CINOP
leren
met impact

ResearchNed
mrt 2022

Onderzoek naar de inhoudsvaliditeit van het toetsproces van twee maatschappijgerichte centrale examens

GESCHIEDENIS VMBO GL/TL 2021-1 EN ECONOMIE HAVO 2021-1

Onderzoek in opdracht van het College voor Toetsen en Examens

Janneke Lommertzen | Judith Janssen | Lieve Heijsters

Onderzoek naar de inhoudsvaliditeit van het toetsproces van twee maatschappijgerichte centrale examens
Janneke Lommertzen, Judith Janssen, Lieve Heijsters
Maart 2022

Inhoudsopgave

1 Achtergrond

1.1 Aanleiding

Het College voor Toetsen en Examens (CvTE) heeft regie over de examenketen in het voortgezet onderwijs. In de Wet College voor toetsen en examens staat vastgelegd dat Stichting Cito (Cito) en DUO wettelijke ketenpartners zijn. Stichting Cito is de ontwikkelaar van de opgaven en de examens en is verantwoordelijk voor het psychometrische onderzoek. DUO verzorgt de logistiek, doet het drukwerk, biedt de examensoftware Facet aan en is verantwoordelijk voor de distributie van de centrale examens. De scholen zijn verantwoordelijk voor de afnames en de beoordeling van de examens. Daarnaast wordt er nauw samengewerkt met organisaties als SLO (het nationaal expertisecentrum leerplanontwikkeling), de Inspectie van het Onderwijs, de Steunpunten taal en rekenen vo en mbo.[1]

Het College voor Toetsen en Examens (CvTE) bestaat uit een College en een bureau. Daarnaast zijn er veel docenten, leerkrachten en vakdeskundigen die in deeltijd meewerken in verschillende vaststellings- of syllabuscommissies. Het College zet de koers uit van het CvTE, en legt verantwoording af aan de minister en/of staatssecretaris van onderwijs. De medewerkers van het bureau van het CvTE zorgen ervoor dat de centrale examens en toetsen tot stand komen zoals bedoeld in de wet, onder meer door sturing te geven aan vaststellingscommissies. In de verschillende vakcommissies/ vaststellingscommissies worden opgaven die door de verschillende constructiegroepen, onder leiding van Cito zijn gemaakt, op kwaliteit beoordeeld en wordt uiteindelijk het examen of de toets in zijn geheel vastgesteld. Deze commissies zijn uiteindelijk ook betrokken bij de normering.[2,3] Daarnaast stelt het CvTE syllabuscommissies samen, die de omschrijving van de te toetsen leerstof verzorgen. De jaarlijkse update van de syllabi wordt door de vaststellings-commissies uitgevoerd. Bij de noodzaak van grote veranderingen wordt een nieuwe syllabuscommissie ingesteld.

Vanuit deze regierol heeft het CvTE aan ResearchNed gevraagd te onderzoeken of de procedures die het CvTE en Cito bij de borging van de validiteit van de maatschappijgerichte examens hanteren aan (inter)nationale kwaliteitsstandaarden voldoen. ResearchNed voert dit onderzoek uit als hoofdaannemer, in samenwerking met CINOP.

Mede naar aanleiding van de Motie Transitie in onderwijstoezicht, ingediend door Jadnanansing[4], laat CvTE de validiteit van centrale examens periodiek onderzoeken. Zo zijn de afgelopen jaren de volgende examens onderzocht door RCEC: de centrale examens Biologie vmbo GL/TL 2016, Duits havo 2016, Engels vwo 2017, Wiskunde vmbo GL/TL 2017 en twee beroepsgerichte examens, te weten: Produceren en Installeren en Energie (PIE) vmbo KB 2019 en Dienstverlening en Producten (D&P) vmbo GL 2019[5]

1 https://www.cvte.nl/over-het-cvte/samenwerking/ketenregie
2 https://www.cito.nl/-/media/files/over-cito/cito-informatie-docenten-rol-constructeur-examenopgaven.pdf
3 https://www.cvte.nl/over-het-cvte/wie-zijn-wij
4 Kamerstukken II 2013/14 33905-7
5 Zie o.a.: Brouwer, A., Sanders, P., & Veldkamp, B. (2019). Onderzoek naar de inhoudsvaliditeit van een tweetal beroepsgerichte centrale examens voortgezet onderwijs 2019. RCEC - Onderzoek in opdracht van het College voor Toetsen en Examens.

Er bestaan verschillende opvattingen over validiteit van toetsen[6,7]. In eerdere onderzoeken naar de Centrale Examens door RCEC wordt validiteit in meer beperkte zin opgevat als 'de mate waarin de inhoud van het examen overeenstemt met het doel van het examen'[5]. In recente literatuur over validiteit wordt benadrukt dat validiteit gaat over het gebruik en de interpretatie van toetsscores. Dit verschil kan worden geïllustreerd aan de hand van een voorbeeld over een toets Franse luistervaardigheid, waar de toetsscore iets zegt over de beheersing van het luisteren naar in het Frans gesproken teksten. En waar de score gebruikt wordt om te beslissen of een leerling kan overgaan naar een volgende klas.[8]

Met andere woorden: validiteit is niet een eigenschap van een toets of een examen, maar hangt nauw samen met waar de toetsscores voor gebruikt worden. Deze opvatting sluit aan bij de Amerikaanse standaarden, '*Standards for Educational and Psychological Testing*' die in 1999 door de American Educational Research Association (AERA), de American Psychological Association (APA) en de National Council on Measurement in Education (NCME) zijn opgesteld[9]. Deze standaarden zeggen het volgende over validiteit: '*Because validity depends on the particular uses and interpretations of test scores, the Standards emphasize the need to elaborate the intended uses and interpretations of test scores and to provide evidence relevant to the intended uses and interpretations*'[10]. In deze opvatting gaat het erom dat toetsontwikkelaars en hun opdrachtgevers de validiteit van de toetsscores voor het beoogde gebruik kunnen onderbouwen door voldoende bewijzen voor validiteit te verzamelen gedurende het toetsontwikkelproces[11]. In paragraaf 2.2 gaan we hier nader op in.

1.2 Onderzoeksvragen

Om een antwoord te kunnen geven op de hoofdvraag, zijn er deelvragen geformuleerd die we eerst zullen beantwoorden. De hoofdvraag is als volgt geformuleerd:

> 'Voldoen de procedures die het CvTE en Cito bij de borging van de inhoudsvaliditeit van de maatschappijgerichte examens hanteren aan (inter)nationale kwaliteitsstandaarden?'

Deze hoofdvraag zal worden beantwoord aan de hand van de volgende drie deelvragen:

1. Wat zijn de internationale kwaliteitsstandaarden voor examens?
2. Welke procedures hanteren het CvTE en Cito bij de borging van de inhoudsvaliditeit van de maatschappijgerichte examens?
3. Voldoen de procedures van het CvTE en Cito aan de internationale kwaliteitsstandaarden?

6 Sluijter, C., Hemker, B., & Eggen, Th. (2018b). Beoordelen van de kwaliteit van toetsen en examens, deel 2: de praktijk. Arnhem: Cito: www.cito.nl/kennis-en-innovatie/kennisplein.
7 Newton, P. E., Shaw, S. D., (2014). Validity and validation. In: P. E. Newton & S. D. Shaw (Eds.) Validity in Educational & Psychological Assessment (pp. 1-26). Sage Publications Ltd.
8 Wools, S., (2013). De validiteit van toetsscores. In: P. F. Sanders (Red.). Toetsen op School (pp. 69 – 83). Arnhem: Cito.
9 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). 1999 Standards for Educational and Psychological Testing, 2014 Edition, Washington, DC: American Educational Research Association
10 Linn, R. L. (2006). The Standards for Educational and Psychological Testing: Guidance in Test Development. In: S. M. Downing & T. M. Haladyna (Eds.). Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
11 Voor een overzicht van historische visies op validiteit, zie: Newton, P. E. & Shaw, S. D. (2014). Validity And Validation. In: Validity in Educational & Psychological Assessment (pp. 1-26). London: Sage Publications.

1.3 Aanpak

Als eerste stap in het onderzoek is er een literatuuronderzoek uitgevoerd om de internationale en nationale kaders voor validiteit in kaart te brengen. Op basis hiervan is een onderzoekskader opgesteld in de vorm van een lijst met criteria waaraan het proces van construeren en vaststellen van toetsen, en de borging van de kwaliteit van toetsen moeten voldoen. Deze lijst wordt beschreven in hoofdstuk 4 en zal in de daaropvolgende hoofdstukken als checklist worden gebruikt om na te gaan of de twee centrale examens voor maatschappijgerichte vakken die in dit onderzoek worden onderzocht aan de gestelde criteria voldoen. Bij het opstellen van het onderzoekskader vormden bestaande nationale en internationale kwaliteitsstandaarden het uitgangspunt. In dit onderzoek kiezen we er bewust voor om verschillende standaarden in elkaar te integreren. Hoewel de standaarden soms in opzet verschillen, zijn er inhoudelijk veel overeenkomsten.[12,13]

De tweede stap bestaat uit het inventariseren en verzamelen van relevante documenten over de examens. Als leidraad bij deze inventarisatiefase gebruiken we daarbij het schema in figuur 3.2. Per fase zijn de bijbehorende documenten opgevraagd.

Als derde stap, tot slot, worden deze bronnen geanalyseerd aan de hand van het bij stap 1 geschetste kader. De uitkomsten zijn per examen systematisch beschreven in hoofdstuk 5 en 6. Hierbij is rekening gehouden met de vergelijkbaarheid met eerdere onderzoeken die zijn uitgevoerd door RCEC.
Indien tijdens het documentenonderzoek zaken naar boven komen waarbij aanvullende informatie gewenst is, zal het CvTE en/of Cito worden gevraagd om toelichting, c.q. aanvullende informatie aan te leveren.

Deze inventarisatie van de examens aan de hand van het internationale kader kan worden gezien als een formatieve evaluatie van de examens, waarbij we inventariseren in hoeverre de onderzochte examens met de internationale kaders in overeenstemming zijn. Resultaten hiervan, vooral daar waar er mogelijk discrepanties zijn tussen de werkwijzen van het CvTE en Cito en de internationale kaders kunnen door het CvTE en Cito als feedback worden gebruikt om eventueel processen te heroverwegen.

12 Sluijter, C., Hemker, B., & Eggen, Th. (2018a). Beoordelen van de kwaliteit van toetsen en examens, deel 1: Systemen en criteria (pp. 1-8). Arnhem: Cito
13 Wools, S. (2012). Towards a comprehensive Evaluation System for the quality of tests and assessments, in: Psychometrics in Practice at RCEC (pp. 95-106).

2 Internationale kwaliteitsstandaarden voor examens

Het is gebruikelijk om de kwaliteit van examens en toetsen te beschrijven met behulp van een viertal kwaliteitscriteria: validiteit, betrouwbaarheid, bruikbaarheid en transparantie.[14] De centrale eindexamens in het voortgezet onderwijs zijn zogenaamde *high-stakes* examens. Er hangt een zwaarwegende beslissing van af. Het centraal examen vormt een belangrijk onderdeel van het eindexamen[15]. Het vormt vijftig procent van het eindresultaat voor een kandidaat (de andere vijftig procent is het schoolexamen). Slaagt een kandidaat voor het eindexamen, dan heeft hij/zij aangetoond over voldoende kennis te beschikken om toegelaten te worden tot een passende vervolgopleiding in het mbo, hbo of wo. Behaalt de kandidaat het examen niet, dan is hij/zij niet (direct) toelaatbaar. Bij *high-stakes* examens is het bij uitstek van belang dat het meetinstrument (= het examen) voldoende betrouwbaar meet, dat de toetsscores uit het meetinstrument valide zijn en dat de examens voor de kandidaten bruikbaar en transparant zijn. Vooral validiteit is belangrijk. Zoals Linn (2006) stelt: "*The higher the stakes associated with test scores, the greater the concern for validity*". Bij minder zwaarwegende toetsen zoals proefwerken en formatieve toetsen om de voortgang in beeld te krijgen kan bij ontwikkeling en samenstelling in de regel aan minder strenge eisen worden voldaan, omdat er van de scores op dergelijke toetsen geen zwaarwegende beslissingen afhangen.[16,17,18]

2.1 Criteria van kwaliteit

We lichten de twee belangrijkste kwaliteitscriteria voor *high-stakes* examens (validiteit en betrouwbaarheid) hier nader toe. Een examen kan gezien worden als een meetinstrument waarmee bepaalde vaardigheden en/of kennis bij een bepaalde doelgroep wordt gemeten (bijvoorbeeld de beheersing van vaardigheden en kennis over Economie door havo-5 leerlingen).

Onder betrouwbaarheid wordt verstaan de mate waarin toetsscores consistent, nauwkeurig en reproduceerbaar zijn. Dit wil zeggen, vrij van meetfouten. Factoren die invloed hebben op de betrouwbaarheid zijn: de lengte van een examen (het aantal vragen), de zorgvuldigheid waarmee opgaven zijn geconstrueerd en de mate waarin de beoordeling vrij is van subjectiviteit. Als een examen te weinig vragen bevat, dan loopt de meting snel het gevaar onbetrouwbaar te worden. En als de beoordeling van de prestaties van een kandidaat op een examen door menselijke beoordelaars plaatsvindt, dan dient dit met zo min mogelijk storende subjectieve invloeden plaats te vinden.[19] Om de betrouwbaarheid van de beoordeling bij open vragen te verhogen wordt er bij *high-stakes* examens vaak een dubbele beoordeling door onafhankelijke beoordelaars uitgevoerd en wordt er veel aandacht besteed aan zorgvuldig opgestelde, eenduidig uit te leggen beoordelingsvoorschriften. De betrouwbaarheid van een toets kan statistisch worden geschat door een betrouwbaarheidscoëfficiënt te berekenen. De waarde van een betrouwbaarheids-coëfficiënt ligt tussen 0 en 1 en kan op verschillende manieren worde vastgesteld. De meest gebruikte methode gaat uit van de Alfa-coëfficiënt, ook wel Cronbach's alfa genoemd. Alfa is een maat voor de interne consistentie van een toets.[20] De berekening van deze alfa is onderdeel van de toets- en itemanalyses die Cito uitvoert.

14 In internationale literatuur worden de laatste twee criteria meestal onder het begrip fairness of test use samengenomen. Zie Linn (2006)
15 Het eindexamen bestaat uit het schoolexamen en het centraal eindexamen.
16 Linn, R. L. (2006). The Standards for Eduational and Psychological Testing: Guidance in Test Development. In: S. M. Downing & T. M Haladyna (Eds.) Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates, Inc. n
17 Maassen, N. et al. (2014). Kwaliteit van toetsen binnen handbereik. Een reviewstudie van onderzoek en onderzoeksresultaten naar de kwaliteit van toetsen, RCEC en Cito.
18 Sluijter, C., Hemker, B., & Eggen, Th. (2018a). Beoordelen van de kwaliteit van toetsen en examens, deel 1: Systemen en criteria. Arnhem: Cito
19 Sluijter, C., Hemker, B., & Eggen, Th. (2018b). Beoordelen van de kwaliteit van toetsen en examens, deel 2: de praktijk. Arnhem: Cito.
20 Cito, Toetstechnische begrippenlijst: geraadpleegd op 17 januari 2022 van: https://www2.cito.nl/static/oenw/ttb/beglist1.htm

Het tweede belangrijke kwaliteitscriterium voor examens is validiteit. Over validiteit bestaan verschillende opvattingen. In engere zin wordt vaak gezegd: een examen dient zo te zijn ontwikkeld en samengesteld dat de vragen precies datgene toetsen wat er getoetst moet worden. Maar een bredere opvatting van validiteit benadrukt dat de toetsscores een valide weergave moeten zijn van de beheersing van het getoetste domein. Om een voorbeeld te geven: als er in een examen wiskunde ook een aantal talige opgaven voorkomen, dan moeten deze opgaven zo zijn vormgegeven dat ze alléén de wiskundige kennis of vaardigheid meten en niet (per ongeluk) het leesbegrip van de examenkandidaten.[21] Als de talige opgaven per ongeluk wél het leesbegrip toetsen, dan is de toetsscore niet als valide te beschouwen voor het aantonen van wiskundige kennis. Hieronder gaan we hier nog nader op in.

2.2 Validiteit

Zoals gezegd, bestaan er verschillende opvattingen over validiteit[19]. De oorspronkelijke definitie van validiteit uit 1937 was "de mate waarin een test meet wat men ermee beoogt te meten"[22]. In deze opvatting wordt validiteit traditioneel afgebakend als 'inhoudsvaliditeit' en wordt er vooral belang gehecht aan de manier waarop het leerstofdomein is geoperationaliseerd in de opgaven in een examen. Hoe nauwkeurig is het leerstofdomein ofwel het 'construct' dat getoetst wordt in het examen omschreven? Hoe goed passen de opgaven daarbij?

Er wordt naast de manier waarop de inhoud van de toets is gedefinieerd, ook gelet op een tweetal subcriteria: in hoeverre vormen de opgaven een goede *representatie* van het leerstofdomein en in welke mate zijn de opgaven *relevant* voor het leerstofdomein? Als veel opgaven over slechts een deel van het leerstofdomein gaan en andere belangrijke delen van de leerstof niet aan bod komen, dan kan er sprake zijn van 'construct-onderrepresentatie' en is het examen niet (volledig) inhoudsvalide. Het kan ook voorkomen dat een examen relatief veel opgaven bevat die buiten het leerstofdomein vallen, dan kan er sprake zijn van 'construct-irrelevantie'.[23]

De "*Standards for Educational and Psychological Testing*", ook wel de 'Amerikaanse standaarden'[24] die in 1999 zijn ontwikkeld en die nog steeds breed erkend worden als de 'gouden standaard'[25] voor het ontwikkelen van goede toetsen en examens, benadrukken dat "*(…) validity depends on the particular uses and interpretations of test scores*". Met andere woorden: het criterium 'validiteit' hangt nauw samen met het doel en is dus afhankelijk van andere factoren. Het gaat niet alleen om de inhoud van het examen zelf, maar over de validiteit van de *interpretatie* voor een bepaald *gebruik* van *toetsscores*. Om validiteit aan te tonen is het volgens deze opvatting nodig om in ieder geval heel nauwkeurig te bepalen welk inhoudelijk construct (bijvoorbeeld wiskunde of leesvaardigheid) er getoetst wordt bij welke doelgroep. Daarnaast dient er vooral ook nauwkeurig omschreven te worden welke *betekenis* een score op dit examen heeft voor de gekozen doelgroep en welke *beslissingen* er op grond van de toetsscores genomen worden.

