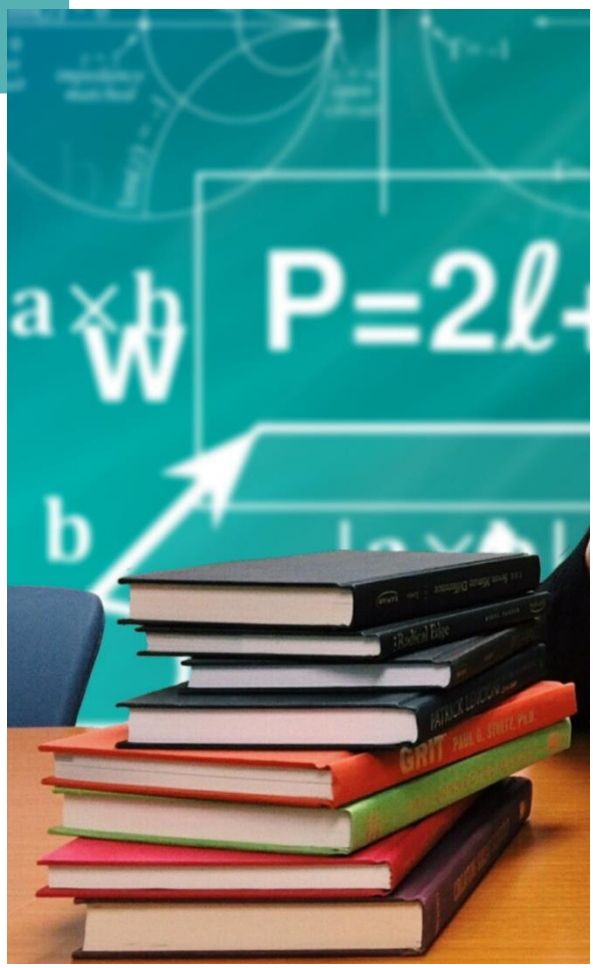
whitepaper

Beoordelen van Complexe Productieve Vaardigheden & Omzetten van Scores naar Cijfers



Stichting Cito

Onderzoek, Kennis, & Innovatie

Dit is een uitgave in de reeks Van Schoolexamens naar Diploma

© Stichting Cito, Arnhem– versie mei 2020

Voor wie is dit document bedoeld?

Vaak maken praktische vaardigheidstoetsen deel uit van de schoolexamens. De vraag is hoe je het werk van de leerlingen op deze toetsen zo objectief en efficiënt mogelijk kunt beoordelen. Met deze bijdrage willen we docenten handvatten bieden die hen kan helpen om de kwaliteit van de vaardigheidstoets zo goed mogelijk te krijgen.

Dit document is ook bedoeld voor alle docenten die een gefundeerde keuze slaag-zakgrens willen maken, en meer willen weten over de verschillende manieren waarop je de scores om kunt zetten in schoolcijfers. In het document staan ook verwijzingen naar handige online hulpmiddelen waarmee de scores naar cijfers om kunt zetten.

Heb je vragen over het schoolexamen of over één van onze uitgaves, of heb je een idee voor een onderwerp, neem dan gerust contact met ons op via **hulpbijexamens@cito.nl**.

© Stichting Cito, Arnhem– versie mei 2020

Wat kun je in dit document lezen?

Deel 1: Beoordelen van Complexe Productieve Vaardigheden

Deel 2: Omzetten van Scores naar Cijfers

© Stichting Cito, Arnhem– versie mei 2020

Deel I

Beoordelen van Complexe Productieve Vaardigheden

1 Inleiding

Productieve vaardigheden, en dan met name spreekvaardigheid en schrijfvaardigheid, zijn belangrijk vaardigheden die we onze leerlingen mee willen geven binnen het voortgezet onderwijs. Met behulp van praktische opdrachten kun je deze vaardigheden toetsen. Leerlingen schrijven bijvoorbeeld een betoog, houden een voordracht, bespreken een boek, of verdedigen een standpunt voor de klas. Aan de hand van het afgeleverde werk moet je dan als docent een oordeel vormen over de mate waarin de leerling over de betreffende vaardigheden beschikt.

