreerd in de schrijfomgeving. Aan de andere kant kan het integreren van een chatbot in de omgeving ervoor zorgen dat docenten niet meer de pure schrijfvaardigheid van de leerling kunnen toetsen en dat leerlingen meer dan normaal gebruik maken van AI. Verschillen tussen leerlingen in AI-geletterd-



Verschillen tussen leerlingen in AI-geletterdheid kunnen daarnaast ongewenste effecten hebben op het maken van opdrachten.

heid kunnen daarnaast ongewenste effecten hebben op het maken van opdrachten. Een docent die voornamelijk gericht is op de facetten van transparantie en authenticiteit zal dus eerder de keuze maken om SchrijfBlik te gebruiken dan een docent die puur geïnteresseerd is in een valide meting van het construct schrijfvaardigheid.

Sociaal contract

De docent krijgt binnen SchrijfBlik een reëler beeld van de leerling en kan hiermee dus mogelijk een beter gesprek voeren over waar de leerling staat. Aan de andere kant, kunnen leerlingen het gevoel hebben dat zij worden gecontroleerd. Het gevoel kan bestaan dat de docent de leerling niet vertrouwd om het werk

op een eerlijke manier te maken en zo kan de ervaren privacy en autonomie van de leerling worden aangetast. De balans tussen transparantie, vertrouwen, privacy en autonomie hangt daarbij af van de doelen van de toets en de waarden die horen bij de context.

Tekstbox 2.

Vragen om te stellen bij de casus Beoordelen van schriftwerk

Onderwijskundige context

Als het AI-systeem niet wordt ingezet, welke onderwijskundige doelen zouden hieronder lijden?

Validiteit

Hoe worden de geprioriteerde facetten van validiteit beïnvloed door de inzet van het AI-systeem in deze context?

Sociaal contract

Wat is de mogelijke invloed op de privacy en autonomie van leerlingen bij de inzet van het AI-systeem in deze context?

3. Ondersteuning bij het nakijken

Het nakijken van het werk van leerlingen is een foutgevoelig proces, waarbij AI docenten mogelijk kan helpen om consistent te beoordelen. CheckMate² is een prototype in ontwikkeling door CitoLab en het Nationaal Onderwijslab AI (NOLAI) dat gebruikmaakt van natuurlijke taal modellen om de beoordeling van open antwoorden te ondersteunen. Daarbij worden antwoorden die gelijk zijn of vergelijkbare woorden bevatten geclusterd. Hierdoor wordt het voor de docent makkelijker om consistent te scoren, kan de docent de beoordeling kalibreren met collega's en

ontstaan er hopelijk minder onbedoelde inconsistenties.

Onderwijskundige context

Doordat CheckMate antwoorden clustert en labelt kan het ondersteunen in een consistente beoordeling door docenten. Ook kan het voor docenten tijdswinst opleveren. De grootste meerwaarde is er bij summatieve toetsen die door veel leerlingen gemaakt worden, zoals de eindexamens. Bij formatieve toetsing in de klas ziet een docent door de gedeeltelijke automatisering misschien juist individuele verschillen tussen leerlingen en leerbehoeften over het hoofd. Het is bij deze casus dus van belang om het niveau van automatisering van het AI-systeem goed af te stemmen op de onderliggende onderwijskundige behoeftes.

Validiteit

CheckMate kan bijdragen aan consistentere en eerlijkere beoordelingen. CheckMate kan aan de ene kant zorgen voor een eerlijke beoordeling door persoonlijke bias van docenten te verminderen. Aan de andere kant kan het ook bias introduceren wanneer het te strikte definities hanteert over wat het goede antwoord is. Dit zou negatief kunnen uitwerken voor leerlingen die antwoorden geven die afwijken van de norm. CheckMate heeft in dit licht dus het meeste meerwaarde in een context waar het terugdringen van persoonlijke bias van docenten veel prioriteit heeft, en het elimineren van algoritmische bias relatief weinig prioriteit.

Sociaal contract

Leerlingen doen moeite om een opdracht te maken en zij zullen over het algemeen verwachten dat docenten ook moeite doen om opdrachten te beoordelen. Wanneer een AI-systeem een deel van de taak van de docent overneemt, kunnen leerlingen op de lange termijn motivatie verliezen of zelfs hun ver-

²<https://cito.nl/onderzoek-innovatie/portfolio/innovatieve-prototypes/de-waarde-van-data/check-mate/>

trouwen in de docent. Docenten zouden op de lange termijn te veel kunnen gaan leunen op het oordeel van het AI-systeem, waardoor ze te weinig ruimte laten voor hun eigen professionele oordeel. In het kader van het sociale contract is het bij deze casus dus van belang om niet alleen na te denken over mogelijke effecten op de korte termijn, maar ook over langetermijneffecten.

Tekstbox 3.

Vragen om te stellen bij de casus *Ondersteuning bij het nakijken*

Onderwijskundige context

Wat is het beoogde niveau van automatisering bij de inzet van het AI-systeem in deze context? Is er nagedacht over het niet inzetten van het AI-systeem?