21 Linn, R. L. (2006). The Standards for Eduational and Psychological Testing: Guidance in Test Development. In: S. M. Downing, & T. M Haladyna (Eds.) Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
22 Oorspronkelijk omschreven door Garret, 1937, p.324, geciteerd in Angoff, 1988. In: Sluijter, C., Hemker, B., & Eggen, Th. 2018b, p. 3.
23 Kane, M., (2006). Content-Related Validity Evidence in Test Development. In: S. M. Downing, & T. M. Haladyna (Eds.), Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
24 American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). 1999 Standards for Educational and Psychological Testing, 2014 Edition, Washington, DC: American Educational Research Association. Geraadpleegd op 5 oktober 2021
25 Sluijter, C., Hemkers, B., Eggen, Th. (2018a), Beoordelen van de kwaliteit van toetsen en examens, deel 1: Systemen en criteria, Arnhem: Cito, p. 3

Het is noodzakelijk om goed onderscheid te maken tussen interpretatie en gebruik van toetsscores. Een score die een leerling behaalt op een bepaald examen kan men *interpreteren* als de mate waarin een kandidaat het getoetste beheerst. Die score kan men vervolgens *gebruiken* om te beslissen of hij/zij voor het examen gezakt of geslaagd is. Volgens de ruimere/modernere opvatting van validiteit dienen verschillende bewijzen te worden gezocht om de *validiteit van de toetsscores* voor een beoogde interpretatie en gebruik aan te tonen. Binnen deze 'argumentgerichte benadering' bepalen de beoogde interpretatie en het beoogde gebruik van toetsscores welke uiteenlopende criteria in welke mate relevant zijn.[26]

Aantonen van validiteit

Er zijn verschillende soorten bewijzen die gebruikt kunnen worden om de validiteit van toetsscores aan te tonen. Voor de centrale examens in het voortgezet onderwijs, waar het gebruiksdoel en de doelgroep nauwkeurig omschreven zijn, zijn vooral *inhoudsbewijzen* van belang. Wools (2013) schrijft hierover: 'Bij inhoudsbewijzen gaat het om de keuzes die gemaakt worden ten aanzien van de onderwerpen of onderdelen uit de leerstof die in de toets opgenomen worden. Deze keuzes bepalen voor een groot deel of een toets representatief is voor het leerstofdomein […] waarover uitspraken gedaan moeten worden'.[5] Dit komt overeen met de Amerikaanse standaarden voor de validiteit van toetsen in het Handbook of Test development[27]. Linn (2006) benadrukt dat validiteit een belangrijke kwaliteitseis is en dat examenontwikkelaars daarom liefst verschillende bewijzen dienen te verzamelen om de validiteit van de toetsscores van examens voor het beoogde gebruik aan te tonen. Het belangrijkste is hierbij een nauwkeurige omschrijving en afbakening van de inhoud van de toets. Deze kan worden beoordeeld door de toetsspecificaties en de toetsmatrijs te analyseren. "*A critical part of the evidence may be based on test content, and the test specifications provide the natural starting place for content-based validity evidence*"[27]. Daarbij moet er aandacht zijn voor de koppeling tussen de leerdoelen of eindtermen binnen het te toetsen leerstofdomein enerzijds en de opgaven in het examen anderzijds. Verder is het belangrijk dat er vakexperts[28] betrokken zijn bij het opstellen en beoordelen van de toetsspecificaties, de toetsmatrijs en bij het construeren van de opgaven. Een belangrijke stap voor het aantonen van validiteit is ook gelegen in de samenstellingsfase: '*Often considered a mundane task, the validity of the final test score interpretation very much relies on a competent and accurate test assembly process*.' Dit betekent onder meer dat het schrijven van de opgaven en het samenstellen van een examen gebeurt met inbreng van een of meerdere inhoudelijke vakexperts en dat er strenge kwalitatieve eisen worden gesteld aan het traject voor het ontwikkelen van opgaven en de samenstellingsfase. '*This is the essence of the content-related validity argument, which, to be taken seriously, must be indepedently verifiable by independent, noninvested content experts*' (Downing 2006, p. 13).[29]

Samenhang tussen validiteit en betrouwbaarheid

Tenslotte is het belangrijk hier te noemen dat de kwaliteitscriteria betrouwbaarheid en validiteit niet los van elkaar kunnen worden gezien omdat ze nauw met elkaar samenhangen. Een examen dat in de toets- en itemanalyse na de afname een lage betrouwbaarheid blijkt te hebben (d.w.z een Cronbach's alfa < 0,7) heeft kennelijk niet op een consistente wijze gemeten waardoor de toetsscores mogelijk niet valide zijn. Wools (2013) merkt op: 'Om er zeker van te zijn dat je met een toets de juiste beslissing neemt, moeten zowel de betrouwbaarheid als de validiteit van toetsscores aangetoond worden. Een meting kan immers

26 Wools, S. (2015). De Validiteit van Toetsscores. In: P.F. Sanders (Red.), Toetsen op School. (pp. 69-83). Arnhem: Cito.
27 Linn, R. L. (2006). The Standards for Eduational and Psychological Testing: Guidance in Test Development. In: S. M. Downing, & T. M Haladyna (Eds.) Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
28 Onder vakexperts worden voor het vak en niveau bevoegde docenten verstaan met ervaring in examenklassen en docenten in het betreffende vak die werkzaam zijn in het hoger onderwijs.
29 Downing, T.M. (2006) Twelve steps for Effective Test Development. In: S. M. Downing, & T. M. Haladyna (Eds.). Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates.

alleen valide zijn als deze ook betrouwbaar is'. Andersom kan een toets waarvan de validiteit van de toetsscores niet overtuigend aangetoond kon worden, wél betrouwbaar hebben gemeten. Ter illustratie: stel, men gebruikt een weegschaal die 5 kg te licht ingesteld staat en men wil weten hoe zwaar men is. De schaal die meet is betrouwbaar (geijkt) en zal met elke meting hetzelfde gewicht aangeven. Maar de meting is niet valide want de weegschaal geeft niet het 'echte' lichaamsgewicht aan, maar 5 kg te weinig.[30]

2.3 Wat betekent bovenstaande voor ons onderzoekskader?

We sluiten aan bij de opvatting dat de validiteit van de toetsscores aangetoond dient te worden door inhoudsbewijzen te verzamelen in verschillende fasen van het toetsproces en door na te gaan of de betrouwbaarheid van een examen zoals gerapporteerd in de analysefase na de afname voldoende is en geen aanleiding geeft om te twijfelen aan de validiteit van de toetsscores.
De interpretatie van en het gebruik van de toetsscores voor de centrale examens zijn heel duidelijk omschreven en zijn duidelijk vastgelegd: de toetsscore wordt geïnterpreteerd als een maat van beheersing van leerstof en het zakken of slagen wordt gebruikt om leerlingen te kwalificeren met als recht om wel of niet mogen doorstromen naar een aansluitende vervolgopleiding[31]. Dit betekent dat door ons nu vooral nagegaan moet worden hoe er in de eerste drie fasen van het toetsontwikkelproces (zie figuur 3.1) is gezorgd voor een goede definitie van de toetsspecificaties en van de inhoud van het examen in relatie tot het te toetsen leerstofdomein. Daarbij zal moeten worden nagegaan hoe de inhoud van het examen 'scoort' op de subcriteria: representatie en relevantie van de inhoud in relatie tot het te toetsen domein.
De Amerikaanse standaarden wijzen er nadrukkelijk op dat examenmakers een systematische en zorgvuldige toetsconstructieprocedure dienen te volgen. Dat is noodzakelijk om de bewijzen voor validiteit en de bewijzen voor betrouwbaarheid op een transparante en systematische wijze te verzamelen en te beargumenteren[32]. Downing (2006) heeft een veel toegepast, praktisch stappenplan voor toetsconstructie ontwikkeld: 'Twelve steps for Effective Test Development'[33]. Door deze twaalf stappen bij de ontwikkeling van een examen te volgen is het mogelijk om per stap in het toetsconstructieproces een kwaliteitscontrole uit te voeren en tevens te plannen hoe de bewijsvoering voor validiteit wordt opgesteld.

Zo is bijvoorbeeld een van de stappen (Downing 2006) dat er na een afname van een examen statistisch onderzoek wordt uitgevoerd over alle antwoorden van de leerlingen die het examen hebben gemaakt, een zogenaamde toets- en itemanalyse. Uit deze toets- en itemanalyse na de afnames van een examen kan worden afgeleid of een examen voldoende betrouwbaar is. De gewenste betrouwbaarheid (Cronbach's alfa) voor '*high stakes*' examens wordt meestal op 0,80 gesteld[34]. De fase van toets- en itemanalyse is een belangrijke controle op de kwaliteit van de opgaven in het examen. Uit de toets- en itemanalyse kan blijken dat bepaalde opgaven te moeilijk of te gemakkelijk waren voor de populatie. Of dat bepaalde opgaven juist slecht gemaakt zijn door leerlingen die op de overige opgaven juist heel goed scoren. Dit kan een aanwijzing zijn dat een dergelijke opgave inhoudelijk niet goed in elkaar steekt. Het is gebruikelijk bij de centrale eindexamens dat de toets- en item analyse plaatsvindt vóórdat aan de hand van de norm en de behaalde toetsscores de echte uitslag van het examen wordt vastgesteld: deze kwaliteitscontrole maakt het mogelijk dat een slecht functionerende opgave eventueel kan worden 'uitgeschakeld' zodat deze niet meetelt voor de uitslag.

30 https://web.cortland.edu/andersmd/STATS/valid.html
31 Sanders, P., Brouwer, A.J., Veldkamp, B.P. (2017). Onderzoek naar de inhoudsvaliditeit van de centrale examens en de afhandeling van onvolkomenheden bij de centrale examens. Enschede: RCEC
32 Linn, R. L. (2006). The Standards for Eduational and Psychological Testing: Guidance in Test Development. In: S. M. Downing, & T. M Haladyna (Eds.) Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
33 Downing, T.M. (2006) Twelve steps for Effective Test Development. In: S. M. Downing, & T. M. Haladyna (Eds.). Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates.
34 Sluijter, C., Hemker, B., & Eggen, Th. (2018). Beoordelen van de kwaliteit van toetsen en examens, deel 2: de praktijk. Arnhem: Cito.

Mede op basis van dit internationaal bekende stappenplan van Downing (2006) introduceerde Veldkamp (2016) een tienstappenplan voor de toetsconstructie van de centrale examens zoals die in Nederland worden afgenomen. Er is grote overlap tussen beide stappenplannen. De tien stappen zoals Veldkamp (2016) die benoemde en zoals deze in voorgaande onderzoeken van RCEC zijn gebruikt, zijn als volgt[35]:

1. Maak een toet plan
2. Definieer de inhoud
3. Stel de toetsmatrijs op
4. Ontwikkel de opgaven (items)
5. Stel het examen samen
6. Neem het examen af
7. Scoor de antwoorden
8. Bepaal de zak/slaaggrens
9. Rapporteer de scores
10. Documenteer de voorgaande stappen

We hebben beide stappenplannen vergeleken. Downing (2006) komt tot twaalf stappen omdat hij naast de genoemde tien stappen stap 6. '*Testproduction*' en stap 11 '*Itembanking*' als separate extra stappen onderscheidt. Terwijl Veldkamp (2016) deze twee activiteiten opneemt als onderdeel van stap 5 '*Stel het examen samen*'. Omwille van de vergelijkbaarheid met eerdere onderzoeken door RCEC[36] sluiten wij in dit onderzoek aan bij het tien-stappenplan van Veldkamp. We gaan na in hoeverre de examenmakers van de door ons geanalyseerde centrale examens een dergelijk stappenplan hanteren en hoe zij binnen de fases van de inhoudelijke afbakening en toets-samenstelling (stappen 1, 2, 3, 4 en 5) voor een kwalitatief goede inhoudsafbakening hebben gezorgd. Hebben de toetsconstructeurs de beschikking over een toetsplan, is er een toetsspecificatie in de vorm van een toetsmatrijs? Komen de constructeurs tot goede, consistente opgaven doordat bijvoorbeeld de vakexperts uit het voortgezet onderwijs die de opgaven schrijven, hierop zijn getraind? Wordt er een kwaliteitscontrole uitgeoefend op de opgaven zelf, bijvoorbeeld via een proces van screening en vaststelling van concept-opgaven waarbij onafhankelijke vakdeskundigen betrokken zijn?

35 Veldkamp, B.P. (2016). De inhoud en constructie van toetsen. In Sanders, P.F. (Red.), Toetsen op School: Hoger onderwijs (pp. 21-30). Arnhem: Cito.

36 Zie: Sanders, P., Brouwer, A.J., Veldkamp, B.P. (2018). Onderzoek naar de inhoudsvaliditeit van een tweetal centrale examens voortgezet onderwijs 2017. Enschede: RCEC. En: Sanders, P., Brouwer, A.J., Veldkamp, B.P. (2017). Onderzoek naar de inhoudsvaliditeit van de centrale examens en de afhandeling van onvolkomenheden bij de centrale examens. Enschede: RCEC

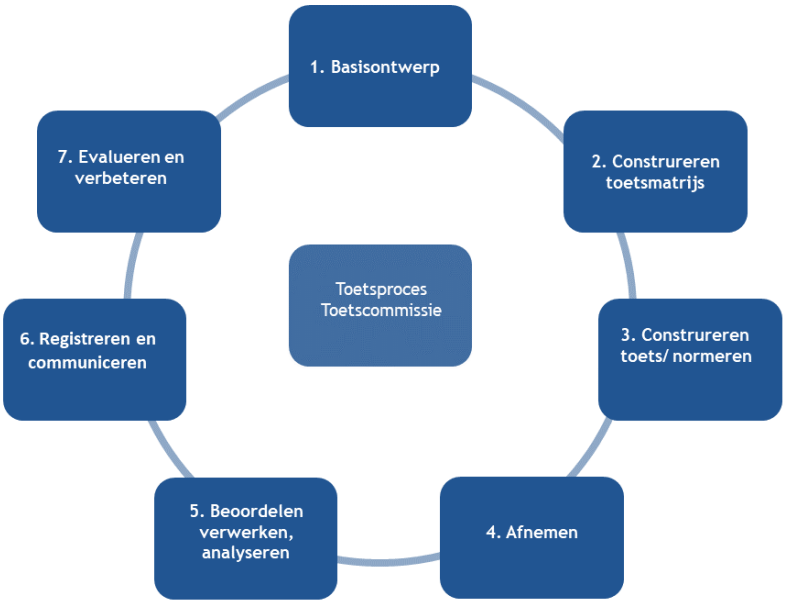
3 Toetsproces

3.1 Toetscyclus

De twaalf stappen van Downing (2006) en de tien stappen van Veldkamp (2016) beschrijven beide de verschillende fasen in het toetsproces. De Europese Association for Educational Assessment, AEA Europe heeft een beknopt raamwerk voor toetsstandaarden opgesteld, dat onder andere is gebaseerd op de Amerikaanse standaarden en het Nederlandse systeem. Ook dit Europese raamwerk gaat uit van de toetscyslus[37].

Het is gebruikelijk om deze fasen grafisch weer te geven in een cirkel omdat het ontwikkelen van examens een sterk cyclisch karakter heeft (figuur 3.1). De cyclus begint met de beslissingen omtrent het ontwerp en inhoud van het examen (doel, functie en inhoud, opgaven) en gaat via de afname, de beoordeling, scores verwerken en analyseren en het registreren van resultaten (scores) naar het einde van de cyclus toe, waar kritisch wordt teruggeblikt op alle voorgaande fasen. Hierna kunnen eventuele noodzakelijke verbeteringen worden doorgevoerd in een nieuwe cyclus.

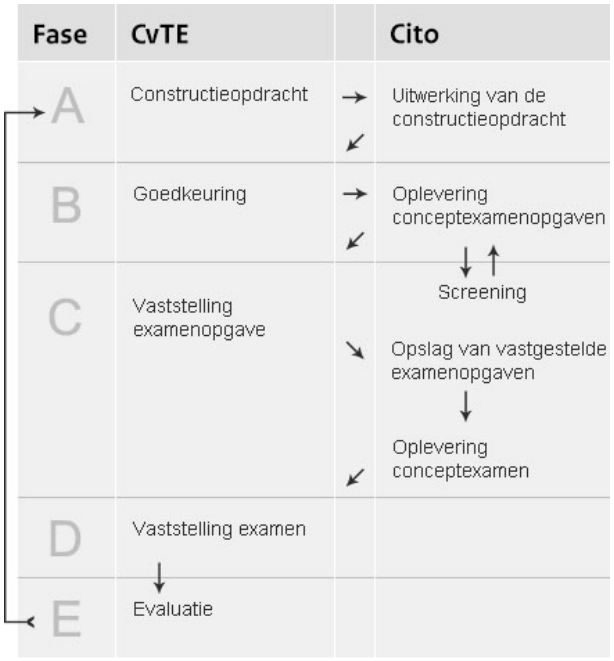
De structuur van het Europese raamwerk beschrijft elke fase van de toetscyclus in drie niveaus:
1) standaardeisen of criteria waar aan moet worden voldaan; 2) methoden om aan deze eisen te voldoen; en 3) mogelijke bewijzen die kunnen worden gebruikt om te contoleren of er aan de standaardeisen wordt voldaan. De toetscyclus is een leidraad die zichtbaar maakt welke producten er per fase (zouden moeten) worden opgeleverd en met welke hulpmiddelen kwaliteit zichtbaar kan worden gemaakt. Wij zullen de toetscyclus volgen om grafisch te weer te geven hoe de toetscyclus van het centraal examen verloopt.



Figuur 3.1: Toetscyclus[38]

37 AEA-Europe European Framework of Standards for Educational Assessment 1.0
38 Overgenomen uit: Expertgroep BKE/SKE, geciteerd in Sanders 2016, p. 76, maar ook in het Framework_of_European_Standards worden deze zeven stappen gehanteerd.

Het CvTE en Cito hanteren een eigen examencyclus. In deze paragraaf worden alle stappen beschreven die zij doorlopen bij de totstandkoming van centrale examens. De totstandkoming van een centraal examen behelst zo'n twee jaar, onderverdeeld in een cyclus van zes herhalende stappen (A t/m F, zie ook figuur 3.2).[39] Deze examencyclus beschrijft vooral de constructie-en samenstellingsfase. Daarom volgen we om de (inhouds)validiteit van de centrale examens te onderzoeken de toetscyclus.



Figuur 3.2: In schema: samenwerking tussen het CvTE en Cito voor constructie en vaststelling van centrale examens vo. Bron: https://www.examenblad.nl/onderwerp/examencyclus.

Examenprogramma

Allereerst worden de examenprogramma's per vak en schooltype door de minister van OCW vastgesteld. Deze examenprogramma's bevatten op hoofdlijnen wat de leerlingen moeten kennen en kunnen.[39]

Syllabus

Het CvTE specificeert vervolgens de eisen voor het centraal examen uit het examenprogramma. Hiervoor geeft het CvTE syllabi uit die zijn samengesteld door een commissie van vakinhoudelijke experts (onder andere vakdocenten met ervaring in eindexamenklassen) van het betreffende vak waarvoor de syllabus geschreven wordt. De syllabus biedt houvast aan examenconstructeurs, docenten en ontwikkelaars van leermiddelen. De syllabus kan naast de beschrijving van de exameneisen ook verdere informatie over het centraal examen bevatten, zoals bijvoorbeeld nadere specificaties van de examenstof, begrippenlijsten, bekend veronderstelde onderdelen van domeinen die verplicht zijn op het schoolexamen, bekend veronderstelde voorkennis uit de onderbouw, bijzondere vormen van examinering, voorbeeldopgaven, toelichting op de vraagstelling, hulpmiddelen. Nieuwe syllabi worden ontwikkeld door een syllabuscommissie die door het CvTE is ingesteld. Een syllabuscommissie bestaat uit een voorzitter, werkzaam in het naast liggende vervolgonderwijs, docenten uit het voortgezet onderwijs en adviseurs van Cito (toetsdeskundigen) en SLO (curriculumspecialisten).[39,40]

39 https://www.examenblad.nl/onderwerp/examencyclus
40 Actuele syllabi voor economie en geschiedenis zijn te vinden op: examenblad.nl

Constructieopdracht

De syllabus is de basis voor de constructieopdracht die het CvTE aan Cito geeft. Hierin zijn zaken vastgelegd als de verhouding tussen open vragen en gesloten vragen, globaal worden er hoofdthema's en vaardigheden verdeeld en er staat aangegeven hoeveel scorepunten het examen moet tellen. Het bevat de toetsmatrijs (de verdeling van de vragen over de leerdoelen), de mate waarin het examenvragen met een reproductief en/of productief karakter moet bevatten, de keuze van teksten en contextmateriaal, de toe te passen vraagvormen en vaardigheidsvragen en de wijze waarop de correctievoorschriften opgesteld dienen te zijn. Op basis van de constructieopdracht inclusief de toetsmatrijs maakt Cito de examenopgaven die in diverse ronden door vakexperts gescreend en beoordeeld worden. Na vaststelling van de examenopgaven door de vaststellingscommissie van het CvTE, doet Cito een voorstel voor samenstelling van het examen op basis van de toetsmatrijs. Om de vaststelling te vergemakkelijken, levert Cito daartoe een toetsrationale aan. In de toetsrationale staat aangegeven hoe het examen voldoet aan wat in de toetsmatrijs staat.[39]

Vaststelling

Vervolgens beoordeelt de vaststellingscommissie van het CvTE of dit examen geschikt en passend is als centraal examen en stelt vervolgens het examen en de correctievoorschriften vast onder verantwoordelijkheid van het college.

Beoordeling, verwerking scores en analyses na de afname

Na afnames en na de beoordelingsrondes (door de examinator en gecommitteerde) van de centrale examens, zijn de scores intern bekend. Ze kunnen nu nog niet aan de kandidaten worden meegedeeld: eerst vindt de analyse van de examenantwoorden plaats via statistisch onderzoek. De zogenaamde toets- en itemanalyses. De uitkomsten van de toets-en itemanalyse worden goed bekeken door de vaststellingscommissie, evenals de feedback van docenten op het examen die via o.m. de Quickscan is verzameld. Dan vindt de normeringsvergadering plaats waarbij de vaststellingscommissie een advies kan geven aan het normeringsoverleg van het CvTE over het bepalen van de definitieve uitslagen. De beslissingen van het normeringsoverleg worden gedocumenteerd in een verslagformulier. Met behulp van een procedure rondom normbepaling en normhandhaving, zoals gepubliceerd in de Staatscourant d.d. 04-02-2016[41] worden de toetsscores van kandidaten omgezet naar cijfers en wordt de zak/slaaggrens bepaald.[39]

Aan het einde van de cyclus, nadat de examenperiode is afgelopen worden de centrale examens geëvalueerd, samen met vertegenwoordigers van scholen. De Quickscan per vak wordt, evenals de toets- en itemanalyse, openbaar gemaakt aan het eind van de examencyclus en gepubliceerd door Cito.