Dit is geen eenvoudige taak! Leerlingen hebben vaak een redelijke grote mate van vrijheid om de opdracht naar eigen inzicht in te vullen. Het is dus vaak weer een verrassing wat de leerlingen produceren. Bovendien doen de opdrachten vaak beroep op meerdere vaardigheden tegelijk. Neem bijvoorbeeld het schrijven van een betoog. Om een goed betoog te kunnen schrijven moet een leerling niet alleen voldoende taalvaardig zijn, maar daarnaast ook logisch kunnen redeneren en analyseren. Vaardigheidstoetsen richten zich daarmee vaak op taken in het hogere cognitieve domein[1]. We spreken dan ook wel van **toetsen van complexe vaardigheden**.

Het beoordelen van de resultaten van een complexe vaardigheidstaak vraagt om zowel vakinhoudelijke kennis als kennis van beoordelingsprocessen (Sluijsmans & Kneyber, 2013). Je moet als docent niet alleen kunnen beoordelen wat inhoudelijk gezien goed en minder goed is, maar je moet je ook bewust zijn van mogelijke externe factoren die de beoordeling kunnen vertekenen en hoe je dit het beste kunt voorkomen. In dit eerste deel van het whitepaper zullen

[1] Een veel gebruikte taxonomie om vaardigheidsniveaus in te delen is die van Bloom. Hierin worden zes beheersingsniveaus onderscheiden: onthouden, begrijpen, toepassen, analyseren, evalueren en creëren..

© Stichting Cito, Arnhem– versie mei 2020

we laten zien die hoe je als docent zo goed mogelijk tot het een objectief beoordeling kunt komen. Dat doen we in een aantal stappen. We zullen eerst de belangrijkste criteria voor goede toetsing vanuit het oogpunt van vaardigheidstoetsen bekijken. In de praktijk probeer je de beoordelingstaak zo in te richten dat je zo goed mogelijk aan deze criteria kunt voldoen. Daarna zullen we een algemeen overzicht geven van verschillende beoordelingsmethodieken die worden gebruikt. We zullen in dit whitepaper niet op alle details in gaan, maar waar mogelijk zullen we referenties naar relevante literatuur geven. Tot slot zullen we dieper ingaan op een speciale methodiek waarvan onderzoek heeft laten zien dat het zeer toepasbaar is in de praktijk, dat is namelijk het gebruik van rubrics.

2 Waaraan Voldoet een Goede Vaardigheidstoets?

Net zoals alle andere toetsen dragen ook vaardigheidstoetsen uiteindelijk bij aan belangrijke beslissingen over de leerling. Denk daarbij aan beslissingen over het wel of niet laten overgaan en het wel of niet halen van het diploma. Het is daarom belangrijk dat de toetsing zo eerlijk en objectief mogelijk gebeurt. Om de kwaliteit van de toets te waarborgen worden doorgaans vier belangrijke criteria gehanteerd: inhoudsvaliditeit, betrouwbaarheid, transparantie, en fairness zie bijvoorbeeld Berkel en Bax, 2002). We zullen elk van deze criteria toelichten met het oog op vaardigheidstoetsen.

2.1 Inhoudsvaliditeit

Een goede toets begint met de juiste inhoud, ook wel aangeduid met inhoudsvaliditeit.[2] Wanneer een toets inhoudsvalide is dan wil dat zeggen dat de toets ook daadwerkelijk een beroep doet op de kennis en vaardigheden die je met de toets wilt meten. De leerdoelen en eindtermen zijn hierbij het uitgangspunt. Inhoudsvaliditeit van een toets moet je dan ook altijd zien in het licht van de gestelde leerdoelen.

Een veelgebruikt middel om de inhoudsvaliditeit te waarborgen is de toetsmatrijs. De toetsmatrijs beschrijft op een systematische manier de verdeling van de vragen over de verschillende leerdelen en beheersingsniveaus. Bij vaardigheidstoetsen ligt het gebruik van een toetsmatrijs uiteraard minder voor de hand omdat er sprake is van een omvangrijkere opdracht (bijvoorbeeld het schrijven van een opstel).. Toch is het belangrijk om al bij het maken van de opdracht zorgvuldig na te gaan of de opdracht voldoende beroep doet op te meten vaardigheden, Daarnaast moet je er op letten dat de opdrachten ook zo zijn opgesteld dat vaardigheidsverschillen tussen leerlingen zichtbaar worden. Dit noemen we ook wel het discriminerend vermogen van een vraag. Het maken van goede opdrachten - dat wil zeggen opdrachten die een beroep doen op relevante vaardigheden, de juiste moeilijkheidsgraad hebben, en ook nog eens onderscheid maken tussen studenten die de stof beheersen en studenten die de stof minder of niet beheersen – vraagt om vakmanschap en de kennis en ervaring van de docent is daarbij onmisbaar.