Validiteit

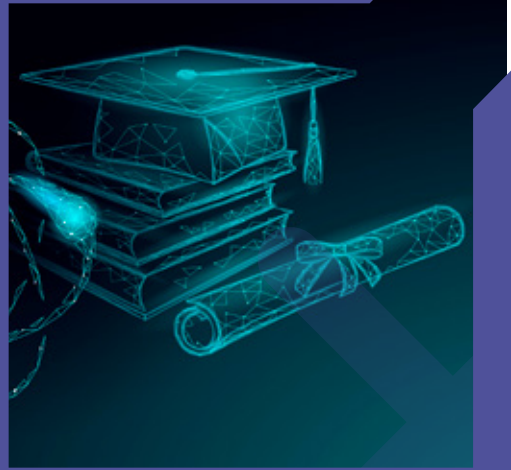
Welke facetten van validiteit hebben minder prioriteit bij de afweging om het AI-systeem in te zetten in deze context en waarom?

Sociaal contract

Wat zijn mogelijke langetermijneffecten voor leerlingen en docenten van de inzet van het AI-systeem in deze context?

Conclusie: Stel vragen en zoek samen naar het antwoord

We hopen dat de kernelementen, casussen en voorbeeldvragen hebben geholpen om onze manier van denken over de waarde van AI voor onderwijskundig meten te verduidelijken. We sporen de lezer aan om in de eigen context vragen te stellen over of en hoe AI-systemen waarde toevoegen. Onze voorbeeldvragen kunnen daarbij een leidraad bieden, maar zijn



Samen kunnen we zorgdragen voor verantwoorde inzet van AI die bijdraagt aan valide toetsing en waardevol onderwijs.

zeker geen uitputtende opsomming. We willen ook het belang aanstippen van het betrekken van alle relevante partijen bij het zoeken naar antwoorden, en dan voornamelijk de leerlingen en docenten die de invloed van AI-interventies het meest direct ervaren. Zo kunnen we samen zorgdragen voor verantwoorde inzet van AI die bijdraagt aan valide toetsing en waardevol onderwijs. ■

Referenties

- Aloisi, C. (2023). The future of standardised assessment: Validity and trust in algorithms for assessment and scoring. *European Journal of Education* 58, 98–110.

- <https://doi.org/10.1111/ejed.12542>
- Biesta, G. J. J. (2015). *Good Education in an Age of Measurement: Ethics, Politics, Democracy* (New York, NY, USA: Routledge)
 - Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education* 32, 347–364. <https://doi.org/10.1007/BF00138871369>
 - Delandshere, G. (2002). Assessment as Inquiry. *Teachers College Record*
 - Earl, L. M. (2003). *Assessment As Learning: Using Classroom Assessment to Maximize Student Learning* (Los Angeles, CA, USA: SAGE Publications)
 - Europese Unie (2024). *AI-verordening: Verordening 2024/1689 van het Europees Parlement en de Raad*. Publicatiedblad van de Europese Unie. <https://eur-lex.europa.eu/legal-content/NL/TXT/?uri=CELEX:32024R1689>
 - van Haastrecht, M., Haas, M., Brinkhuis, M., & Spruit, M. (2024). Understanding validity criteria in technology-enhanced learning: a systematic literature review. *Computers & Education*. 220:105128. <https://doi.org/10.1016/j.compedu.2024.105128>
 - van Haastrecht, M., De Groot, L., Buytenhuijs, F., Jongbloed-Pereboom, M., & Kruis, J. (2025). Artificial Intelligence for Educational Measurement: Where is the Value for Education? *Frontiers in Education* 10:1677255. <https://doi.org/10.3389/feduc.2025.1677255>
 - Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement* 50, 1–73. <https://doi.org/10.1111/jedm.12000384>
 - Kruis, J., Pera, M. S., Napel, Z. t., Landoni, M., Murgia, E., Huibers, T., & Feskens, R. (2024). Toward Personalised Learning Experiences: Beyond Prompt Engineering. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference* (New York, NY, USA: Association for Computing Machinery), 644–649. <https://doi.org/10.1145/3628516.3659367>
 - Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher* 18, 5–11. <https://doi.org/10.3102/0013189X018002005>
 - Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *European Journal of Education* 57, 632–645. <https://doi.org/10.1111/ejed.12527392>
 - 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 (Springer Nature Switzerland), 71–80. https://doi.org/10.1007/978-3-031-99261-2_6
 - UNESCO (2021). *Reimagining our futures together: a new social contract for education*. Tech. rep., 405 Educational and Cultural Organization of the United Nations, Paris, France
 - Williamson, B., Bayne, S., & Shay, S. (2020). The datafication of teaching in Higher Education: critical issues and perspectives. *Teaching in Higher Education* 25, 351–365. <https://doi.org/10.1080/13562517.2020.1748811>
 - Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice* 31, 2–13. <https://doi.org/10.1111/j.1745-3992.2011.00223.x>