In tabel 3.1 staat de relatie tussen de toetscyclus en de examencyclus van Cito en CvTE weergegeven.

41 Regeling omzetting scores in cijfers centrale examens en rekentoets VO 2016. In: Stcrt. nr. 4817, 4 februari 2016. Geraadpleegd op 1 november van: http://www.stilus.nl/examen/n-term-2016.pdf

Tabel 3.1: Relatie tussen Toetscyclus en Examencyclus Cito en het CvTE.

	Toetscyclus	**Examencyclus Cito/CvTE**
1.	Basisontwerp	A: Constructieopdracht
2.	Constructie toetsmatrijs	A & B: uitwerking constructieopdracht
3.	Constructie opgaven, samenstelling van examen en normering	B & C: Screening concept examenopgaven, C: vaststelling examenopgaven, oplevering conceptexamen en D: vaststelling examen en evt. een standaardbepaling op een referentie examen.
4.	Afname	
5.	Beoordeling, verwerking analyse	E: Evaluatie
6.	Registratie en communicatie van resultaten	E: Evaluatie
7.	Evaluatie en verbeteringen	E: Evaluatie

Normering tijdens corona

Als gevolg van corona zijn bepaalde examencondities aangepast. In 2021 waren er twee herkansingen en een extra tijdvak waardoor leerlingen hun examens konden spreiden. Ook hadden leerlingen de mogelijkheid om één niet-kernvak niet mee te laten tellen. Dit had gevolgen voor de normering omdat de populatie niet vergelijkbaar was met voorgaande jaren.[42]

42 https://www.rijksoverheid.nl/actueel/nieuws/2021/02/12/eindexamens-gaan-door-met-drie-extra-maatregelen

4 Onderzoekskader inhoudsvaliditeit

Op basis van de bronnen beschreven in hoofdstukken 2 en 3, zijn we tot een lijst criteria gekomen, die samen het onderzoekskader vormen waarmee centrale examens kunnen worden beoordeeld op validiteit. Het kader dient als checklist om te inventariseren welke bewijzen het CvTE en Cito leveren voor de (inhouds)validiteit van de maatschappijgerichte examens. In het onderzoekskader zijn de kwaliteitsstandaarden ingedeeld binnen de fasen uit de eerder beschreven toetscyclus.

Om op basis van de Amerikaanse standaarden tot een checklist te komen die toepasbaar is op de centrale schriftelijke eindexamens in het Nederlandse vo, zijn alle standaarden bestudeerd en beoordeeld op relevantie. Om niks te missen is dit door twee onderzoekers gedaan, onafhankelijk van elkaar. Vervolgens zijn de relevante standaarden besproken in het projectteam en opgenomen in de checklist. Omwille van de vergelijkbaarheid met eerdere onderzoeken door RCEC zijn deze standaarden vervolgens vergeleken met de criteria die door RCEC zijn gehanteerd. Daar waar de criteria van RCEC en de Amerikaanse standaarden overlappen, is dit in de tabellen aangegeven.

Bij centrale examens is het vooral relevant hoe de examenontwikkelaars zorgdragen voor het goed definiëren van het domein, de domeinrepresentatie en -relevantie. Dit betekent dat we ons focussen op de eerste drie fasen van de toetscyclus en de wijze waarop examenontwikkelaars in deze fasen de bewijzen voor validiteit aannemelijk maken. Verder is het van belang dat een examen voldoende betrouwbaar is. De bewijzen die de daarvoor worden geleverd komen vooral in fase 5 van de toetscyclus aan bod.

4.1 Fase 1. Basisontwerp

Allereerst moet er een zogenaamd toetsplan worden opgesteld, met daarin in ieder geval beschreven: de doelgroep, het meetdoel en het gebruiksdoel. Het specificeren van de doelgroep kan onder andere van belang zijn bij het beoordelen van de inhoud van het examen zoals het taalgebruik en de gehanteerde normen of cesuren. Het meetdoel beschrijft wat de examenkandidaten geacht worden te beheersen. Dit kan de beheersing van een bepaald construct zijn of de beheersing van een exameneenheid van een examenprogramma. Door het formuleren van het meetdoel wordt duidelijk wat wel en wat niet tot het te meten domein wordt gerekend. Het gebruiksdoel bij centrale examens is diplomering. Afhankelijk van de toetsscores van een kandidaat wordt wel of niet een diploma verstrekt.

Volgens de Amerikaanse kwaliteitsstandaarden[43] is het daarnaast gebruikelijk om ook de interpretatie en het gebruik van toetsscores in het plan op te nemen. De examenmaker moet duidelijk aangeven hoe de scores geïnterpreteerd moeten worden door de contexten te beschrijven waarin de scores worden gebruikt. Voor elke beoogde interpretatie is nodig: een motivering, een samenvatting van het bewijs en de theorie met betrekking tot de beoogde interpretatie. Bij de centrale examens is dit eenvoudig, omdat er maar één interpretatie van de toetsscores is; er is een minimaal aantal scorepunten nodig voor een voldoende, waarmee wordt aangegeven dat de kandidaat de examenstof voldoende beheerst.

43 American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). 1999 Standards for Educational and Psychological Testing, 2014 Edition, Washington, DC: American Educational Research Association

Verder zouden toetsontwikkelaars conform de Amerikaanse standaarden (standaard 1.2[43]) een toetsplan moeten opstellen waarin de beoogde interpretatie van de toetsscores (i.c. beheersing van een bepaald omschreven leerstofdomein) wordt onderbouwd, gegeven het beoogde gebruik i.c. de beslissing zakken/slagen. In Nederland worden er verschillende begrippen gebruikt die gezien kunnen worden als onderdeel van het toetsplan. Bijvoorbeeld de *syllabus* en de *constructieopdracht* spelen bij het ontwikkelen van de centrale examens door Het CvTE en Cito een cruciale rol. In de *syllabus* specificeert het CvTE op basis van het examenprogramma de eisen voor het centraal examen. In de *constructieopdracht* die het CvTE opstelt voor Cito, is vastgesteld op welke wijze de examenstof uit de syllabus wordt getoetst in het centraal examen. De *constructieopdracht* geeft nadere toetsspecificaties, waarbij de link tussen de examenstof en het examen wordt vastgelegd. De toetsspecificaties gaan onder meer over de inhoud, vorm, de lengte van de toets, psychometrie, scoring en afname van het examen.

Tabel 4.1: Checklist Fase 1: Basisontwerp

	Checklist Fase 1	**Bron**
1.1	Er is een toetsplan.	Amerikaanse standaarden: st. 4.1
1.2	In het toetsplan is aangegeven wat de doelgroep van het examen is.	Amerikaanse standaarden: st. 4.1 RCEC: criterium 1.1
1.3	In het toetsplan is aangegeven wat het meetdoel het examen is oftewel hoe het domein gedefinieerd is.	Amerikaanse standaarden: st. 4.1 RCEC: criterium 1.2
1.4	In het toetsplan is aangegeven wat het gebruiksdoel van het examen is.	Amerikaanse standaarden: st. 4.1 RCEC: criterium 1.3
1.5	In het toetsplan is in duidelijke taal gespecificeerd in welke contexten toetsscores gebruikt moeten worden.	Amerikaanse standaarden: st. 4.1
1.6	In het toetsplan is in duidelijke taal gespecificeerd met welke processen de toets moet worden afgenomen en gescoord (bv. papieren of digitale toets, wijze van afname, wijze van beoordeling, hoe komt score tot stand?).	Amerikaanse standaarden: st. 4.2
1.7	In het toetsplan staat de theorie met betrekking tot de beoogde interpretatie beschreven.	Amerikaanse standaarden: st. 1.2

4.2 Fase 2. Construeren toetsmatrijs

Onderdeel van de toetsspecificaties is de toetsmatrijs. Een toetsmatrijs is een schematische weergave van de inhoud van het examen waarin voor de verschillende inhoudscategorieën is vastgelegd op welk niveau de toetsing ervan plaatsvindt. Het is een nuttig hulpmiddel om de beoogde leerdoelen uit te werken tot toetsbare leerdoelen. Voor examenmakers kan het prettig zijn als in de toetsmatrijs percentages staan in plaats van absolute aantallen. Dus als er sprake is van een relatieve toetsmatrijs.

Het is raadzaam om in de toetsmatrijs ook het vereiste beheersingsniveau van het betreffende meetdoel op te nemen. Dit kan worden verduidelijkt door het gebruik van werkwoorden, bijvoorbeeld: kennen, kunnen, toepassen of begrijpen. De taxonomieën van Bloom of Romizowsky zijn veelgebruikt[44].

Conform de Amerikaanse kwaliteitsstandaarden is het noodzakelijk dat vakinhoudelijke experts de toetsspecificaties beoordelen op hun geschiktheid voor het beoogde gebruik van de toetsscores en de transparantie naar examenkandidaten. Hiermee wordt de inhoudsvaliditeit geborgd. Het doel van de beoordeling, het proces waarmee de beoordeling wordt uitgevoerd en de resultaten van de beoordeling moeten worden gedocumenteerd. Evenals de kwalificaties, relevante ervaringen en demografische kenmerken van de vakinhoudelijke experts.

44 Berkel, H. van, Bax, A. & Joosten-Ten Brinke, D. (2013). Toetsen in het hoger onderwijs, 3e herziene druk (p. 6). Houten: Bohn Stafleu van Loghum.

Tabel 4.2: Checklist Fase 2: Construeren toetsmatrijs

	Checklist Fase 2	**Bron**
2.1	Er zijn toetsspecificaties (bv. in de vorm van een constructieopdracht)	Amerikaanse standaarden: st. 4.1
2.2	In de toetsspecificaties staan het aantal vragen met bijbehorende scorepunten	Amerikaanse standaarden: st. 4.2 RCEC: criterium 3.1
2.3	In de toetsspecificaties staat de toetsvorm en/of het soort vragen (bijvoorbeeld gesloten en/of open)	Amerikaanse standaarden: st. 4.2 RCEC: criterium 3.1
2.4	In de toetsspecificaties staan de toegestane hulpmiddelen.	Amerikaanse standaarden: st. 4.2 RCEC: criterium 3.1
2.5	In de toetsspecificaties staat de voorgestelde lengte van de toets.	Amerikaanse standaarden: st. 4.2
2.6	In de toetsspecificaties staat de hoeveelheid tijd die is toegestaan voor de toets.	Amerikaanse standaarden: st. 4.2
2.7	In de toetsspecificaties staan de gewenste psychometrische eigenschappen van de items en de toets als geheel.	Amerikaanse standaarden: st. 4.2
2.8	In de toetsspecificaties wordt de volgorde van items gedefinieerd.	Amerikaanse standaarden: st. 4.2
2.9	Er is een toetsmatrijs.	Amerikaanse standaarden: st. 4.2 RCEC: criterium 3.1
2.10	In de toetsmatrijs wordt de relevantie van de inhoud van de toets of het examen voor het beoogde meetdoel aannemelijk gemaakt.	Amerikaanse standaarden: st. 4.2 RCEC: criterium 1.2
2.11	De toetsmatrijs is een adequate representatie van het meetdoel.	Amerikaanse standaarden: st. 4.2 RCEC: criterium 3.1
2.11a	De eind- en toetstermen representeren het meetdoel.	Amerikaanse standaarden: st. 4.2 RCEC: criterium 3.1
2.11b	De eind- en toetstermen sluiten aan op de inhoud en het vereiste beheersingsniveau van het betreffende meetdoel	Amerikaanse standaarden: st. 4.2 RCEC: criterium 3.1
2.11c	Het werkwoordgebruik in de eind- en toetstermen is eenduidig en sluit goed aan bij de gebruikte taxonomie (bijvoorbeeld van Bloom of Romiszowsky)	Amerikaanse standaarden: st. 4.2 RCEC: criterium 3.1
2.11d	Het aantal vragen geeft een voldoende dekking van het meetdoel.	Amerikaanse standaarden: st. 4.2 RCEC: criterium 3.1
2.12	Relevante vakinhoudelijke experts beoordelen de toetsspecificaties op hun geschiktheid.	Amerikaanse standaarden: st. 4.6
2.13	Het doel van deze beoordeling van toetsspecificaties, het proces waarmee deze beoordeling wordt uitgevoerd en de resultaten van deze beoordeling worden gedocumenteerd.	Amerikaanse standaarden: st. 4.6
2.14	De kwalificaties, relevante ervaringen en demografische kenmerken van relevante vakinhoudelijke experts worden gedocumenteerd.	Amerikaanse standaarden: st. 4.6

4.3 Fase 3. Construeren toets en normeren

Bij het construeren van het examen moeten eerst opgaven worden ontwikkeld. Vervolgens kan het examen worden samengesteld. Dan moet ook de norm (ofwel cesuur, zak/slaaggrens genoemd) van het examen worden bepaald. Daarnaast dienen regels opgesteld te worden om de toetsscores in cijfers of een waardering om te zetten. Het bepalen van de norm dient volgens de Amerikaanse standaarden te gebeuren met een methode voor standaardbepaling (ook wel ‘standardsetting’) van een examen. Een standaardbepaling dient volgens een vooraf bepaalde procedure te worden uitgevoerd door een onafhankelijk orgaan[45].

45 Cizek, J. K. (2006). Standard Setting. In S. M. Downing, & T. M. Haladyna (Eds.), Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bij vakken met een nieuw examenprogramma wordt bij Cito een standaardbepaling uitgevoerd door vakexperts (docenten uit het voortgezet onderwijs en/of uit het hoger onderwijs) onder begeleiding van een toetsdeskundige van Cito. Zij kunnen als experts vooraf de moeilijkheidsgraad van een examen inschatten ten opzichte van eerder afgenomen examens[46]. Tijdens de standaardbepaling wordt vastgesteld wat kandidaten zouden moeten beheersen en hoeveel punten er moeten worden behaald om te kunnen slagen. Bij deze vakken wordt docenten gevraagd wat een leerling zou moeten kennen en hoeveel punten een leerling zou moeten behalen voor een onvoldoende.[45,46] Een standaardbepaling is een methode om een norm te bepalen. Vervolgens zijn er ook methoden nodig om de norm te handhaven. Bij de centrale eindexamens gebeurt dit via de procedure met de N-term.[47]

In de (inter)nationale kwaliteitsstandaarden staan aandachtspunten voor het ontwikkelen en evalueren van vragen beschreven. Bewijzen die tonen dat hierop is gelet, tonen vooral de betrouwbaarheid van het examen aan. Daarentegen hangen criteria zoals “De stam bevat geen overbodige informatie (behalve als selectie van informatie het doel is)” en “Het taalniveau is afgestemd op de doelgroep” samen met de vraag of er getoetst wordt wat getoetst moet worden (domeinrepresentatie). Verder is het belangrijk dat kandidaten de vraag kunnen beantwoorden zoals bedoeld. Daarvoor is het essentieel dat de vragen voldoende gedetailleerd zijn en dat er indien nodig vooraf oefenvragen worden verstrekt.

De Amerikaanse standaarden schrijven voor dat de kwalificaties van de personen die de vragen ontwikkelen en op kwaliteit beoordelen gedocumenteerd moeten worden. Ook moeten de processen die worden gebruikt om hen te trainen en te begeleiden bij deze activiteiten worden vastgelegd.
Bij het bepalen of de vragen kwalitatief goed zijn geconstrueerd is het belangrijk dat er psychometrische analyses worden uitgevoerd (bijv. door middel van een pretest met analyses) en/of dat er gebruik wordt gemaakt van relevante experts om de opgaven en bijbehorende beoordelingscriteria op kwaliteit te beoordelen. In deze constructiefase is het belangrijk de geconstrueerde vragen op kwaliteit te screenen. Dit kan door de opgaven aan relevante onafhankelijke vakexperts voor te leggen ter beoordeling. Het kan ook (en de Amerikaanse standaarden adviseren dat) door het organiseren van een pretest bij leerlingen, zodat empirische gegevens over de opgaven bij de juiste doelgroep worden verzameld en geanalyseerd. De Amerikaanse standaarden benadrukken dat het betrekken van vakexperts in deze fase bijdraagt aan de validiteit van een examen. De standaarden beschrijven waar experts in het beoordelingsproces allemaal voor kunnen worden ingezet. Ze kunnen gevraagd worden om de score van de items te controleren en materiaal te identificeren dat mogelijk ongepast, verwarrend of aanstootgevend is voor bepaalde groepen examenkandidaten. Onafhankelijke deskundigen kunnen beoordelen in welke mate de inhoud van de items overeenkomen met de inhoudscategorieën in de toetsspecificaties en of de toets in zijn geheel een evenwichtige dekking van de beoogde inhoud vormt.

Tenslotte adviseren de Amerikaanse standaarden dat er een toetsrationale wordt gemaakt waarin verantwoording over de samenstelling van het examen wordt afgelegd. Deze wordt mogelijk aan het eind van de constructiefase, dus na de samenstelling van het examen, geschreven.

46 Cito: Een kijkje achter de schermen van het normeren: https://www.cito.nl/-/media/files/voortgezet-onderwijs/centrale-examens/achtergrondinformatie-artikelen-discussie-over-examens/cito-achtergrondinformatie-ce-publicatie-een-kijkje-achter-de-schermen-van-het-normeren.pdf

47 Regeling van het College voor Toetsen en Examens van 30 november 2015, nummer CvTE-15.02159. In: Staatscourant, nr. 4817, 4 februari 2016. Geraadpleegd op 1 november van http://www.stilus.nl/examen/n-term-2016.pdf

Tabel 4.3: Checklist fase 3: Construeren van de toets en normeren

	Checklist Fase 3	**Bron**
3.1	De vragen zijn correct geformuleerd	Amerikaanse standaarden: st. 4.7 RCEC: criterium 2.6
3.1a	De vragen of opdrachten zijn gestandaardiseerd.	Amerikaanse standaarden: st. 4.7 RCEC: criterium 2.1
3.1b	De vragen of opdrachten zijn zodanig ontworpen dat fouten bij invulling voorkomen worden.	Amerikaanse standaarden: st. 4.7 RCEC: criterium 2.3
3.1c	De instructie (vaak eerste bladzijde toetsboekje) voor de kandidaat is gestandaardiseerd, volledig en duidelijk.	Amerikaanse standaarden: st. 4.7 RCEC: criterium 2.5
3.1d	De instructie voor de kandidaat dient minimaal te bevatten: aantal vragen, wijze waarop antwoorden worden gegeven, (deel) score per vraag of opdracht, de maximaal te behalen score en de cesuur; toegestane hulpmiddelen; beschikbare tijd en wat ingeleverd moet worden bij afronding; beoordelingspunten bij open vragen.	Amerikaanse standaarden: st. 4.16 RCEC: criterium 2.5
3.1e	De kwaliteit van de lay-out en vormgeving is in orde.	Amerikaanse standaarden: st. 4.16 RCEC: criterium 2.7
3.2	De vragen zijn voldoende gedetailleerd, zodat examenkandidaten de vraag kunnen beantwoorden zoals bedoeld.	Amerikaanse standaarden: st. 4.16
3.3	De moeilijkheidsgraad van de vraag is afgestemd op de beoogde doelgroep.	Amerikaanse standaarden: st. 4.10 RCEC: criterium 3.2
3.3a	Er heeft een pretest plaatsgevonden om informatie te verkrijgen over de moeilijkheidsgraad (kwantitatieve evaluatie).	Amerikaanse standaarden: st. 4.10 RCEC: criterium 3.2
3.3b	Er heeft een kwalitatieve evaluatie met deskundigen plaatsgevonden om informatie te verkrijgen over de moeilijkheidsgraad.	Amerikaanse standaarden: st. 4.8 RCEC: criterium 3.2
3.4	Er wordt een norm(cesuur) verstrekt die via een wetenschappelijke methodiek (standaardbepaling) is bepaald.	Amerikaanse standaarden: st. 5.21 RCEC: criterium 5.1
3.5	De manier waarop de norm wordt bepaald is gedocumenteerd.	Amerikaanse standaarden: st. 5.21
3.6	De standaardbepaling is correct uitgevoerd.	Amerikaanse standaarden: st. 5.21 RCEC: criterium 5.2
3.6a	De standaardbepalingsmethode is op de juiste wijze uitgevoerd.	Amerikaanse standaarden: st. 5.21 RCEC: criterium 5.2
3.6b	De vakdeskundigen/experts die bij de standaardbepaling betrokken zijn, zijn naar behoren geselecteerd en getraind.	Amerikaanse standaarden: st. 5.21 RCEC: criterium 5.2
3.6c	Er is voldoende overeenstemming tussen de vakdeskundige experts.	Amerikaanse standaarden: st. 5.22 RCEC: criterium 5.2
3.7	De procedures die worden gebruikt om items te ontwikkelen, beoordelen en uitproberen en om items uit de vragenbank te selecteren, zijn gedocumenteerd.	Amerikaanse standaarden: st. 4.7
3.8	Er is gedocumenteerd in hoeverre het inhoudsdomein van de toets het domein vertegenwoordigt, dat is gedefinieerd in de toetsspecificaties.	Amerikaanse standaarden: st. 4.12
3.9	Er is gedocumenteerd wat de kwalificaties, relevante ervaringen en demografische kenmerken, evenals de instructies en training die de beoordelaars t.b.v. het beoordelingsproces krijgen, zijn.[48]	Amerikaanse standaarden: st. 4.8
3.10	Er is een rationale voor de toets waarin verantwoording wordt afgelegd over de samenstelling van de toets of het examen.	Amerikaanse standaarden: st. 1.2

48 Bij het gebruik van empirische analyses gelden andere aanvullende criteria. Zie in de Amerikaanse standaarden de criteria 4.09 t/m 4.11.