[2] Naast inhoudsvaliditeit zijn er ook nog andere vormen van validiteit zoals bijvoorbeeld begripsvaliditeit en criteriumvaliditeit. Deze vormen van validiteit zijn met name van belang bij psychologische toetsen.

© Stichting Cito, Arnhem– versie mei 2020

2.2 Betrouwbaarheid

Een toets is betrouwbaar als de resultaten niet worden beïnvloed door allerlei toevalligheden. Betrouwbaarheid is belangrijk omdat je graag wil dat als groep even vaardige leerlingen de toets maakt, dat zij dan ook allemaal dezelfde score hebben.
Wanneer echter toevalligheden een grote rol spelen, dan zullen de scores door toeval uiteen gaan lopen. De ene leerling heeft geluk en haalt een hoge score, terwijl de andere leerling met precies dezelfde vaardigheid misschien pech heeft. Wanneer toevalligheden een te grote rol gaan spelen dan kunnen we niet meer vertrouwen op de meting. We weten dan immers niet meer of een leerling een hoge score heeft puur door geluk, of omdat hij of zij gewoon heel vaardig. Het omgekeerde geldt natuurlijk ook: we weten dan niet of leerling pech had of de vaardigheid nog onvoldoende beheerst.

Wanneer bij de toetsing gebruik wordt gemaakt van beoordelaars is het nuttig om een onderscheid te maken tussen de betrouwbaarheid van een individuele beoordelaar en de overeenstemming tussen beoordelaars. Het eerste aspect, ook wel de *intra-beoordelaarsbetrouwbaarheid* genoemd, heeft betrekking op de mate waarin een individuele beoordelaar consistent is in zijn beoordelingen. De vraag die centraal staat is: zou je als docent tot dezelfde beoordelingen komen als je de werkstukken helemaal opnieuw zou beoordelen alsof je ze voor het eerst ziet?

Je zou denken dat wanneer docenten dezelfde werkstukken nog een keer zouden bekijken dat de beoordeling dan hetzelfde zal zijn, maar dat is helemaal niet evident. Beoordelaars laten zich mogelijk onbewust leiden door irrelevant toevallige omgevingsfactoren zoals hun gemoedstoestand, de volgorde waarin de werkstukken worden nagekeken, recente gebeurtenissen in de klas. Het kan ook komen doordat de beoordelingscriteria onduidelijk en ambigu zijn en daardoor niet consistent toegepast worden. Helaas is het nauwelijks mogelijk om voor een individuele beoordelaar vast te stellen hoe consistent hij of zij te werk gaat. Daarvoor zou je hem of haar daadwerkelijk werkstukken opnieuw moeten laten bekijken, onafhankelijk van de eerste keer. Dat is praktische gezien vaak onhaalbaar. Toch is het goed om je hiervan bewust te zijn en de beoordelingstaken zo in te richten dat je als beoordelaar gedwongen wordt om zo consistent mogelijk te werk gaan. Daarmee voorkom je dat de beoordeling niet afhangt van het toevallige moment waarop de beoordeling plaatsvindt.

De beoordelaarsovereenstemming gaat over de mate waarin *verschillende* beoordelaars tot eenzelfde beoordeling komen. [3] Wanneer de overeenstemming hoog is dan maakt het voor de leerling dus niet uit welke docent de beoordeling geeft. Onderzoek (Sanders, 2013) laat zien dat er grote verschillen in de beoordelingen kunnen bestaan *tussen* beoordelaars. [4] Dit betekent dat de ene beoordelaar (consistent) veel strenger is dan de ander. Dit soort factoren wil je natuurlijk zoveel mogelijk uitsluiten. Een objectieve en eerlijke beoordeling betekent dat de beoordeling niet moet afhangen van de toevallige docent die de beoordeling geeft. Een hoge

[3] In de literatuur wordt onderscheid gemaakt tussen relatieve en absolute overeenstemming. Van relatieve overeenstemming is sprake als elke beoordelaar de leerlingen in dezelfde volgorde zet, maar de scores hoeven niet perse overeen te komen. Van absolute overeenstemming is sprake als de beoordelaars ook tot dezelfde behoordeling komen. Bij afsluitende schoolexamens streeft men naar absolute overeenstemming.
[4] Zie SLO rapport over beoordelen van taalvaardigheid; en POK Memorandum "Het scoren van open vragen: beoordelen van beoordelaars" (Bechger en Maris, 2006) en Maris G., Bechger, T. M. (2004). *Het scoren van open vragen: Theorie en Praktijk.* POK memorandum 2004-2. Cito, Arnhem.