4.4 Fase 4. Afnemen

In de (inter)nationale kwaliteitsstandaarden staan verschillende criteria beschreven voor het afnemen van een examen. Deze criteria hebben voornamelijk te maken met de betrouwbaarheid en de transparantie naar examenkandidaten. Maar daar waar het *high stakes* examens betreft, kunnen er in deze fase ook aanzienlijke bedreigingen voor de validiteit van de toetsscores optreden. Bijvoorbeeld wanneer de geheime opgaven voortijdig uitlekken of als er mogelijkheden zijn tot fraude, dan is de validiteit van de toetsscores in het geding. Immers, als een kandidaat van tevoren op de hoogte is van de opgaven, kan hij/zij zich met of zonder externe hulpmiddelen perfect voorbereiden op het geven van een goed antwoord, zonder dat hij/zij daadwerkelijk de betreffende leerstof beheerst. Op basis van de aan ons geleverde documenten is van dergelijke incidenten bij de afname van deze twee examens echter geen sprake geweest. We concluderen daarom dat een check op de criteria voor deze fase verder buiten beschouwing kan blijven.

4.5 Fase 5. Beoordelen, verwerken, analyseren

In deze fase van de toetscyclus worden de antwoorden van open vragen beoordeeld door beoordelaar(s), worden de antwoorden op de gesloten meerkeuze vragen (automatisch in het geval van digitale afname of met behulp van schrapkaarten) gescoord. De scores op de examens worden vervolgens geanalyseerd met behulp van psychometrisch onderzoek. Hiermee wordt onderzocht hoe nauwkeurig de score voor kandidaten op het examen is. Bij de centrale examens betekent dit dat de standaardmeetfout wordt berekend. Dat gekeken wordt naar de kwaliteit[49] van de individuele opgaven en dat de kwaliteit van het examen als geheel wordt onderzocht op basis van de resultaten van kandidaten. Hoe meer kandidaten, hoe betrouwbaarder dit psychometrische onderzoek. Het voert te ver om hier diep in te gaan op de verschillende methoden van onderzoek. Belangrijk is dat in alle internationale standaarden wordt genoemd dat de nauwkeurigheid van de toetsscores moet worden onderzocht in de empirie. Afhankelijk van het toetsdoel wordt gekozen voor de meest geschikte uit verschillende methoden om betrouwbaarheid aan te tonen[50]. In de Amerikaanse standaarden (met name standaard 4.13) wordt erop gewezen dat een toetsontwikkelaar, voorzover mogelijk, moet onderzoeken hoe construct-irrelevante variatie in de examens kan worden voorkomen en gereduceerd. Construct-irrelevante variatie betekent dat opgaven niet meten wat ze beogen te meten, en dit is een bedreiging voor de validiteit van de toetsscore. Dit dient eigenlijk alleen te gebeuren als in deze fase 5 de statistische data aanleiding geven tot verdenking van construct-irrelevante variatie. Het is ook gebruikelijk in deze fase 5 om te analyseren of de hoeveelheid examentijd voldoende was voor de kandidaten.

Volgens de Amerikaanse standaarden moeten er door de examenmaker duidelijke en voldoende gedetailleerde procedures voor het scoren en beoordelingscriteria worden opgesteld. Met het oog op de betrouwbaarheid van de score is het van belang om ervoor te zorgen dat het scoren op een zo duidelijk en nauwkeurig mogelijke manier gebeurt. De beoordeling van de antwoorden op de open opgaven dient op een zo objectief mogelijke manier plaats te vinden en bij *high stakes* examens bovendien door meer dan één beoordelaar, onafhankelijk van elkaar. De Amerikaanse standaarden noemen het gebruikelijk om in beoordelingsvoorschriften meerdere voorbeelden van antwoorden op elk scoreniveau te geven (standaard 4.18). Indien van toepassing dienen examenmakers selectiecriteria op te stellen voor degene die verantwoordelijk zijn voor het scoren van de examens. Ook de procedures voor training, kwalificatie en monitoring van beoordelaars moeten worden gedocumenteerd.
In het onderhavige onderzoek met de focus om na te gaan of de examenontwikkelaars bewijzen voor validiteit leveren, is het van belang om in fase 5 na te gaan of en hoe de betrouwbaarheid van het examen wordt berekend en hoe er is omgegaan met de beoordeling van de open vragen.

49 De diverse parameters van opgaven blijken uit de Toets- en Itemanalyse via het programma TIA-plus dat Cito gebruikt in fase 5. Dit zijn onder meer: de moeilijkheid van een opgaven voor de populatie en het vermogen van een opgave om te onderscheiden tussen sterke en zwakke kandidaten. Daarnaast worden de gemiddelde moeilijkheid en de betrouwbaarheid van het examen berekend.

50 Sluijter, C., Hemker, B., & Eggen, Th. (2018b). *Beoordelen van de kwaliteit van toetsen en examens, deel 2: de praktijk*. Arnhem: Cito.

Tabel 4.4: Checklist Fase 5: Beoordelen, verwerken, analyseren

	Checklist Fase 5	Bron
5.1	Procedures voor het scoren en scorecriteria zijn gedocumenteerd	Amerikaanse standaarden: st. 4.18
5.2	Het proces van het selecteren, trainen, kwalificeren en monitoren van de beoordelaar(s) die scoort (scoren) is gedocumenteerd.	Amerikaanse standaarden: st. 4.20
5.3	Er zijn twee onafhankelijke beoordelaars die het examen apart van elkaar scoren.	Amerikaanse standaarden: st. 4.18
5.4	De twee onafhankelijke scores worden allebei gerapporteerd (n.b. huidige procedure is wettelijke procedure[51])	Amerikaanse standaarden: st. 6.9
5.5	Er is gedocumenteerd hoe de uiteindelijke score voor het examen wordt berekend.	Amerikaanse standaarden: st. 6.10
5.6	Op elk scoreniveau worden in de beoordelingsvoorschriften meerdere voorbeelden van antwoorden gegeven.	Amerikaanse standaarden: st. 4.18
5.7	Er wordt een toets- en itemanalyse uitgevoerd onder voldoende kandidaten	Amerikaanse standaarden: st. 4.10
5.8	De toets- en itemanalyse maakt de betrouwbaarheid van de toetsscores zichtbaar met behulp van Cronbach's alfa of een andere algemeen geaccepteerde maat voor betrouwbaarheid.	Amerikaanse standaarden: st. 2.3
5.9	De betrouwbaarheid die berekend is geeft geen aanleiding om te twijfelen aan de validiteit van de toetsscores.	Amerikaanse standaarden: st. 2.3
5.10	Indien sprake is van open vragen, wordt via statistisch onderzoek de mate van beoordelaarsovereenstemming onderzocht.	Amerikaanse standaarden: st. 4.20

4.6 Fase 6. Registreren en communiceren

In de (inter)nationale kwaliteitsstandaarden staan verschillende criteria beschreven voor het registreren van en communiceren over examenresultaten. Deze criteria hebben echter voornamelijk te maken met de transparantie naar examenkandidaten. Volgens de (inter)nationale standaarden worden in deze fase geen bewijzen geleverd voor validiteit.

4.7 Fase 7. Evalueren en verbeteren

Voor het evalueren en verbeteren is het belangrijk dat de stappen die tijdens het ontwerp- en ontwikkelingsproces zijn genomen om bewijs te leveren voor validiteit en betrouwbaarheid worden gedocumenteerd. Volgens de Amerikaanse standaarden zijn de ondersteunende documenten voor examens het belangrijkste middel waarmee examenmakers met -kandidaten communiceren. Het is van belang dat deze documenten worden beoordeeld op hun volledigheid, nauwkeurigheid en duidelijkheid.

Tabel 4.5: Checklist Fase 7: Evalueren en verbeteren

	Checklist Fase 7	Bron
7.1	Toetsspecificaties worden aangepast als er significante veranderingen zijn in het vertegenwoordigde domein.	Amerikaanse standaarden: st. 4.24
7.2	Bewijzen voor (inhouds)validiteit zijn vastgelegd	Amerikaanse standaarden: st. 4.0
7.3	Bewijzen voor betrouwbaarheid (waaronder de mate van beoordelaarsovereenstemming) zijn vastgelegd	Amerikaanse standaarden: st. 4.0
7.4	Procedures rondom toetsontwikkeling zijn vastgelegd	Amerikaanse standaarden: st. 4.0
7.5	Rollen, verantwoordelijkheden en taakverdeling rondom toetsontwikkeling zijn vastgelegd	Amerikaanse standaarden: st. 4.0
7.6	Er wordt onderzoek gedaan onder de stakeholders (belangenvertegenwoordigers van kandidaten en/of van opleiders) die zijn betrokken bij het gebruik van de toets.	Amerikaanse standaarden: st. 1.2
7.7	Er is in een verslag (jaarrapport, evaluatieverslag) waarin verantwoording is afgelegd over de ontwikkeling, samenstelling en afname van de toets.	Amerikaanse standaarden: st. 7.4

51 Eindexamenbesluit VO art. 41 en 42.

5 Onderzoek validiteit CE geschiedenis vmbo gl/tl 2021 eerste tijdvak

5.1 Beschrijving van het examen

Het centraal examen geschiedenis vmbo gl/tl 2021 eerste tijdvak is een papieren examen bestaande uit 51 vragen. Voor dit examen zijn maximaal 65 punten te behalen. Het is opgebouwd uit 29 open, 15 meerkeuze en 7 voorgestructureerde vragen. De afname van het examen vond plaats op dinsdag 25 mei 2021 tussen 09.00 en 11.00.

5.1.1 Domein definitie

De examenstof voor het centraal examen geschiedenis vmbo gl/tl 2021 bestaat uit vijf exameneenheden. Elke exameneenheid bestaat uit één of meer eindtermen.

Exameneenheden

- Leesvaardigheden in het vak geschiedenis en staatsinrichting (GS/K/3)
- De kandidaat kan strategische vaardigheden toepassen die bijdragen tot de ontwikkeling van het eigen leervermogen en het vermogen om met voor geschiedenis en staatsinrichting geëigende vaktaal en methodieken te communiceren en onderzoek te doen
- Staatsinrichting van Nederland (GS/K/5)
- De kandidaat kan herkennen en beschrijven hoe de Nederlandse rechtsstaat/staatsinrichting zich vanaf 1848 tot nu ontwikkeld heeft en deze ontwikkelingen in verband brengen met belangrijke gebeurtenissen en ontwikkelingen in de Nederlandse geschiedenis vanaf 1848.
- Historisch overzicht vanaf 1900 (GS/K/10)
- De kandidaat kan herkennen en beschrijven welke belangrijke gebeurtenissen en ontwikkelingen zich in de Nederlandse en (West-)Europese geschiedenis vanaf 1900 hebben voorgedaan.
- Vaardigheden in samenhang (GS/V/8)
- De kandidaat kan de vaardigheden uit het kerndeel in samenhang toepassen
- Verrijkingsdeel bij historisch overzicht vanaf 1900 (GS/V/9)
- De kandidaat kan de volgende thema's herkennen, beschrijven, verklaren en plaatsen in het kader van het Historisch Overzicht vanaf 1900: Het ontstaan en de gevolgen van het communisme in de Sovjetunie (1917-1941) en het ontstaan en de gevolgen van Indonesië als voorbeeld van dekolonisatie (1942-1949)

Specificatie van eindtermen

In de syllabus zijn de globale eindtermen van de exameneenheden verder gespecificeerd.

- Leesvaardigheden in het vak geschiedenis en staatsinrichting (GS/K/3)
  De kandidaat kan:
  1. Verschillende typen historische vragen herkennen en zo zelfstandig mogelijk formuleren
  2. Bij gegeven of zelf geformuleerde historische vragen informatie verwerven
  3. Verworven of aangereikte informatie verwerken
  4. Principes en procedures die kenmerkend zijn voor de benaderingswijzen van het vak geschiedenis en staatsinrichting toepassen

- Overige exameneenheden
  In de syllabus wordt bij het beschrijven van wat een kandidaat moet kunnen geen onderscheid gemaakt tussen exameneenheden GS/K/5, GS/K/10, GS/V/8 en GS/V/9. De begrippen die de leerlingen moeten kennen zijn gerubriceerd. Enerzijds naar inhoudsdomein (de politieke geschiedenis van Nederland, de wereld en het functioneren van het politieke bestel) en anderzijds naar zes tijdsperioden (de periode 1848-1914, de Eerste Wereldoorlog, het Interbellum, de Tweede Wereldoorlog, de periode 1945-1989 en de Nieuwe Wereldorde). De begrippen gaan bijvoorbeeld over (het verklaren van) de staatkundige kaart en kenmerkende gebeurtenissen, personen en ontwikkelingen.

5.1.2 Domein representatie

Domein representatie betreft de mate waarin een examen het domein zoals dat gedefinieerd is door de toetsspecificaties en in de toetsmatrijs adequaat representeert en meet. Het behoort tot de expertise van de vakdeskundigen in de vaststellingscommissie om te bepalen of de opgaven in voldoende mate het beoogde domein representeren. De toetsdeskundige van Cito stelt voorafgaand aan deze beoordeling de toetsrationale op, zodat de vaststellingscommissie kan vergelijken of het daadwerkelijk samengestelde examen voldoet aan de toetsmatrijs qua verdeling van de vragen over de leerstofonderdelen. In de toetsrationale staat hoeveel scorepunten per onderdeel maximaal behaald kunnen worden en hoe het centraal examen zich verhoudt tot de toetsspecificaties en de toetsmatrijs uit de constructieopdracht.

In de kolomtotalen van de toetsrationale voor het examen geschiedenis vmbo gl/tl 2021-1 is onderscheid gemaakt tussen drie onderdelen. Van de in totaal 65 scorepunten konden er 16 behaald worden op het onderdeel Staatsinrichting, 40 bij het Historisch overzicht en 9 bij het Verrijkingsdeel. Het centraal examen wijkt in dit opzicht enigszins af van de constructieopdracht. Er konden meer punten behaald worden op het onderdeel Staatsinrichting en minder op het Verrijkingsdeel.

Het totaal aantal scorepunten is in de toetsrationale ook uitgesplitst naar de domeinen, namelijk de zes tijdsperioden en chronologie:
- Nederland (1848-1914): 11 scorepunten
- De Eerste Wereldoorlog (1914-1918): 5 scorepunten
- Het Interbellum (1918-1939): 15 scorepunten
- De Tweede Wereldoorlog (1939-1945): 16 scorepunten
- Europa en de wereld (1945-1989): 10 scorepunten
- De nieuwe wereldorde (na 1990): 5 scorepunten
- Chronologie: 3 scorepunten

5.1.3 Domein relevantie

Domein relevantie betreft de mate waarin elke opgave van een examen relevant is voor het beoogde domein. Ook om de relevantie van opgaven te bepalen, is de inzet van vakdeskundigen nodig. In de constructiefase gebeurt dit doordat screeners en constructiegroepleden van Cito zijn betrokken bij de constructie van de opgaven. En ook de vakdeskundigen van de vaststellingscommissie kijken hier grondig naar: ze keuren opgaven die in hun ogen niet relevant zijn af.

Voor alle examens geldt dat er na afname ook gegevens van andere vakexperts beschikbaar zijn over het examen. Middels een enquête (de *Quickscan*) hebben docenten/ examinatoren geschiedenis hun waardering voor het examen geschiedenis vmbo gl/tl 2021 uitgesproken. In deze *Quickscan*[52] werd hun oordeel

52 https://www2.cito.nl/vo/examenverslag/waardering2021/pdf/21193.pdf

gevraagd over de moeilijkheidsgraad van het examen, de lengte van het examen en de aansluiting bij het gegeven onderwijs. De resultaten van 961 respondenten van 711 scholen waren als volgt:

- 43% vond het examen moeilijk, 48% vond het niet te moeilijk/ niet te makkelijk en 5% vond het makkelijk;
- 46% vond het examen te lang en 54% vond het examen qua lengte precies goed;
- 33% vond de aansluiting bij het gegeven onderwijs goed, 55% vond de aansluiting bij het gegeven onderwijs voldoende en 9% vond de aansluiting onvoldoende;
- 13% gaf het examen het cijfer 5, 35% gaf het examen het cijfer 6, 40% gaf het cijfer 7 en 8% gaf het examen het cijfer 8.

5.2 Samenvatting onderzoek validiteit CE geschiedenis vmbo gl/tl 2021 eerste tijdvak

Aan de hand van het onderzoekskader, beschreven in hoofdstuk 4, is het centraal examen geschiedenis vmbo gl/tl uit het eerste tijdvak van 2021 geanalyseerd. In deze paragraaf beschrijven we de belangrijkste bevindingen. Gedetailleerde bevindingen per criterium staan beschreven in Bijlage 2: Checklist Centraal examen geschiedenis vmbo gl/tl 2021 eerste tijdvak. Fasen 4 en 6 uit het toetsproces (Afname en Registratie en communicatie) worden hier buiten beschouwing gelaten.

Fase 1: Basisontwerp

In het onderzoekskader staat dat er een toetsplan moet zijn en wat daarin moet worden gespecificeerd. De term ‘toetsplan’ wordt door het CvTE echter niet gebruikt. Documenten zoals het examenprogramma, de syllabus en de constructieopdracht kunnen echter samen als zodanig worden aangemerkt. Aan de criteria voor een toetsplan wordt in grote mate voldaan. De doelgroep van het examen is vastgesteld, het domein is gedefinieerd, het gebruiksdoel van het examen is gedefinieerd en het is duidelijk in welke contexten de toetsscores gebruikt moeten worden. Een ander criterium in deze fase is het specificeren van hoe de toets moet worden afgenomen en gescoord. We concluderen dat vrijwel volledig aan dit criterium wordt voldaan. Het CvTE beschrijft in de constructieopdracht de examencondities kort en bondig. Verder is een van de criteria het beschrijven van de theorie met betrekking tot de beoogde interpretatie van de scores. In de syllabus zijn vijf globale eindtermen beschreven. De benodigde kennis en vaardigheden bij deze eindtermen zijn nader gespecificeerd. In de syllabus staat niet vermeld op basis van welke specifieke theorie de examenstof is beschreven. Wel is het document opgesteld door een syllabuscommissie van vakexperts, die verantwoorden hoe de syllabus tot stand is gekomen. Hiermee wordt naar onze mening grotendeels aan criterium 1.7 voldaan.
Onze suggestie is om in de syllabus ook te refereren aan de beheersingsniveaus van Bloom (of Romiszowski) en/of aan de Canon geschiedenis. Dit is gebruikelijk en kan behulpzaam zijn om bij het opstellen van de toetsmatrijs in de volgende fase het gewenste niveau van cognitieve handeling bij opgaven te monitoren.

Fase 2: Opstellen toetsspecificaties en toetsmatrijs

In de constructieopdracht worden zowel de algemene toetsspecificaties en bijzonderheden speciaal voor dit examen toegelicht. Het beschrijft de toetsvorm, het soort vragen, de toegestane hulpmiddelen en de voorgestelde lengte en duur van het examen. Het maximaal aantal scorepunten voor het geschiedenisexamen vmbo-gl/tl wordt vermeld, maar het maximaal aantal vragen niet. Het aantal scorepunten is afhankelijk van welke inhoudscomponent er in die vraag getoetst wordt. Elke inhoudscomponent wordt beloond met 1 scorepunt. Hierdoor kan jaarlijks het aantal scorepunten per examen variëren. In de constructieopdracht is beschreven dat de scoreschaal een marge van 10% bevat.
De psychometrische eigenschappen van de opgaven worden niet vooraf in de constructieopdracht vastgelegd, maar spelen wel een rol in de voorbereiding op de normering (na afname van het examen). Voor de beoordeling van de moeilijkheidsgraad van het examen wordt bij de vaststelling van het examen gekeken naar het examen als geheel, en niet per opgave. De Amerikaanse standaarden adviseren om de

psychometrische eigenschappen wel voor samenstelling te bekijken. Hiermee krijg je vooraf al zicht op de moeilijkheidsgraad van de items en is de kans kleiner dat er achteraf items af moeten vallen. Maar dit zou een organisatie van een pretest vergen. In de constructieopdracht wordt de volgorde van items niet gedefinieerd. Het CvTE/Cito licht schriftelijk toe dat het bij geschiedenis vmbo gl/tl de gewoonte is om de vragen op chronologische volgorde te zetten.

In de constructieopdracht wordt in meerdere toetsmatrijzen (zie paragraaf 5.6 van de Constructieopdracht geschiedenis vmbo gl/tl) toegelicht hoe de scorepunten verdeeld moeten worden. Er zijn meerdere tabellen met een voorstel voor de procentuele verdeling van scorepunten over domeinen, voor gedragsaspecten en voor vraagvormen. Eén toetsmatrijs, in de vorm van één overzicht waarin deze verdelingen als een blauwdruk voor het examen zijn samengevoegd, ontbreekt in de constructieopdracht geschiedenis. Opvallend was dat cognitieve beheersingsniveaus (zoals die van Bloom bijvoorbeeld) in de syllabus niet worden genoemd, maar dat in de toetsmatrijs wel onderscheid wordt gemaakt tussen reproductie en productie (gedragsaspecten uit de taxonomie van Romizovski). Het CvTE en Cito verduidelijken dit als volgt: *"Bij sommige vakken wordt vanuit het veld de wens uitgesproken om de beheersingsniveaus en de gebruikte taxonomie uit te werken in de syllabus. Bij geschiedenis vmbo is dit geen wens geweest. Een taxonomie is vooral een onderwijshulpmiddel en niet per se een toetsmiddel. In de toetsmatrijs wordt een grove indeling van de beheersingsniveaus toegepast. Examenkandidaten moeten beschikken over reproductieve en productieve vaardigheden, die aan verschillende taxonomieën te koppelen zijn."*

Uit de aangeleverde documenten is door de onderzoekers in eerste instantie onvoldoende op te maken of de inhoud van het examen voor het beoogde meetdoel relevant is (criterium 2.10) en in hoeverre de toetsmatrijzen een adequate representatie zijn van het meetdoel (criterium 2.11). Uit de schriftelijke informatie die het CvTE/Cito hierover heeft aangeleverd, blijkt echter dat zowel bij het opstellen als de goedkeuring van de toetsmatrijzen diverse vakexperts betrokken zijn. Zij waarborgen *de relevantie* van de toetsinhoud. De vaststellingscommissie ontwerpt de toetsmatrijzen op basis van het examenprogramma en de syllabus; dit waarborgt dat er inhoudelijk is nagedacht hoe representatief het examen is voor het leerstofdomein. Bij de vaststelling wordt bekeken of het aantal vragen voldoende dekking geeft voor het meetdoel. De Amerikaanse standaarden benadrukken dat het belangrijk is dat vakexperts betrokken zijn in deze inhoudelijke fase. Dit komt ten goede aan de validiteit. De leden van de vaststellingscommissie geschiedenis zijn bevoegd voor het vak waarvoor zij in de vaststellingscommissie zitten en hebben ervaring in examenklassen. De benoemingsduur van de vakexperts als lid van een vaststellingscommissie is meestal voor een jaar, maar met mogelijkheid te verlengen. In de praktijk zijn de vakexperts meerdere jaren werkzaam in een vaststellingscommissie. Dit laatste zorgt mede voor continuïteit bij het vaststellen van de centrale examens.

Fase 3: Constructie en Normering

Fase 3 gaat over het construeren en normeren van examenopgaven. Het algemene deel van de constructieopdracht verheldert hoe de opgaven moeten worden vormgegeven door te verwijzen naar handboeken die zijn opgesteld voor de constructeurs. Zowel de vragen als de instructies voor de examenkandidaten zijn daarin gedetailleerd voorgeschreven en gestandaardiseerd. De instructie die in het opgavenboekje wordt gegeven is zeer compleet. Het enige wat ontbreekt, als we kijken naar de criteria in de gehanteerde Amerikaanse standaarden, is dat de cesuur niet in het opgavenboekje vermeld wordt. Bij het centraal examen wordt de N-term (de cesuur) pas achteraf bekend gemaakt. De N-term wordt namelijk pas achteraf vastgesteld, afhankelijk van de moeilijkheidsgraad van het examen zoals gebleken uit de toets- en itemanalyse. Dit is een punt om in overweging te nemen; door het ontbreken van de cesuur weet een kandidaat vooraf niet exact hoeveel en welke vragen hij/zij goed moet hebben om het examen te behalen. In een schriftelijke toelichting stelt het CvTE dat er wel voorlichting wordt gegeven over de

betekenis van de N-term[53] en dat examenkandidaten en docenten in het vo een globale inschatting kunnen maken van de normering op het examen op basis van eerdere examens.

Een van de criteria in de Amerikaanse standaarden luidt dat de moeilijkheidsgraad van de examenopgaven is afgestemd op de beoogde doelgroep. Dit aspect is gewaarborgd door de constructiegroep, die bestaat uit drie of vier docenten geschiedenis met ervaring in eindexamenklassen én een toetsdeskundige met ervaring in het maken van examens. Doordat deze vakexperts de opgaven construeren, wordt aangesloten bij wat de doelgroep aankan qua moeilijkheidsgraad. Zij brengen hun praktijkervaring in bij de constructie van de opgaven. De moeilijkheidsgraad wordt echter niet op basis van empirische gegevens bepaald voorafgaande aan de afname middels een pretest. Om verschillende redenen (o.a. vanwege actualiteit van de examenopgaven) is pretesten niet haalbaar bij deze vakken. De constructiegroep doet zelf geen uitspraak over de moeilijkheidsgraad van een examen; deze taak ligt bij de vaststellingscommissie. De vaststellingscommissie stelt voorafgaand aan de afname van het examen de moeilijkheidsgraad vast in vergelijking met het referentie-examen. Na de afname blijkt de daadwerkelijke moeilijkheidsgraad uit de toets- en itemanalyse.
De Amerikaanse standaarden adviseren dat een examen wordt samengesteld uit gepreteste opgaven (criterium 3.3a). Voor dit examen geschiedenis vmbo gl/tl is er geen pretest geweest. Hiervoor geeft CvTE/Cito de volgende reden:
"Voor niet alle examens is echter het pretesten van opgaven mogelijk, omdat bijvoorbeeld de opgaven als toets […] moeten worden uitgezet bij een doelgroep die voorbereid is op deze onderwijsstof, maar waarvoor tegelijk geldt dat deze groep zelf niet het examen met deze opgaven in het examenjaar krijgt."

Volgens het CvTE/Cito is een groot nadeel van pretesten dat de opgaven twee jaar eerder geproduceerd moeten worden. Er heeft voor het geschiedenisexamen vmbo gl/tl wel een standaardbepaling plaatsgevonden ten behoeve van de normering. Er wordt een wetenschappelijke procedure gehanteerd om de norm te bepalen. De standaardbepaling vindt plaats volgens de Angoff methode, om een inschatting te maken van de moeilijkheidsgraad van het examen ten behoeve van de normering en het bepalen van de N-term. De Angoff-methode gaat uit van de gemiddelde kandidaat. De manier waarop de N-term wordt bepaald is gedocumenteerd in de Staatscourant.

Voor de productie van examens zijn er vertrouwelijke handboeken beschikbaar voor de constructeurs van examens en de vaststellingscommissie (criterium 3.7). In de constructieopdracht en via de toetsmatrijzen is gedocumenteerd in hoeverre het inhoudsdomein van de toets het domein vertegenwoordigt dat is gedefinieerd in de toetsspecificaties (criterium 3.8). De toetsrationale verantwoordt de samenstelling van het examen in relatie tot de toetsmatrijzen die beschreven staan in de constructieopdracht. Ook wordt er in de toetsrationale een inschatting gegeven van de moeilijkheidsgraad per vraag door de toetsdeskundige van Cito. De enige discrepantie die we hier vonden, is dat er geen beschrijving is gegeven van hoe lang de kandidaat over een vraag doet (inschatting van de tijd), terwijl deze volgens de constructieopdracht 1.6.1 wel gegeven zou moeten worden (criterium 3.10). Verder worden de kwalificaties en relevante ervaring van de beoordelaars in het beoordelingsproces vastgelegd en gedocumenteerd (criterium 3.9).

Fase 5: Beoordeling, verwerking en analyse

Er wordt aan alle criteria voor deze fase voldaan. De procedures voor het scoren van de opgaven en de beoordelingscriteria bij opgaven zijn vastgelegd in het correctievoorschrift. Het correctievoorschrift bevat de regels voor de beoordeling, algemene regels, vakspecifieke regels, een beoordelingsmodel, regels over het aanleveren van scores en bronvermeldingen. Daarnaast is er een toets- en itemanalyse uitgevoerd waarin de betrouwbaarheid is bepaald. Opvallende bevindingen uit de toets- en itemanalyse na afname van

53 Een voorbeeld van uitleg over de N-term: https://www.youtube.com/watch?v=EZxx6BdzXmw

het examen worden in het Evaluatieformulier (fase E) met het CvTE gedeeld. De gemiddelde p-waarde in de toets- en itemanalyse is een maat voor de moeilijkheidsgraad. De toets- en itemanalyse na afname van dit examen heeft uitgewezen dat de moeilijkheidsgraad van het examen redelijk te noemen is (gemiddelde p = 57,5). Uit de toets- en itemanalyse blijkt dat dit examen een betrouwbaar beeld oplevert, de coëfficiënt alfa is 0,81. Er is daarmee dus ook geen aanleiding om aan de validiteit van de toetsscores te twijfelen.

Er zijn twee beoordelaars die het examen apart van elkaar scoren: de examinator (een docent van de examenkandidaat) en een “gecommitteerde” (tweede corrector) van een andere school. De examinator en de tweede corrector dienen in onderling overleg het behaalde aantal scorepunten vast te stellen. In de applicatie Wolf voert de eerste beoordelaar de score in, waarna de tweede beoordelaar zijn/haar score ook in kan voeren (maar dit is niet verplicht). Tijdens het invullen van de scores kan de tweede beoordelaar ook de scores van de eerste beoordelaar inzien. Wij leiden hieruit af dat in het (overigens wettelijk vastgestelde) examenproces van het CvTE de oordelen van de twee beoordelaars op deze manier niet onafhankelijk van elkaar geregistreerd en gerapporteerd worden. De in overleg bepaalde scores worden uiteindelijk door de eerste beoordelaar (de examinator) gefiatteerd in de applicatie Wolf. Dit is de standaard werkwijze bij de centrale examens. Een consequentie van deze werkwijze is dat de mate van beoordelaarsovereenstemming niet (voorafgaand aan de uitslag) door psychometrisch onderzoek onderzocht kan worden. De monitoring van het beoordelingsproces verloopt daarmee niet volgens de adviezen van de Amerikaanse standaarden, waarin grote waarde gehecht wordt aan een zo objectief mogelijk beoordelingsproces met monitoring van de overeenstemming. Dit komt ten goede aan de betrouwbaarheid van de uitslag voor kandidaten. In een toelichting van het CvTE wordt opgemerkt dat de praktijk van de eerste en tweede correctie door Cito wel na afloop van een afnameperiode van de examens periodiek onderzocht wordt.

Fase 7: Evaluatie

De wijze waarop de toetscyclus wordt geëvalueerd voldoet in principe bijna volledig aan de criteria uit het onderzoekskader. Bij wijzigingen worden de toetsspecificaties aangepast en de bewijzen voor betrouwbaarheid evenals de processen, rollen en verantwoordelijkheden zijn vastgelegd. Bovendien vindt er een onderzoek plaats onder de stakeholders die bij de afname betrokken zijn. Deze Quickscan is uitgevoerd onder 961 docenten geschiedenis op vmbo-scholen in Nederland. De inhoudelijke aansluiting bij het gegeven onderwijs (en daarmee de domeinrepresentatie en domeinrelevantie) wordt door de meerderheid positief bevonden. Dit is een belangrijk bewijs voor inhoudsvaliditeit.

Een van de criteria was dat de bewijzen voor (inhouds)validiteit zijn vastgelegd. Dit gebeurt nu niet op expliciete manier, bijvoorbeeld in de vorm van een verantwoordingsdocument. De heldere en gedetailleerde constructieopdracht, de toetsmatrijzen en de procedures zoals het opstellen en terugverwijzen naar de toetsmatrijs via de toetsrationale wijzen er echter wel op dat er impliciet bewijsvoering voor validiteit wordt verzameld. Door de inzet van een groot aantal vakexperts bij het opstellen en goedkeuren van de syllabi en de constructieopdracht, door het volgen van een duidelijke toetsconstructieprocedure en door het gebruik van handboeken voor de constructie, wordt duidelijk dat er veel aandacht is voor representatieve en relevante opgaven. Dit alles maakt dat de inhoudsvaliditeit geborgd wordt.

Hoe het evaluatieproces aan het einde van de toetscyclus in zijn werk gaat, konden we uit de aangeleverde documenten echter niet precies opmaken. Het CvTE/Cito heeft schriftelijk toegelicht dat dit gebeurt door voorafgaand aan de normering het verslagformulier voor de normering in te vullen en na de normering door middel van het Evaluatieformulier fase E de besluiten te documenteren. Op het verslagformulier voor de normering geeft de vaststellingscommissie een eerste reflectie die relevant is voor het vaststellen van de N-term, bijvoorbeeld als er gecompenseerd moet worden voor fouten of tijdnood. Met het Evaluatieformulier

fase E wordt aan de hand van een aantal gerichte vragen nagegaan hoe het examen op basis van de psychometrische analyses heeft gefunctioneerd. Hierbij wordt teruggeblikt op zaken waar bij de constructie en afname tegenaan gelopen is en wordt aangegeven in hoeverre dit aanpassing van toekomstige examens, constructieopdrachten, syllabi en communicatie nodig maakt.
Daarbij is er aandacht voor eventuele gevolgen voor de normering. Het lijkt ons aanbevelingswaardig om een verslag op te stellen waarin verantwoording wordt afgelegd over de ontwikkeling en afname van het examen en de conclusies over wat dit betekent voor de ontwerpfase van nieuwe examens. Een dergelijke verantwoording ontbreekt momenteel nog. Dit zou aan het Evaluatieformulier E kunnen worden toegevoegd.

Samenvattend

Het examen geschiedenis voldoet in sterke mate aan het onderzoekskader dat op internationale standaarden is gebaseerd. Er is sprake van een goede dekking van het inhoudsdomein, het examen is representatief en de opgaven zijn relevant voor het domein. Volgens een meerderheid van bevraagde docenten uit het veld sluit het examen sluit goed aan bij het onderwijs. Hiermee lijkt er voldoende bewijs voor de inhoudsvaliditeit van het examen. Er is sprake van zeer gedegen, hoogwaardige toetsontwikkelprocedures met ‘checks and balances’ door verschillende vakexperts en door CvTE.
Slechts op een criterium scoort het examen een ‘nee’: er is bewust geen pretest gehouden (3.3a) om overigens begrijpelijke redenen. Dit betekent dat er geen empirische gegevens over de moeilijkheidsgraad bekend waren voorafgaand aan de afname. Dit wordt echter goed gemaakt door de procedure dat er na afname gelegenheid is om naar aanleiding van de toets en itemanalyses kritisch naar de opgaven te kijken. Indien de uitkomsten daar aanleiding toe geven, biedt deze procedure nog mogelijkheid om eventueel slecht functionerende opgaven te verwijderen voor de uitslag. Veel onderzochte aspecten zijn impliciet in orde, maar niet alle stappen en details zijn in die mate expliciet gedocumenteerd zoals de criteria voorschrijven. Op een aantal criteria (1.6, 1.7, 2.2, 2.7, 2.9, 2.10, 3.10, 5.4, 5.10 en 7.3) scoorde dit examen een ‘(groten)deels’. De belangrijkste daarvan lijkt ons het gegeven dat de oordelen van de twee beoordelaars (correctoren niet onafhankelijk van elkaar kunnen worden gerapporteerd (5.4). Hierdoor kan de beoordelaarsovereenstemming niet worden onderzocht en als kwaliteitscontrole worden gebruikt voordat de uitslag aan de kandidaat wordt gegeven.

6 Onderzoek validiteit CE economie havo 2021 eerste tijdvak

6.1 Beschrijving van het examen

Het centraal examen economie havo 2021 eerste tijdvak bestaat uit 29 vragen, verdeeld over 6 opgaven. Voor dit examen zijn maximaal 58 punten te behalen. Het betreft 20 open en 9 voorgestructureerde vragen. Er is sprake van een opgavenboekje en een bronnenboekje. De te gebruiken bronnen zijn gegroepeerd per opgave. Het examen is op vrijdagmiddag 28 mei 2021 afgenomen en de beschikbare afnametijd was 3 uur.

6.1.1 Domein definitie

Het examenprogramma voor het centraal examen economie havo bestaat uit de volgende domeinen:

- Vaardigheden (A)
  De kandidaat kan economische concepten herkennen en toepassen in uiteenlopende contexten.
- Concept Markt (D)
  De kandidaat kan in contexten analyseren dat keuzes en ruil die plaatsvinden worden gecoördineerd via de markt. Prijsvorming is het coördinatiemechanisme waarmee vraag en aanbod op elkaar worden afgestemd. De manier waarop prijsvorming plaatsvindt is afhankelijk van de marktstructuur (marktvormen) en heeft gevolgen voor toetreding, welvaart en economische politiek.
- Concept Ruilen over de tijd (E)
  De kandidaat kan, binnen de contexten van gezinshuishoudingen, bedrijfshuishoudingen en overheidshuishoudingen, analyseren dat ruil niet alleen op één moment in de tijd plaatsvindt, maar ook over de tijd. De prijs die deze intertemporele ruil coördineert is de rente.
- Concept Samenwerken en onderhandelen (F)
  De kandidaat kan in contexten analyseren dat, wanneer belangen van individuele actoren conflicteren, samenwerken en onderhandelen meer oplevert voor (markt)partijen dan vertrouwen op individuele acties. Centralisatie, waarbij (collectieve) dwang het middel is om acties tot stand te brengen, kan een alternatief coördinatiemechanisme zijn voor keuzes.
- Concept Risico en Informatie (G)
  De kandidaat kan in contexten analyseren dat gezinnen en bedrijven bij het maken van keuzes informatie verzamelen ten einde onzekerheid te verkleinen. Aangezien de informatie vaak een beperkt karakter zal hebben moeten transactiepartijen een inschatting maken van mogelijke gebeurtenissen (risico) en de mate waarin transactiepartners gebeurtenissen beïnvloeden of informatie achterhouden die relevant is voor het tot stand brengen van een transactie (asymmetrische informatie).
- Concept Welvaart en groei (H)
  De kandidaat kan in contexten analyseren wat op nationaal en op mondiaal niveau de oorzaken zijn van economische groei en van de verdeling van inkomen en welvaart. Keuzes op microniveau werken door op macroniveau in elke economie die gekenmerkt wordt door wederzijds afhankelijke markten.
- Concept Goede tijden, slechte tijden (I)
  De kandidaat kan in contexten analyseren waarom er sprake is van korte termijn schommelingen in economische activiteiten en welke mogelijkheden en grenzen er zijn voor conjunctuurbeleid. Markten laten zich niet gemakkelijk reguleren mede door toedoen van rigiditeiten.