© Stichting Cito, Arnhem– versie mei 2020

overeenstemming bereik je overigens alleen als beoordelaars ook consistent zijn. De beoordelaarsovereenstemming kan goed worden onderzocht door leerlingen door meerdere willekeurig gekozen docenten te laten beoordelen (Heuvelmans & Sanders, 1993). Met behulp van statistische modellen kunnen we nagaan hoe groot de overeenstemming is en ook kunnen beoordelaars identificeren die erg afwijken (Bechger & Maris, 2006).

Helaas spelen toevalligheden altijd een rol, en zullen er altijd verschillen tussen docenten bestaan. Met een aantal eenvoudige maatregelen kun je de betrouwbaarheid en overeenstemming wel zo hoog mogelijk te maken. Je kunt bijvoorbeeld twee (of meer) beoordelaars inschakelen, in ieder geval bij leerlingen die op de grens zitten tussen zakken en slagen. De gemiddelde beoordeling geeft een betrouwbaarder beeld dan de individuele beoordelingen. Een andere maatregel is het gebruik van zorgvuldig geconstrueerde beoordelingsinstrumenten voor de beoordeling. We zullen zien dat met name analytische instrumenten, waarbij de beoordeling plaatsvindt aan de hand van verschillende objectieve criteria, in hoge mate bijdragen aan de betrouwbaarheid en overeenstemming.

2.3 Transparantie

De toetsing is transparant als het voor de leerling duidelijk is hoe de beoordeling tot stand is gekomen. Dit betekent onder andere dat het voor een leerling duidelijk moet zijn welke criteria zijn gebruikt, hoe de criteria zijn meegewogen in de eindbeoordeling, en hoe het eindcijfer tot stand is gekomen. Het gaat hierbij vooral om *transparantie* over de toetsingsprocedure en *uitleg* bij de uitslag. Het is goed om te benadrukken dat bij summatieve toetsing de uitleg vooral inzicht moet geven in hoe de beoordeling tot stand is gekomen, maar het hoeft geen uitgebreide feedback te zijn. Dat is anders bij formatieve toetsing waarbij de toetsing wordt gebruikt om het leertraject verder vorm te geven. Het geven van gerichte feedback is daarbij een belangrijke bepalende factor. [5]

2.4 Fairness

Het begrip *fairness* kun je het beste interpreteren als “onpartijdigheid”. Strikt genomen spreken we van een onpartijdige toets als persoonlijke kenmerken van leerlingen die irrelevant zijn voor de te meten vaardigheid geen invloed hebben op de prestaties of de uitslag (COTAN addendum 2015 [6]). Met andere woorden, bepaalde groepen leerlingen mogen door de toetsing niet systematisch worden benadeeld. Je kunt dan denken aan jongens en meisjes, maar ook om culturele verschillen. Fairness begint bij een neutrale vraagstelling. Vragen die specifiek aansluiten bij de belevingswereld van een bepaalde selectieve groep kunnen nadelig uitpakken voor andere leerlingen. Ter illustratie: Laat je leerlingen voor het vak Frans een essay schrijven

[5] Dit geldt met name voor summatieve toetsing, waarbij de uitleg vooral inzicht moet geven in hoe de beoordeling tot stand is gekomen. Dat is anders bij formatieve toetsing, waarbij de toetsen wordt gebruikt om het (individuele) leertraject verder vorm te geven. Het geven van gerichte feedback is daarbij een belangrijke bepalende factor.

[6] https://www.psynip.nl/wp-content/uploads/2019/01/Fairness-addendum-per-01-07-15-def_met-logo-en-liggende-paginas.pdf

© Stichting Cito, Arnhem– versie mei 2020

over het belang van mode in de Franse cultuur, dan is dat voor meisjes gemiddeld genomen misschien gemakkelijker dan voor jongens.

Eén belangrijke bedreiger van onder andere fairness is het zogenaamde halo-effect, waarbij een beoordelaar zich laat leiden op basis van prestaties op andere vakken of op slechts één (uitspringend) kenmerk. Halo-effecten kunnen bij onderwijskundige toetsen om verschillende redenen optreden. Een beoordelaar kan zich op basis van de eerste uitingen een indruk vormen van de vaardigheid van een kandidaat om op basis van die indruk de overige beoordelingen te doen zonder nog verder te luisteren naar wat de kandidaat zegt. Halo-effecten treden ook op als aan bepaalde groepen leerlingen (bijv. rustige leerlingen, leerlingen met een bepaald voorkomen) a priori hogere vaardigheden worden toegedicht. Het risico van halo-effecten neemt over het algemeen toe naarmate de docent de leerling beter kent. Het is ook om die reden raadzaam om als het kan een tweede beoordelaar in te schakelen en dan bij voorkeur een collega die wat verder op afstand van de leerling staat.