Op het centraal examen moeten de kandidaten aantonen dat zij met behulp van economische concepten een context kunnen beschrijven of analyseren. Een context is een voor een kandidaat herkenbare situatie of gebeurtenis. Bijvoorbeeld het wel of niet toelaten van Poolse arbeiders tot de Nederlandse arbeidsmarkt. Kennis over de contexten mag niet van de kandidaten worden verwacht. Daarom worden

‘casus-achtige’ opgaven ontwikkeld, waarbij een beroep wordt gedaan op het verwerken van meerdere soorten bronnen. De te toetsen kennis bestaat uit de conceptuele begrippen en verbanden uit de zes inhoudelijke domeinen (D, E, F, G, H en I) in combinatie met de vaardigheden uit domein A. Voor het examen Economie havo 2021 zijn, gerubriceerd per exameneenheid en subdomein, verschillende toetsspecificaties uitgewerkt.

Domein A: Vaardigheden

De vaardigheden waar kandidaten bij het centraal examen op worden getoetst zijn onder te verdelen in vier subdomeinen: (A1) informatievaardigheden, (A2) rekenkundige en grafische vaardigheden, (A3) standpuntbepaling en (A4) strategisch inzicht. Bij informatievaardigheden gaat het om de actieve beheersing van het economische begrippenapparaat. Voor de toetsing moeten kandidaten problemen oplossen of standpunten toelichten op basis van informatie uit aangeboden bronnen.
Met actieve beheersing wordt bedoeld: herkennen én toepassen. Bij het tweede subdomein worden de beheersing van rekenkundige technieken en algoritmen getoetst en daarnaast worden in elke context bij voorkeur een of meer grafische elementen (grafiek, figuur) opgenomen. Standpuntbepaling houdt in dat kandidaten in staat moeten zijn om gegeven standpunten te herkennen, te beschrijven en/ of te beargumenteren. De syllabus geeft ook aan dat bij ‘een analyseopdracht’ de kandidaat ‘een langer antwoord’ moet produceren ‘dat meerdere aspecten belicht van een of meerdere keuzes’. Standpunt bepalen hangt hier samen met het beheersingsniveau analyseren. Bij het laatste subdomein ‘strategisch inzicht’ geeft de syllabus voorbeelden waarbij sprake is van ‘herkennen’, ‘hanteren’ en ‘onderscheiden’. Bijvoorbeeld het onderscheiden van oorzaak en gevolg of van evenwichtige en onevenwichtige situaties. Het heeft betrekking op meerdere beheersingsniveaus, waarbij ‘hanteren’ geïnterpreteerd mag worden als toepassen.

Overige domeinen

Bij het verder specificeren van de globale eindtermen is ieder domein opgesplitst in een aantal subdomeinen. De bijbehorende economische aspecten staan in de syllabus uitgebreid beschreven. Een voorbeeld hiervan is het Concept Markt (Domein D). Domein D bestaat uit vier subdomeinen, namelijk: (D1) Vraag en aanbod, (D2) Toetreding, (D3) De marktstructuur en (D4) Welvaart en economische politiek. Bij het eerste subdomein ‘Vraag en aanbod’ horen vervolgens negen economische aspecten die kandidaten actief moeten beheersen:

“De kandidaat kan in contexten herkennen en toepassen:

1.1 de wijze waarop consumenten een maximaal verschil nastreven tussen de te betalen prijs en de betalingsbereidheid (de prijs die de consument maximaal bereid is te betalen); (1)
1.2 marktevenwicht (evenwichtsprijs en evenwichtshoeveelheid) dat ontstaat als vraag en aanbod aan elkaar gelijk zijn;
[…]
1.9 winstgevende/ verliesgevende uitbreiding van productie, wanneer de marginale kosten lager/ hoger zijn dan de marginale opbrengsten. (2)”

Daarbij wordt ook de manier waarop het economisch aspect moet worden onderbouwd (grafisch en/of rekenkundig) aangegeven per eindterm. Dergelijke specifieke aanwijzingen per eindterm zijn echter alleen bij domein D gegeven.

6.1.2 Domein representatie

Domein representatie betreft de mate waarin een examen het domein zoals dat gedefinieerd is door de toetsspecificaties en in de toetsmatrijs adequaat representeert en meet. Het behoort tot de expertise van de vaststellingscommissie om te bepalen of de opgaven in voldoende mate het beoogde domein representeren. De toetsdeskundige van het Cito stelt als onderdeel van deze beoordeling de toetsrationale op: zo kan de vaststellingscommissie het daadwerkelijk samengestelde examen vergelijken met wat was aangegeven in de toetsmatrijs qua verdeling van de vragen over de leerstofonderdelen. In de toetsrationale staat hoeveel scorepunten per onderdeel maximaal behaald kunnen worden en hoe het centraal examen zich verhoudt tot de toetsspecificaties en de toetsmatrijs uit de constructieopdracht.

In de kolomtotalen van de toetsrationale voor het examen Economie havo 2021-1 is af te lezen dat er van de 58 scorepunten 21 punten te behalen waren voor het Concept Markt (D), zes op het concept Ruilen over tijd (E), zes voor Samenwerken en onderhandelen (F), vijf voor Risico en informatie (G), vijf voor Welvaart en groei (H) en tot slot vijftien voor het Concept Goede tijden, slechte tijden (I). Voor de domeinen D en I waren dus relatief veel scorepunten te behalen.
In de toetsrationale is ook de verdeling van scorepunten naar vaardigheid opgenomen als percentage van het totale aantal te behalen scorepunten. Van het totale aantal te behalen scorepunten hing 34 procent samen met de vaardigheden ‘Noem, citeer en kies’, 29 procent met ‘Leg uit en verklaar’ en 36 procent met ‘Bereken, teken en arceer’.

6.1.3 Domein relevantie

Domein relevantie betreft de mate waarin elke opgave van een examen relevant is voor het beoogde domein. Ook om de relevantie van opgaven te bepalen, is de inzet van vakdeskundigen nodig. In de constructiefase gebeurt dit doordat ervaren docenten economie in de constructiegroep van Cito zitten, die ervaring hebben in examenklassen. Deze vakexperts construeren de opgaven. En ook de vaststellingscommissie, die bestaat uit vakexperts, kijkt hier grondig naar: ze keuren opgaven die in hun ogen niet relevant zijn af.

In het geval van het examen economie havo 2021-1 zijn er ook gegevens van andere vakexperts beschikbaar. Na afname van het examen hebben docenten en examinatoren in een korte enquête, de zogenaamde Quickscan, de relevantie van het examen economie havo 2021 geëvalueerd[54]. In totaal hebben er 1.174 respondenten van 598 scholen hun oordeel gegeven over de moeilijkheidsgraad van het examen, de lengte van het examen en de aansluiting bij het gegeven onderwijs. Daaruit kwamen de volgende resultaten naar voren:

- 39% vond het examen moeilijk, 53% vond het niet te moeilijk/ niet te makkelijk en 4% vond het makkelijk;
- 28% vond het examen te lang, 71% vond de lengte precies goed en 1% vond het examen te kort;
- 28% vond de aansluiting bij het gegeven onderwijs goed, 56% vond de aansluiting bij het gegeven onderwijs voldoende en 13% vond de aansluiting onvoldoende;
- 13% gaf het examen het cijfer 5, 35% gaf het examen het cijfer 6, 36% gaf het cijfer 7 en 11% gaf het examen het cijfer 8.

54 https://www.examenblad.nl/evaluatie/quick-scan-havo-economie-2021/2021/f=/21151.pdf

6.2 Samenvatting onderzoek validiteit CE economie havo 2021 eerste tijdvak

Het havo-examen economie uit het eerste tijdvak van 2021 is geanalyseerd aan de hand van het onderzoekskader uit Hoofdstuk 4. In deze paragraaf beschrijven we per fase in het toetsproces de belangrijkste bevindingen. Onze bevindingen per criterium staan samengevat in Bijlage 3: Checklist Centraal examen economie havo 2021-1. Fasen 4 en 6 van de toetscyclus, respectievelijk *Afname* en *Registratie en communicatie* zijn hier buiten beschouwing gelaten.

Fase 1: Basisontwerp

Er wordt aan alle criteria voor deze fase voldaan. Het begrip toetsplan wordt niet gehanteerd door het CvTE/ Cito. De combinatie van het examenprogramma, vakspecifieke informatie, syllabus en constructieopdracht bevatten alle gevraagde informatie.
De doelgroep van het examen is vastgesteld, het domein is gedefinieerd, het gebruiksdoel van het examen is gedefinieerd, er is gespecificeerd hoe de toets moet worden afgenomen en gescoord en de theorie met betrekking tot de beoogde interpretatie van de scores staat beschreven in de Syllabus.

Fase 2: Opstellen toetsspecificaties en toetsmatrijs

Hoewel niet alle aspecten die volgens de Amerikaanse standaarden in de constructieopdracht behoren te staan, daar ook in te vinden zijn, zijn deze zaken wel impliciet aanwezig, dan wel vastgelegd in andere documenten. Dit blijkt uit aanvullende schriftelijke informatie die door het CvTE en Cito is aangeleverd.
Er zijn voor dit examen toetsspecificaties opgesteld, verdeeld over de syllabus en de constructieopdracht. Deze bevatten een beschrijving van het aantal vragen met bijbehorende scorepunten, de toetsvorm en de soort vragen, de toegestane hulpmiddelen de voorgestelde lengte van de toets en de hoeveelheid tijd die ervoor is toegestaan.

Er zijn vooraf geen gewenste psychometrische eigenschappen van opgaven vastgesteld via een pretest. Voor de beoordeling van de moeilijkheidsgraad wordt er door het Cito in de toetsrationale een inschatting van de moeilijkheidsgraad per opgave gegeven. Door de vaststellingscommissie van CvTE wordt er gekeken naar het examen als geheel, en niet per vraag. Een belangrijke kwaliteitscontrole wordt uitgevoerd na de afname. De empirische gegevens over de examenopgaven worden, via de toets- en itemanalyse (TIA) na afname, meegenomen in de voorbereiding op de normering (Fase 5).
De constructieopdracht beschrijft dat het examen dient te beginnen en eindigen met een relatief eenvoudige opgave. Het CvTE/Cito vult schriftelijk aan dat een bepaalde volgorde niet gewenst is bij dit examen: *“contexten staan centraal en er komen bij elke context diverse concepten (c.q. domeinen uit het examenprogramma) aan de orde”*.

Er is een erg basale toetsmatrijs in de vakspecifieke constructieopdracht Economie waarin in percentages een verdeling van de vragen over de leerstofeenheden is voorgeschreven. Opgemerkt wordt dat ieder domein wordt getoetst. Er wordt in de constructieopdracht globaal gerefereerd aan de beheersingsniveau's van Bloom, maar die zijn niet in de toetsmatrijs opgenomen. Het gaat om "memoriseren, begrijpen, toepassen, analyseren, evalueren en creëren. De indeling geeft handvatten om globaal aan te geven welke beheersingsniveaus in het centraal examen aan de orde moeten komen en om een niveauonderscheid te maken tussen havo en vwo”[55].

De toetsmatrijs die onderdeel is van de constructieopdracht voldoet niet op alle punten expliciet aan het door ons gehanteerde onderzoekskader. Zo staat niet expliciet in de toetsmatrijs beschreven hoe de relevantie van de inhoud aannemelijk wordt gemaakt (criterium 2.10) en is ook onvoldoende op te maken in hoeverre de toetsmatrijs een adequate representatie is van het meetdoel (criterium 2.11). Er ontbreekt een expliciete beschrijving van de procedures om te toetsmatrijs te maken en vast te stellen. CvTE vult schriftelijk aan:

55 Zie Constructieopdracht vakspecifiek voor Economie, p. 1

"toetsmatrijs sluit aan en is opgenomen in de constructieopdracht (beschrijving van de inhoud van de toets met de verschillende onderdelen, met waardes daaraan gehangen, aanduiding van omvang domein). Er zijn verschillen in de toetsmatrijzen tussen vakken. De constructie opdracht (aan Cito) wordt opgesteld door CvTE op basis van de syllabus. en vastgesteld door de vaststellingscommissie waar 4 vakexperts zitting in hebben. De constructie opdracht (aan Cito) wordt opgesteld door CvTE op basis van de syllabus. en vastgesteld door de vaststellingscommissie waar 4 vakexperts zitting in hebben"

Aanvullende schriftelijke informatie van het CvTE/ Cito leert dat de relevantie van de inhoud van het examen en de adequate representatie in de praktijk geborgd worden door de vakexpertise van de leden van de vaststellingscommissie (vakinhoudelijke experts) die de toetsmatrijs ontwikkeld hebben op basis van het examenprogramma en de syllabus. De leden van de vaststellingscommissie zijn bevoegde docenten voor ecomomie en hebben ervaring in examenklassen. De benoemingsduur van de vakexperts als lid van een vaststellingscommissie is meestal voor een jaar, maar met mogelijkheid te verlengen. In de praktijk zijn de vakexperts meerdere jaren werkzaam in een vaststellingscommissie. Dit laatste zorgt mede voor continuïteit bij het vaststellen van de centrale examens.
CvTE licht verder nog toe:

" Tijdens het constructieproces van individuele vragen en opgaven én bij het samenstellen van een compleet examen worden vragen ingedeeld naar drie categorieën van beheersingsniveau, zoals zichtbaar wordt in de toetsmatrijs. Deze indeling sluit aan bij het gebruik van handelingswerkwoorden in de syllabus. Deze indeling naar drie categorieën functioneert in de praktijk goed. Onderzocht zou kunnen worden of een indeling volgens de taxonomie van Bloom voordelen op kan leveren voor de kwaliteit van de toetsing."

Het aantal vragen geeft voldoende dekking van het meetdoel naar het oordeel van de betrokken toetsdeskundige en de vakinhoudelijke experts uit de vaststellingscommissie. Als er inhoudelijk iets moet worden gewijzigd in de constructieopdracht wordt dit door de vaststellingscommissie, na overleg met de toetsdeskundige van Cito, vastgelegd in het Evaluatieformulier fase E.

Fase 3: Constructie en Normering

In de Constructieopdracht wordt verwezen naar handboeken voor productie van examens waarin staat beschreven hoe de opgaven moeten worden vormgegeven. Deze handboeken, gebruikt door deskundige vakexperts, waarborgen standaardisatie (criterium 3.1).

De instructie voor de kandidaat is gestandaardiseerd en duidelijk. Deze wordt als onderdeel van het examen op diverse manieren bekeken en beoordeeld. De instructie voor de kandidaten bevat alle vereiste gegevens, behalve de cesuur (zak/ slaaggrens). Dit is een punt waarop de werkwijze van Cito/ CvTE afwijkt van de Amerikaanse standaarden; ons is bekend dat de definitieve N-term bekend wordt gemaakt na afname, na analyse van de toetsresultaten, op de normeringsvergadering. Dit is een punt om in overweging te nemen, door het ontbreken van de cesuur weet een kandidaat vooraf niet exact hoeveel en welke vragen hij/zij goed moet hebben om het examen te behalen In een schriftelijke toelichting stelt het CvTE dat er wel voorlichting wordt gegeven over de betekenis van de N-term en dat examenkandidaten en docenten in het vo een globale inschatting kunnen maken van de normering op het examen op basis van eerdere examens.

Op basis van het examenboekje en het handboek productie concluderen we dat de kwaliteit van de lay-out en vormgeving in orde is en de vragen voldoende gedetailleerd zijn, zodat kandidaten de vragen kunnen beantwoorden zoals bedoeld. De constructieopdracht beschrijft hoe er wordt gezorgd dat de moeilijkheidsgraad van vragen is afgestemd op de beoogde doelgroep (criterium 3.3): Bij de constructie

waarborgt de constructiegroep met daarin vier vakexperts met ervaring in de eindexamenklassen dat de moeilijkheidsgraad van de vragen is afgestemd op de doelgroep. Deze vakexperts brengen hun praktijkervaring in bij de constructie van de opgaven. De moeilijkheidsgraad wordt echter niet op basis van empirische gegevens bepaald voorafgaand aan de afname door middel van een pretest. Om verschillende redenen (o.a. vanwege de belangrijk geachte actualiteit van de examenopgaven) is pretesten niet haalbaar gebleken bij Economie. Wel is er sprake van een ander evaluatief inhoudelijk oordeel: de vaststellingscommissie, bestaande uit vakexperts, doet ook een inschatting van de moeilijkheidsgraad van een examen. Zij doen dit voorafgaand aan de afname. Na de afname blijkt vervolgens de daadwerkelijke moeilijkheidsgraad uit de toets- en itemanalyse (zie ook hieronder fase 5). Bij het examen Economie heeft er daarnaast een standaardbepaling met onafhankelijke vakexperts plaatsgevonden (criterium 3.4). De vakexperts beoordeelden dit examen op moeilijkheid en gaven een advies over de normering. Op basis van werkdocumenten die het CvTE ons heeft getoond, hebben we kunnen vaststellen dat de standaardbepaling volgens een wetenschappelijke methode correct is uitgevoerd[56]. Er is toegelicht dat er voorafgaand aan de afname voldoende overeenstemming was tussen deze vakexperts over de moeilijkheidsgraad van het examen. De standaardbepaling leidt vervolgens tot een advies voor een norm (de zak/slaaggrens). Deze gevolgde procedures geven voldoende vertrouwen dat er meerdere onafhankelijke vakexperts een evaluatief oordeel hebben gegeven over de moeilijkheidsgraad en dat CvTE zorgvuldig aandacht besteedt aan de afstemming van de moeilijkheidsgraad op de doelgroep via het oordeel van diverse inhoudelijke vakexperts. Dit is in overeenstemming met wat de Amerikaanse standaarden adviseren.

Met een standaardbepaling kan een norm worden vastgesteld. Er staat ook beschreven hoe de norm wordt gehandhaafd zodat de examens over de jaren heen voor meerdere cohorten leerlingen even moeilijk zijn. De normhandhaving gebeurt via toepassing van de N-term. De betekenis van de N-term en de procedure rondom de N-term is gepubliceerd in de Staatscourant[57]. Deze procedure voorziet er onder andere in dat de zak-slaaggrens niet bij de afname aan de kandidaten bekend wordt gemaakt (bijvoorbeeld via vermelding in het examenboekje), zoals de Amerikaanse standaarden adviseren. De definitieve N-term kan pas op basis van de afnamegegevens worden vastgesteld. Na de afname zijn er daarom normeringsvergaderingen waarop CvTE bekijkt op basis van de empirische gegevens over o.a. de moeilijkheid en betrouwbaarheid uit de toets- en itemanalyses of de norm eventueel dient te worden bijgesteld. En na definitieve vaststelling van de N-term volgt de uitslag voor de kandidaten.
Zoals eerder beschreven is de procedure rond de normering in 2021 in aangepaste vorm verlopen, omdat de populatie in 2021 niet overeenkwam met die van eerdere jaren. De examencondities zijn door het Ministerie van OCW en het CvTE voor dit cohort aangepast in verband met de coronacrisis. Deze aanpassing heeft geen gevolgen voor de hierboven beschreven procedures.

Aan de criteria 3.7 tot en met 3.9 wordt voldaan: er is een uitgebreide documentatie in de vorm van een handboek voor constructeurs en een handboek voor de vaststellingscommissie zodat de opgaven uniform en gestandaardiseerd worden geconstrueerd. Ook voor de selectie van opgaven uit een itembank zijn er procedures. In de aangeleverde toetsrationale (criterium 3.10) is voor de afname gedocumenteerd in hoeverre het inhoudsdomein van het examen dekkend is met de toetsspecificaties en de toetsmatrijs. In het geval van het onderzochte examen Economie kwam de inhoud van het examen bijna geheel overeen met wat in de constructieopdracht en toetsmatrijs als gewenste spreiding over het domein werd omschreven. Het domein in de volle breedte gedekt in dit examen. De toetsrationale sluit mogelijk niet helemaal aan bij wat in criterium 3.10 met een toetsrationale wordt bedoeld. Een verantwoording van gemaakte keuzes bij

56 CvTE vermeldt: ‘Er is een standaardbepaling gehouden voor Economie 2021. We gebruiken de Angoff methode, waarbij we niet de grens kandidaat maar de gemiddelde kandidaat gebruiken.’

57 Staatscourant 2021, 18168, Regeling van het College voor Toetsen en Examens van 15 maart 2021, nummer CvTE-21.00446, houdende wijziging van Regeling omzetting scores in cijfers centrale examens en rekentoets VO 2016 in verband met de gevolgen van COVID19 voor de wijze van normering in het examenjaar 2021xx geraadpleegd d.d. 14-1-2022 van: https://zoek.officielebekendmakingen.nl/stcrt-2021-18168.html

de samenstelling van het examen dient volgens de Amerikaanse standaarden meer te omvatten. Daar gaat het om verantwoording van alle keuzes rondom het examen, niet alleen over de realisatie van de toetsmatrijs. In de toetsrationale die de toetsdeskundige van het Cito opstelt bij oplevering van het examen, wordt bijvoorbeeld niet vastgelegd dat er geen pretest is gehouden. En ook de geschatte moeilijkheidsgraad volgens de vaststellingscommissie (een tweede expertoordeel) ontbrak in expliciete vorm in de toetsrationale. Het CvTE lichtte mondeling toe dat de vaststellingscommissie de geschatte moeilijkheidsgraad in de praktijk wel documenteert in de vorm van het vaststellingsformulier fase D of met een apart inschattingsformulier.

Fase 5: Beoordeling, verwerking en analyse

De procedures voor het scores en de beoordelingscriteria zijn gedocumenteerd in het correctievoorschrift, dat zes onderdelen bevat: 1. regels voor de beoordeling, 2. Algemene regels, 3. Vakspecifieke regels, 4. Beoordelingsmodel, 5. Aanleveren scores en 6. Bronvermeldingen.
De beoordelingsvoorschriften in het correctievoorschrift bevatten meerdere voorbeelden van de antwoorden (criterium 5.6)
Het beoordelen (scoren) van een examen gebeurt door twee docenten Economie: een docent van de kandidaat is de examinator en in die hoedanigheid de eerste beoordelaar. Daarna is er een tweede beoordelaar, de gecommitteerde, van een andere school, die het werk van de kandidaat ook beoordeelt. Vervolgens vindt er overleg plaats tussen de eerste en tweede corrector waarin het behaalde aantal scorepunten wordt vastgesteld. Dit is de wettelijke procedure, in deze procedure is niet voorzien dat de oordelen van twee beoordelaars (correctors) onafhankelijk van elkaar worden geregistreerd en gerapporteerd (criterium 5.4).

Er is na de afname van het examen economie havo 2021 een toets- en itemanalyse uitgevoerd die in het evaluatieformulier als volgt wordt samengevat: "Het examen is met een gemiddelde p'-waarde van 0,53 moeilijk voor deze groep kandidaten. 14 vragen hebben een Rir-waarde lager dan 22 waarvan 5 een hoge p'-waarde hebben; deze vragen waren relatief gemakkelijk. 3 vragen hebben een lage p'-waarde (Rir-waarden resp.: 21/17/16; vraag 9 (1-punts rekenvraag), 11 (concept-vraag overheidsschuld en overheidstekort) en 20 (PS; aanvulling). Er staan relatief veel vragen met een hoge p'-waarde tegenover. En er waren relatief veel rekenvragen." (criteria 5.7 en 5.8).

Uit de Quickscan (de enquete die is afgenomen onder de docenten economie) en uit de toets- en itemanalyse na de afname van het examen economie blijkt dat de betrouwbaarheid aan de lage kant is (coëfficiënt alfa = 0,7) voor een high-stakes examen met open vragen met polytome scoring[58]. Dit geeft ons aanleiding om enige twijfel te hebben over de validiteit (criterium 5.9). Ook het gegeven dat het examen moeilijk was voor de populatie en dat er sprake is van een onevenwichtige samenstelling, gezien enkele zeer moeilijke en ook vijf gemakkelijke vragen (vragen met juist hele hoge p-waarden), maakt dat er aanleiding is te twijfelen over de validiteit van de toetsscores.

Er zijn enkele plausibele verklaringen denkbaar voor de lage betrouwbaarheid:

- het examen is niet homogeen genoeg. Er is sprake van te grote spreiding over het domein. Er worden te veel verschillende onderwerpen en contexten bevraagd.[59]
- het examen bevat mogelijk te weinig meetmomenten (scorepunten); het is bekend dat het langer maken van een examen de betrouwbaarheid in positieve zin beïnvloedt.
- het examen bevatte enkele extreem moeilijke vragen (zoals vraag 11 en 20) en juist die dragen niet bij aan de betrouwbaarheid[60].

58 De COTAN acht een betrouwbaarheid lager dan 0,80 onvoldoende voor high stakes examens, zie Sluijter, C., Hemker, B., & Eggen, Th. 2018b, p. 4.
59 Berkel, H. van, (2011). Meten is weten, vergeet het maar! Over het zoeken naar de ware score. In: Examens augustus, nr 3, pp 10-14.
60 Berkel, H. van, (2011) Meten is weten, vergeet het maar! Over het zoeken naar de ware score. In: Examens augustus, nr 3, pp 10-14.

Mede op basis van de gegevens uit de toets- en itemanalyse kan tijdens een normeringsvergadering na afname door de vaststellingscommissie het advies gegeven worden om de N-term aan te passen. Het is mogelijk om bijvoorbeeld een of twee opgaven met slechte psychometrische eigenschappen (zoals bijvoorbeeld zeer moeilijke opgaven met lage p-waarden) te schrappen voor de uitslag. Uit de aangeleverde documentatie van het CvTE blijkt dat de N-term niet is aangepast. Men zag daar voor Economie geen reden toe. De overwegingen waren de volgende: het examen was moeilijk voor de populatie, maar het gemiddelde cijfer was een 6.2. En het percentage kandidaten met een onvoldoende was 24%.

Fase 7: Evaluatie

Het examen voldoet in hoge mate aan de criteria uit de internationale kaders voor de evaluatiefase. Aan de syllabus is met markeringen te zien dat deze voor het examen 2021 is geactualiseerd (criterium 7.1). Er zijn geen wijzigingen in de gebruiksomstandigheden van het examen. De bewijzen voor betrouwbaarheid zijn vastgelegd in de toets- en itemanalyse, maar bewijzen voor de validiteit van de toetsscores (incl. de inhoudsvaliditeit) zijn niet expliciet gedocumenteerd. (criterium 7.2). Er zijn echter duidelijke aanwijzingen dat deze bewijsvoering impliciet wél verzameld wordt. De gedegen toetsontwikkelprocedures en de checks door vakexperts wijzen daarop. Zo zijn er toetsspecificaties in de vorm van een syllabus en een uitgebreide constructieopdracht. En er is een basale toetsmatrijs als blauwdruk voor gewenste samenstelling van het examen. Vervolgens wordt aan het eind van de constructiefase een toetsrationale gemaakt waarin er wordt teruggekoppeld naar de toetsmatrijs. Hierdoor worden belangrijke voorwaarden voor de validiteit: de representativiteit en waarschijnlijk ook relevantie, transparant in kaart gebracht. Aan de criteria 7.4 tot en met 7.6 wordt voldaan: in de constructieopdracht worden procedures en handboeken vermeld ten behoeve van de productie van examens havo. Ook staan de rollen, verantwoordelijkheden en taakverdeling rondom toetsontwikkeling vastgelegd in de constructieopdracht. Er heeft een Quickscan onder examinatoren (vakdocenten) in het veld plaatsgevonden, dit is een evaluatieonderzoek onder 804 docenten economie van 527 scholen die bij het eindexamen Economie betrokken waren. De Quickscan bevraagt de docenten op vier punten:

- moeilijkheidsgraad;
- lengte;
- aansluiting bij het onderwijs;
- waardering.

Uit deze Quickscan blijkt dat 39% van de respondenten het examen moeilijk vonden en 53% het examen als ‘precies goed’ (niet moeilijk/niet makkelijk) waardeerden. Een meerderheid vond de lengte precies goed (71%). Een minderheid vond het examen te lang (28%). En men ervoer de aansluiting bij het onderwijs als voldoende (56%) of goed (28%). Men waardeerde het examen als volgt: 35% het cijfer 7 en 36% het cijfer 8. En slechts 13% vond het examen onvoldoende (cijfer 5). Deze evaluatiegegevens kunnen door de toetsontwikkelaars als een bewijs voor validiteit van de toetsscores opgevoerd worden.

Samenvattend

Het examen Economie voldoet in hoge mate aan de internationale standaarden. Er is sprake van een goede dekking van het inhoudsdomein, het examen is representatief en de opgaven zijn relevant voor het domein. Het examen sluit goed aan bij het onderwijs, volgens een meerderheid van de bevraagde docenten uit het veld. Hiermee lijkt er voldoende bewijs voor de inhoudsvaliditeit van het examen. Er is sprake van zeer gedegen, hoogwaardige toetsontwikkelprocedures met ‘*checks and balances*’ door verschillende vakexperts en door CvTE.

Slechts op een criterium scoort het examen een ‘nee’: er is geen pretest gehouden (3.3a) om overigens begrijpelijke redenen. Dit betekent dat er geen empirische gegevens over de moeilijkheid waren voorafgaand aan de afname. Dit wordt echter goed gemaakt door de procedure dat er na afname gelegenheid is om naar aanleiding van de toets en itemanalyses kritisch naar de opgaven te kijken. Het is in deze procedure mogelijk om slecht functionerende opgaven nog te verwijderen voor de uitslag.
Op een aantal andere criteria (2.7, 2.10, 3.1d, 3.10, 5.4, 5.9, 5.10, 7.3, 7.4) scoorde het examen een ‘deels’. De belangrijkste daarvan lijkt ons het gegeven dat de oordelen van de twee beoordelaars (correctoren) niet volledig onafhankelijk worden gerapporteerd (5.4). Hierdoor kan de beoordelaarsovereenstemming niet worden onderzocht en als kwaliteitscontrole worden gebruikt voordat de uitslag aan de kandidaat wordt gegeven. Daarnaast is de lage betrouwbaarheid van 0,7 een aandachtspunt. We bevelen aan om nader te onderzoeken of dit mogelijk samenhangt met een te brede dekking van het leerstofdomein.

7 Conclusie en aanbevelingen

In de hoofdstukken 5 en 6 werden de bevindingen gepresenteerd van het onderzoek naar twee maatschappijgerichte examens dat is uitgevoerd met behulp van het onderzoekskader uit hoofdstuk 4. Hier presenteren we de antwoorden op de onderzoeksvragen en sluiten af met enkele aanbevelingen.
Op basis van de geanalyseerde internationale standaarden is er een onderzoekskader ontwikkeld waarmee twee maatschappijgerichte centrale examens uit het eerste tijdvak van 2021 zijn onderzocht: Economie havo en geschiedenis vmbo gl/tl. Het onderzoekskader omvat 48 criteria ingedeeld naar vijf relevante fasen van de toetscyclus. Het gehanteerde kader is daarmee uitgebreider dan het kader dat door RCEC in eerdere onderzoeken is gebruikt[61]. Dit kader biedt een breder perspectief op de validiteit van de toetsscores (waaronder ook inhoudsvaliditeit) van de examens. Dit kan ertoe leiden dat er aspecten van het toetsproces komen bovendrijven die bij het beperktere onderzoek buiten beschouwing zouden zijn gelaten.

De onderzoeksvraag luidde: '*Voldoen de procedures die het CvTE en Cito bij de borging van de inhoudsvaliditeit van de maatschappijgerichte examens hanteren aan (inter)nationale kwaliteitsstandaarden?*'

Op basis van dit onderzoek is het mogelijk de bovengenoemde onderzoeksvraag positief te beantwoorden. Samenvattend kan worden gesteld dat beide examens in hele grote mate voldoen aan de internationale kwaliteitsstandaarden. De volgende aspecten onderbouwen de inhoudsvaliditeit en de validiteit van de toetsscores:

1. In de eerste drie fasen van de toetscyclus wordt een groot aantal vakexperts betrokken bij het opstellen van de toetsspecificaties, de syllabi en de constructie van de opgaven en de samenstelling van de examens. Ook bij de vaststelling van opgaven en examens zijn (andere) vakexperts in vaststellingscommissies betrokken.
2. De inzet van een groep bevoegde vakexperts met ervaring in examenklassen als constructeurs onder leiding van een toetsdeskundige van Cito waarborgt dat in hoge mate kan worden aangesloten bij wat de doelgroep aankan: de moeilijkheidsgraad van de opgaven is een aandachtspunt bij constructie. En ook bij de vaststelling wordt een inschatting van de moeilijkheid gemaakt door een onafhankelijke groep van experts. Dat diverse vakexperts zijn betrokken is een belangrijke waarborg voor validiteit van de toetsscores volgens de Amerikaanse standaarden.
3. Er is sprake van zeer gedegen toetsontwikkel- en samenstellingsprocedures die ondersteund worden door handboeken voor de constructie en de vaststelling.
4. Er wordt een wetenschappelijk verantwoorde manier voor de standaardbepaling gebruikt en correct uitgevoerd.
5. Er is een wetenschappelijke methode voor de normhandhaving via de N-term procedure.
6. De opgaven worden na afname in de empirie op kwaliteit beproefd via een toets- en itemanalyse voordat de definitieve uitslag aan kandidaten wordt gegeven. Hiermee wordt een essentiële kwaliteitscontrole uitgevoerd: opgaven die om welk reden dan ook niet goed hebben gefunctioneerd kunnen zo nog verwijderd worden uit het examen, zodat ze niet meetellen voor de uitslag. Mocht de gerapporteerde coëfficiënt alfa (maat voor betrouwbaarheid) in de toets- en itemanalyse aanleiding zijn om zorgen te hebben over de validiteit, dan wordt dit ook in deze fase, nog voor het bepalen van de definitieve uitslag, duidelijk en zou dit bijgesteld kunnen worden als nodig.

61 Sanders, P., Brouwer, A.J., Veldkamp, B.P. (2018). Onderzoek naar de inhoudsvaliditeit van een tweetal centrale examens voortgezet onderwijs 2017. Enschede: RCEC

Op een aantal punten wijkt de werkwijze van het CvTE en Cito af van wat de Amerikaanse standaarden adviseren. Deze afwijkingen zijn de volgende:

- Er wordt geen cesuur vermeld (bijvoorbeeld in het opgavenboekje) aan de kandidaten. Het is belangrijke informatie voor kandidaten en het komt de transparantie over de examens ten goede. In een schriftelijke toelichting geeft het CvTE aan dat er de laatste jaren voor kandidaten en docenten in het veld meer voorlichting wordt gegeven over de betekenis van de N-term[62]. De N-termen van voorgaande examens zijn bekend, waardoor kandidaten van tevoren wel een inschatting kunnen maken van wat ze ongeveer zouden kunnen verwachten (met een kleine marge)
- De twee oordelen (1e corrector en 2e corrector) worden niet onafhankelijk geregistreerd en gerapporteerd. Dit is overigens in overeenstemming met het wettelijke geregelde beoordelingsproces. Hierdoor is het niet mogelijk om de beoordelaarsovereenstemming bij de examens met een groot aantal open vragen (zoals het examen economie havo) te onderzoeken en de uitkomsten daarvan te laten meewegen in beslissingen over kandidaten of om beoordelaars te monitoren.
- Er is een groot aantal documenten om het werk aan de ontwikkeling van de opgaven en examens te ondersteunen. Deze documenten bestrijken elk een deel van de toetsontwikkeling of een deel van de evaluatie. Dit maakt het lastig om een transparant beeld van de gemaakte keuzes in de toetscyclus te verkrijgen.

Het is duidelijk dat de toetscyclus die het CvTE en Cito voor deze twee examens hebben doorlopen kwalitatief op een hoog peil staat. Het proces komt in hoge mate overeen met wat de Amerikaanse ‘gouden’ standaard voor ogen staat. Dit positieve gegeven willen we hier graag benadrukken. We kunnen daarnaast de volgende aanbevelingen in overweging geven:

- Overweeg voor het examen Economie havo om in de syllabus en in ieder geval in de toetsmatrijs ook de beheersingsniveaus van Bloom (of Romiszowski) op te nemen. Dit is een hulpmiddel om de cognitief erg moeilijke opgaven in het stadium van constructie al op te sporen en te vermijden.
- Overweeg voor de examens met open vragen of de twee oordelen van de correctoren onafhankelijk geregistreerd en gerapporteerd kunnen worden, zodat ook de mate van beoordelaarsovereenstemming een kwaliteitscontrole instrument wordt vóórdat de uitslag gegeven wordt.
- Overweeg om een kort verantwoordingsdocument per examen op te (laten) stellen in de evaluatie fase (fase 7) van de toetscyclus waarin expliciet beargumenteerd wordt welke keuzes in de verschillende fasen van de toetscyclus zijn gemaakt. Hierin kan geëxpliciteerd worden hoe de inhoudsvaliditeit geborgd is, hoe valide de toetsscores zijn en hoe de betrouwbaarheid van het examen was. Hierin kan ook worden beschreven of er wel of niet een pretest was en of al dan niet een standaardbepaling heeft plaatsgevonden. De Amerikaanse standaarden hechten hier veel waarde aan. En het is ook de tiende stap van Veldkamp en de twaalfde stap bij Downing.

62 Een voorbeeld van uitleg over de N-term: https://www.youtube.com/watch?v=EZxx6BdzXmw

Literatuur

AEA-Europe. (2017) European Framework of Standards for Educational Assessment 1.0: https://www.aea-europe.net/wp-content/uploads/2017/07/SW_Framework_of_European_Standards.pdf

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *1999 Standards for Educational and Psychological Testing, 2014 Edition*, Washington, DC: American Educational Research Association. Geraadpleegd op 5 oktober 2021 van: https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Berkel, H. van, (2011) *Meten is weten, vergeet het maar! Over het zoeken naar de ware score*. In: Examens augustus, nr 3, pp 10-14. Geraadpleegd op 10 februari 2022 van http://benwilbrink.nl/EXAMENS_2011-3.pdf

Berkel, H. van, Bax, A. & Joosten-Ten Brinke, D. (Red.) (2013) *Toetsen in het hoger onderwijs, 3e herziene druk*, Houten: Bohn Stafleu van Loghum.

Bijkerk, L. (2015). Basiskwalificatie Examinering in het Hoger Beroepsonderwijs. Bohn Stafleu van Loghum/Springer Media. Geraadpleegd 5 oktober 2021 van: https://link.springer.com/book/10.1007/978-90-368-0933-7

Brouwer, A., Sanders, P., Veldkamp, B. (2019) Onderzoek naar de inhoudsvaliditeit van een tweetal beroepsgerichte centrale examens voortgezet onderwijs 2019. RCEC - Onderzoek in opdracht van het College voor Toetsen en Examens.

Cizek, J. K. (2006). Standard Setting. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Downing, S. M. (2006) (Online edition 2011). Twelve steps for effective test development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Jaspers, M., & Schade, M. (2002). Toets & beleid: toetsbeleid en geautomatiseerde toetsing. [Fontys Hogescholen], Facilitair Bedrijf, afdeling Onderwijs.

Kane, M., (2006). Content-Related Validity Evidence in Test Development. In: S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Linn, R. L. (2006). The Standards for Eduational and Psychological Testing: Guidance in Test Development. In: S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Maassen, N., Otter, D.den, Wools, S., Hemker, B. Straetmans, G., & Eggen, Th. (2014). *Kwaliteit van toetsen binnen handbereik. Een reviewstudie van onderzoek en onderzoeksresultaten naar de kwaliteit van toetsen*, RCEC en Cito.

Newton, P.E., Shaw, S.D., (2014). Validity and validation. In: P. E. Newton, & S. D. Shaw (Eds.), *Validity in Educational & Psychological Assessment*. Sage Publications Ltd. Geraadpleegd op 1 november 2021 van: https://sk.sagepub.com/books/validity-and-educational-assessment/n1.i400.xml

Regeling van het College voor Toetsen en Examens van 30 november 2015, nummer CvTE-15.02159, houdende vaststelling van regels voor de omzetting van scores in cijfers bij centrale examens en de rekentoets in het voortgezet onderwijs (Regeling omzetting scores in cijfers centrale examens en rekentoets VO 2016). In: Staatscourant, nr. 4817, 4 februari 2016. Geraadpleegd op 1 november van http://www.stilus.nl/examen/n-term-2016.pdf

Roelofs, E., & Visser, J. (2017). De inhoud van toetsen. In: P. Sanders (Red.), *Toetsen op school*. Arnhem: Cito.

Sanders, P. (Red.). (2016). Toetsen op school. Hoger Onderwijs. Arnhem: Cito. Geraadpleegd op 2 november van https://www.cito.nl/kennis-en-innovatie/kennisplein

Sanders, P., Brouwer, A.J., Veldkamp, B.P. (2017). Onderzoek naar de inhoudsvaliditeit van de centrale examens en de afhandeling van onvolkomenheden bij de centrale examens. Enschede: RCEC

Sanders, P., Brouwer, A.J., Veldkamp, B.P. (2018). Onderzoek naar de inhoudsvaliditeit van een tweetal centrale examens voortgezet onderwijs 2017. Enschede: RCEC

Sanders, P. (Red.) (2017). Toetsen op school. Arnhem: Cito. 1e druk 2011, herziene versie 2017. Geraadpleegd op 5 oktober van https://www.cito.nl/kennis-en-innovatie/kennisplein

Sanders, P., Brouwer, A., Vegt, A.L. van der, (2019). Hfst 6 Het beoordelen en borgen van de kwaliteit van (studie)toetsen en examens. In: J. Scheerens, A. Brouwer, P. Sanders, B. Veldkamp, & A. L. Vegt, van der (Red.). *Fundamentele vragen over examens en toetsing. Eindrapportage*. Utrecht: Oberon. Geraadpleegd op 5 oktober 2021 van www.oberon.eu

Sluijter, C., Hemker, B., & Eggen, Th. (2018a). *Beoordelen van de kwaliteit van toetsen en examens, deel 1: Systemen en criteria*. Arnhem: Cito. Geraadpleegd op 5 oktober 2021 van www.cito.nl/kennis-en-innovatie/kennisplein

Sluijter, C., Hemker, B., & Eggen, Th. (2018b). *Beoordelen van de kwaliteit van toetsen en examens, deel 2: de praktijk*. Arnhem: Cito. Geraadpleegd op 5 oktober 2021 van www.cito.nl/kennis-en-innovatie/kennisplein

Veldkamp, B. P. (2016). De inhoud en constructie van toetsen. In P. F. Sanders (Red.), *Toetsen op school hoger onderwijs* (pp. 21–30). Cito. https://www.cito.nl/-/media/Files/kennis-en-innovatie-onderzoek/toetsen-op-school/cito-toetsen-op-school-ho.pdf

Wools, S. (2012). Towards a comprehensive Evaluation System for the quality of tests and assessments, in: *Psychometrics in Practice at RCEC* (pp. 95 - 106).

Wools, S. (2015). *All About Validity: towards a new evaluation system for educational assessments*. [Dissertatie. Universiteit van Twente]. Geraadpleegd op 12 oktober 2021 van https://www.researchgate.net/publication/281434844_All_About_Validity_An_evaluation_system_for_the_quality_of_educational_assessment

Wools, S. (2017). De validiteit van Toetsscores. In: P. F. Sanders (Red.), *Toetsen op school*. (pp. 69-83). Herziene versie. Arnhem: Cito. Geraadpleegd op 15 oktober 2021 via www.cito.nl/kennis-en-innovatie/kennisplein

Overige bronnen:

Begrippen uit de TIA: geraadpleegd van Cito: https://www2.cito.nl/vo/share/Begrippen uit de TIA.pdf

Opbouw toets- en itemanalyse: geraadpleegd van Cito: https://www2.cito.nl/vo/share/Opbouw toets-en itemanalyse.pdf

Over de normering bij de Centrale Eindexamens: https://www.examenblad.nl/onderwerp/normering-centrale-examens/2021

Een kijkje achter de schermen van het normeren, geraadpleegd van Cito: https://www.cito.nl/-/media/files/voortgezet-onderwijs/centrale-examens/achtergrondinformatie-artikelen-discussie-over-examens/cito-achtergrondinformatie-ce-publicatie-een-kijkje-achter-de-schermen-van-het-normeren.pdf

Bijlage 1: Overzicht van figuren en tabellen

Bijlage 2: Checklist Centraal examen geschiedenis vmbo gl/tl 2021 eerste tijdvak

	Fase 1: Basisontwerp geschiedenis vmbo gl/tl 2021-1	Aanwezig?
1.1	Er is een toetsplan.	ja
1.2	In het toetsplan is aangegeven wat de doelgroep van het examen is.	ja
1.3	In het toetsplan is aangegeven wat het meetdoel van het examen is oftewel hoe het domein gedefinieerd is.	ja
1.4	In het toetsplan is aangegeven wat het gebruiksdoel van het examen is.	ja
1.5	In het toetsplan is in duidelijke taal gespecificeerd in welke contexten toetsscores gebruikt moeten worden.	ja
1.6	In het toetsplan is in duidelijke taal gespecificeerd met welke processen (papieren toets of digitale toets, welke wijze van afname, welke wijze van beoordeling, hoe komt score tot stand) de toets moet worden afgenomen en gescoord.	grotendeels
1.7	In het toetsplan staat de theorie met betrekking tot de beoogde interpretatie beschreven.	grotendeels

	Fase 2: Opstellen toetsspecificaties en toetsmatrijs geschiedenis vmbo gl/tl 2021-1	Aanwezig?
2.1	Er zijn toetsspecificaties (bv. in de vorm van een constructieopdracht).	ja
2.2	In de constructieopdracht staat het aantal vragen met bijbehorende scorepunten.	deels
2.3	In de constructieopdracht staat de toetsvorm en/of het soort vragen (bijvoorbeeld gesloten en/of open).	ja
2.4	In de constructieopdracht staat de toegestane hulpmiddelen.	ja
2.5	In de constructieopdracht staat de voorgestelde lengte van de toets.	ja
2.6	In de constructieopdracht staat de hoeveelheid tijd die is toegestaan voor de toets.	ja
2.7	In de constructieopdracht staan de gewenste psychometrische eigenschappen van de items en de toets als geheel.	deels
2.8	In de constructieopdracht wordt de volgorde van items gedefinieerd.	ja
2.9	Er is een toetsmatrijs.	deels
2.10	In de toetsmatrijs wordt de relevantie van de inhoud van de toets of het examen voor het beoogde meetdoel aannemelijk gemaakt.	deels
2.11	De toetsmatrijs is een adequate representatie van het meetdoel.	ja
	2.11a) De eind- en toetstermen representeren het meetdoel.	+
	2.11b) De eind- en toetstermen sluiten aan op de inhoud en het vereiste beheersingsniveau van het betreffende meetdoel.	+
	2.11c) Het werkwoordgebruik in de eind- en toetstermen is eenduidig en sluit goed aan bij de gebruikte taxonomie (bijvoorbeeld van Bloom of Romiszowsky).	-
	2.11d) Het aantal vragen geeft een voldoende dekking van het meetdoel.	+
2.12	Relevante vakinhoudelijke experts beoordelen de toetsspecificaties op hun geschiktheid.	ja
2.13	Het doel van deze beoordeling, het proces waarmee de beoordeling wordt uitgevoerd en de resultaten van de beoordeling worden gedocumenteerd.	ja
2.14	De kwalificaties, relevante ervaringen en demografische kenmerken van relevante vakinhoudelijke experts worden gedocumenteerd.	ja

	Fase 3: Constructie en normering geschiedenis vmbo gl/tl 2021-1	Aanwezig?
3.1	De vragen zijn correct geformuleerd (zie richtlijnen in RCEC; die richtlijnen zijn hieronder ingevoegd als subcriteria).	ja
	3.1a) De vragen of opdrachten zijn gestandaardiseerd.	+
	3.1b) De vragen of opdrachten zijn zodanig ontworpen dat fouten bij invulling voorkomen worden.	+
	3.1c) De instructie voor de kandidaat (vaak eerste bladzijde van toetsboekje) is gestandaardiseerd, volledig en duidelijk.	+
	3.1d) De instructie voor de kandidaat dient minimaal te bevatten: aantal vragen, wijze waarop antwoorden worden gegeven, (deel) score per vraag of opdracht, de maximaal te behalen score en de cesuur; toegestane hulpmiddelen; beschikbare tijd en wat ingeleverd moet worden bij afronding; beoordelingspunten bij open vragen.	+/-
	3.1e) De kwaliteit van de lay-out en vormgeving is in orde.	+

	Fase 3: Constructie en normering geschiedenis vmbo gl/tl 2021-1	Aanwezig?
3.2	De vragen zijn voldoende gedetailleerd, zodat examenkandidaten de vraag kunnen beantwoorden zoals bedoeld.	ja
3.3	De moeilijkheidsgraad van de vraag is afgestemd op de beoogde doelgroep.	ja
	3.3a) Er heeft een pretest plaatsgevonden om informatie te verkrijgen over de moeilijkheidsgraad (kwantitatieve evaluatie).	-
	3.3b) Er heeft een kwalitatieve evaluatie met deskundigen plaatsgevonden om informatie te verkrijgen over de moeilijkheidsgraad.	+
3.4	Er wordt een norm/cesuur verstrekt die via een wetenschappelijke methodiek (standaardbepaling) is bepaald.	ja
3.5	De manier waarop de norm wordt bepaald is gedocumenteerd.	ja
3.6	De standaardbepaling is correct uitgevoerd.	ja
	3.6a) De standaardbepalingsmethode is op de juiste wijze uitgevoerd.	+
	3.6b) De vakdeskundigen/experts die bij de standaardbepaling betrokken zijn, zijn naar behoren geselecteerd en getraind.	+
	3.6c) Er is voldoende overeenstemming tussen de beoordelaars.	+/-
3.7	De procedures die worden gebruikt om items te ontwikkelen, beoordelen en uitproberen en om items uit de vragenbank te selecteren, zijn gedocumenteerd.	ja
3.8	Er is gedocumenteerd in hoeverre het inhoudsdomein van de toets het domein vertegenwoordigt, dat is gedefinieerd in de toets specificaties.	ja
3.9	Er is gedocumenteerd wat de kwalificaties, relevante ervaringen en demografische kenmerken, evenals de instructies en training die de beoordelaars t.b.v. het beoordelingsproces krijgen, zijn.	ja
3.10	Er is een rationale voor de toets waarin verantwoording wordt afgelegd over de samenstelling van het examen.	grotendeels
	Fase 4: Afname (n.v.t.)	

	Fase 5: Beoordeling, verwerking en analyse geschiedenis vmbo gl/tl 2021-1	Aanwezig?
5.1	Procedures voor het scoren en scorecriteria zijn gedocumenteerd.	ja
5.2	Het proces van het selecteren, trainen, kwalificeren en monitoren van degene die scoren is gedocumenteerd.	ja
5.3	Er zijn twee onafhankelijke beoordelaars die het examen apart van elkaar scoren.	ja
5.4	De twee onafhankelijke scores worden allebei gerapporteerd.	deels
5.5	Er is gedocumenteerd hoe de uiteindelijke score voor het examen wordt berekend.	ja
5.6	Op elk scoreniveau worden in de beoordelingsvoorschriften meerdere voorbeelden van antwoorden gegeven.	ja
5.7	Er wordt een Toets- en itemanalyse uitgevoerd onder voldoende kandidaten.	ja
5.8	De Toets- en Itemanalyse maakt de betrouwbaarheid van de toetscores zichtbaar met behulp van Cronbachs coëfficiënt alfa of een andere algemeen geaccepteerde maat voor betrouwbaarheid.	ja
5.9	De betrouwbaarheid die berekend is geeft geen aanleiding om te twijfelen aan de validiteit van de toetsscores.	ja
5.10	Indien sprake is van open vragen wordt via statistisch onderzoek de mate van beoordelaarsovereenstemming onderzocht.	deels

	Fase 6: Registreren en communiceren: n.v.t.	

	Fase 7: Evaluatie geschiedenis vmbo gl/tl 2021-1	Aanwezig?
7.1	Toetsspecificaties worden aangepast als er significante veranderingen zijn in het vertegenwoordigde domein.	ja
7.2	Bewijzen voor (inhouds)validiteit en betrouwbaarheid zijn vastgelegd.	ja
7.3	Bewijzen voor betrouwbaarheid (waaronder de mate van beoordelaarsovereenstemming) zijn vastgelegd.	deels
7.4	Procedures rondom toetsontwikkeling zijn vastgelegd.	ja
7.5	Rollen, verantwoordelijkheden en taakverdeling rondom toetsontwikkeling zijn vastgelegd.	ja
7.6	Er wordt onderzoek gedaan onder de stakeholders (belangenvertegenwoordigers van kandidaten en/of van opleiders) die zijn betrokken bij het gebruik van de toets.	ja
7.7	Er is een verslag (jaarrapport, evaluatieverslag) waarin verantwoording is afgelegd over de ontwikkeling, samenstelling, afname, beoordeling, scores en uitgevoerde analyses van de toets.	ja

Bijlage 3: Checklist Centraal examen economie havo 2021 eerste tijdvak

	Fase 1: Basisontwerp economie havo 2021-1	Aanwezig?
1.1	Er is een toetsplan.	ja
1.2	In het toetsplan is aangegeven wat de doelgroep van het examen is.	ja
1.3	In het toetsplan is aangegeven wat het meetdoel van het examen is oftewel hoe het domein gedefinieerd is.	ja
1.4	In het toetsplan is aangegeven wat het gebruiksdoel van het examen is.	ja
1.5	In het toetsplan is in duidelijke taal gespecificeerd in welke contexten toetsscores gebruikt moeten worden.	ja
1.6	In het toetsplan is in duidelijke taal gespecificeerd met welke processen (papieren toets of digitale toets, welke wijze van afname, welke wijze van beoordeling, hoe komt score tot stand) de toets moet worden afgenomen en gescoord.	ja
1.7	In het toetsplan staat de theorie met betrekking tot de beoogde interpretatie beschreven.	deels

	Fase 2: Opstellen toetsspecificaties en toetsmatrijs economie havo 2021-1	Aanwezig?
2.1	Er zijn toetsspecificaties (bv. in de vorm van een constructieopdracht).	ja
2.2	In de constructieopdracht staat het aantal vragen met bijbehorende scorepunten.	ja
2.3	In de constructieopdracht staat de toetsvorm en/of het soort vragen (bijvoorbeeld gesloten en/of open).	ja
2.4	In de constructieopdracht staat de toegestane hulpmiddelen.	ja
2.5	In de constructieopdracht staat de voorgestelde lengte van de toets.	ja
2.6	In de constructieopdracht staat de hoeveelheid tijd die is toegestaan voor de toets.	ja
2.7	In de constructieopdracht staan de gewenste psychometrische eigenschappen van de items en de toets als geheel.	deels
2.8	In de constructieopdracht wordt de volgorde van items gedefinieerd.	ja
2.9	Er is een toetsmatrijs.	ja
2.10	In de toetsmatrijs wordt de relevantie van de inhoud van de toets of het examen voor het beoogde meetdoel aannemelijk gemaakt.	deels
2.11	De toetsmatrijs is een adequate representatie van het meetdoel.	ja
	2.11a) De eind- en toetstermen representeren het meetdoel.	+
	2.11b) De eind- en toetstermen sluiten aan op de inhoud en het vereiste beheersingsniveau van het betreffende meetdoel.	+
	2.11c) Het werkwoordgebruik in de eind- en toetstermen is eenduidig en sluit goed aan bij de gebruikte taxonomie (bijvoorbeeld van Bloom of Romiszowsky).	+
	2.11d) Het aantal vragen geeft een voldoende dekking van het meetdoel.	+
2.12	Relevante vakinhoudelijke experts beoordelen de toetsspecificaties op hun geschiktheid.	ja
2.13	Het doel van deze beoordeling, het proces waarmee de beoordeling wordt uitgevoerd en de resultaten van de beoordeling worden gedocumenteerd.	ja
2.14	De kwalificaties, relevante ervaringen en demografische kenmerken van relevante vakinhoudelijke experts worden gedocumenteerd.	ja

	Fase 3 Constructie en normering economie havo 2021-1	Aanwezig?
3.1	De vragen zijn correct geformuleerd (zie richtlijnen in RCEC; die richtlijnen zijn hieronder ingevoegd als subcriteria):	ja
	3.1a) De vragen of opdrachten zijn gestandaardiseerd.	+
	3.1b) De vragen of opdrachten zijn zodanig ontworpen dat fouten bij invulling voorkomen worden.	+
	3.1c) De instructie voor de kandidaat (vaak eerste bladzijde van toetsboekje) is gestandaardiseerd, volledig en duidelijk.	+
	3.1d) De instructie voor de kandidaat dient minimaal te bevatten: aantal vragen, wijze waarop antwoorden worden gegeven, (deel) score per vraag of opdracht, de maximaal te behalen score en de cesuur; toegestane hulpmiddelen; beschikbare tijd en wat ingeleverd moet worden bij afronding; beoordelingspunten bij open vragen.	+/-
	3.1e) De kwaliteit van de lay-out en vormgeving is in orde.	+
3.2	De vragen zijn voldoende gedetailleerd, zodat examenkandidaten de vraag kunnen beantwoorden zoals bedoeld.	ja

	Fase 3 Constructie en normering economie havo 2021-1	Aanwezig?
3.3	De moeilijkheidsgraad van de vraag is afgestemd op de beoogde doelgroep.	ja
	3.3a) Er heeft een pretest plaatsgevonden om informatie te verkrijgen over de moeilijkheidsgraad (kwantitatieve evaluatie).	-
	3.3b) Er heeft een kwalitatieve evaluatie met deskundigen plaatsgevonden om informatie te verkrijgen over de moeilijkheidsgraad.	+
3.4	Er wordt een norm/cesuur verstrekt die via een wetenschappelijke methodiek (standaardbepaling) is bepaald.	ja
3.5	De manier waarop de norm wordt bepaald is gedocumenteerd.	ja
3.6	De standaardbepaling is correct uitgevoerd.	ja
	3.6a) De standaardbepalings-methode is op de juiste wijze uitgevoerd.	+
	3.6b) De vakdeskundigen/experts die bij de standaardbepaling betrokken zijn, zijn naar behoren geselecteerd en getraind.	+
	3.6c) Er is voldoende overeenstemming tussen de beoordelaars.	+/-
3.7	De procedures die worden gebruikt om items te ontwikkelen, beoordelen en uitproberen en om items uit de vragenbank te selecteren, zijn gedocumenteerd.	ja
3.8	Er is gedocumenteerd in hoeverre het inhoudsdomein van de toets het domein vertegenwoordigt, dat is gedefinieerd in de toets specificaties.	ja
3.9	Er is gedocumenteerd wat de kwalificaties, relevante ervaringen en demografische kenmerken, evenals de instructies en training die de beoordelaars t.b.v. het beoordelingsproces krijgen, zijn.	ja
3.10	Er is een rationale voor de toets waarin verantwoording wordt afgelegd over de samenstelling van het examen.	deels
	Fase 4: Afname (n.v.t.)	

	Fase 5: Beoordeling, verwerking en analyse economie havo 2021-1	Aanwezig?
5.1	Procedures voor het scoren en scorecriteria zijn gedocumenteerd.	ja
5.2	Het proces van het selecteren, trainen, kwalificeren en monitoren van degene die scoren is gedocumenteerd.	ja
5.3	Er zijn twee onafhankelijke beoordelaars die het examen apart van elkaar scoren.	ja
5.4	De twee onafhankelijke scores worden allebei gerapporteerd.	deels
5.5	Er is gedocumenteerd hoe de uiteindelijke score voor het examen wordt berekend.	ja
5.6	Op elk scoreniveau worden in de beoordelingsvoorschriften meerdere voorbeelden van antwoorden gegeven.	ja
5.7	Er wordt een Toets- en itemanalyse uitgevoerd onder voldoende kandidaten.	ja
5.8	De Toets- en Itemanalyse maakt de betrouwbaarheid van de toetscores zichtbaar met behulp van Cronbachs coëfficiënt alfa of een andere algemeen geaccepteerde maat voor betrouwbaarheid.	ja
5.9	De betrouwbaarheid die berekend is geeft geen aanleiding om te twijfelen aan de validiteit van de toetsscores.	deels
5.10	Indien sprake is van open vragen wordt via statistisch onderzoek de mate van beoordelaarsovereenstemming onderzocht.	deels

	Fase 6: Registreren en communiceren: n.v.t.	

	Fase 7: Evaluatie economie havo 2021-1	Aanwezig?
7.1	Toetsspecificaties worden aangepast als er significante veranderingen zijn in het vertegenwoordigde domein.	ja
7.2	Bewijzen voor (inhouds)validiteit en betrouwbaarheid zijn vastgelegd.	ja
7.3	Bewijzen voor betrouwbaarheid (waaronder de mate van beoordelaarsovereenstemming) zijn vastgelegd.	deels
7.4	Procedures rondom toetsontwikkeling zijn vastgelegd.	ja
7.5	Rollen, verantwoordelijkheden en taakverdeling rondom toetsontwikkeling zijn vastgelegd.	ja
7.6	Er wordt onderzoek gedaan onder de stakeholders (belangenvertegenwoordigers van kandidaten en/of van opleiders) die zijn betrokken bij het gebruik van de toets.	ja
7.7	Er is een verslag (jaarrapport, evaluatieverslag) waarin verantwoording is afgelegd over de ontwikkeling, samenstelling, afname, beoordeling, scores en uitgevoerde analyses van de toets.	deels