Het is belangrijk om te benadrukken dat bovengenoemde criteria niet los van elkaar gezien kunnen worden. Zo is een onbetrouwbare toets per definitie niet valide. Even zo geldt dat een partijdige toets zorgt voor een lagere overeenstemming en lage validiteit. Je moet de standaarden dan ook altijd in onderling verband bekijken.

3 Beoordelingsinstrumenten

Een belangrijk hulpmiddel bij het (summatief) beoordelen van productieve vaardigheden zijn beoordelingsinstrumenten. Dit kan een eenvoudige checklist zijn waarop je als docent voor een aantal aspecten aangeeft of het werk voldoende of onvoldoende is. Het kan ook een uitgebreid formulier zijn waarin gedetailleerd staat omschreven hoe het werk beoordeeld moet worden. Welke vorm het beste werkt hangt af van verschillende factoren zoals de complexiteit van de opdracht en het doel van de toetsing. Je kunt gebruik maken van bestaande instrumenten, maar vaak is het beter nog om je eigen beoordelingsformulier te ontwerpen waarmee je het naadloos aansluiten op het gegeven onderwijs." (Sanders, 2013, hoofdstuk 8, p. 11). De uitwerking van Baack (2018) is daarvan een mooi voorbeeld.

In deze paragraaf zullen we een aantal praktische tips geven over hoe je een beoordelingsinstrument in elkaar kunt zetten. Welke keuzes heb je? Welke keuze past het best bij welke vorm? Om de tips goed toe te kunnen passen is het zinvol om eerst kort stil te staan bij de vraag: Wat is de rol van de beoordelaar bij schoolexamens? Daarna zullen we enkele varianten van instrumenten bespreken en laten we zien hoe de instrumenten de beoordelaar kunnen ondersteunen. Tot slot zullen we maken van rubrics (Ragupathi & Lee, 2020) toelichten.

© Stichting Cito, Arnhem– versie mei 2020

4 De Rol van de Beoordelaar

Beoordelen is het toekennen van een waardering –verbaal of met een cijfer – aan een geleverde prestatie. In die zin is elke toets en schoolexamen een vorm van beoordelen. Kenmerkend aan beoordelen in het onderwijs is dat je in feit twee moeten doen. Allereerst moet je bepalen wat de juistheid, volledigheid en kwaliteit van het geleverde werk is. Een simpel voorbeeld is nagaan of de leerling alle werkwoorden goed gespeld heeft. Als tweede moet je bepalen of het afgeleverde werk ook *voldoende* is. Het gaat hierbij om het geven van een normatief oordeel. Terugkomend op het voorbeeld over spelling: Is het aantal spelfouten nog acceptabel, of worden er zoveel spelfouten gemaakt dat je een onvoldoende moet geven?

Bij het corrigeren van meerkeuzetoetsen is het onderscheid tussen de twee beoordelingstaken duidelijk. Eerst kijk je welke vragen de leerling goed heeft beantwoord. Vervolgens zet je de score om in een schoolcijfer waarbij je rekening houdt met de slaag-zakgrens. Als de leerling evenveel of meer punten heeft dan de slaag-zakgrens dan is het cijfer minimaal een 5,5, en anders is het cijfer lager. Doordat je de cijfers afstemt op de slaag-zakgrens is er sprake van een normatief oordeel op het moment dat je cijfers toekent.

Bij beoordelen van complexe vaardigheden lopen de twee beoordelingstaken vaak door elkaar. Neem bijvoorbeeld de situatie waarbij de docent direct aangeeft of hij of zij een mondeling voordracht *onvoldoende, voldoende* of *goed* vond. In dit geval wordt direct een normatief oordeel gegeven maar blijft de onderbouwing ervan impliciet. De leerling ziet dan niet *wat* de docent beoordeeld heeft, wat hij of zij anders had kunnen of moeten doen, en *hoe* het normatieve oordeel tot stand is gekomen. Nu is dit misschien een wat gekunsteld voorbeeld omdat je als docent de beoordeling doorgaans wel motiveert, maar als dit achteraf gebeurt dan bestaat het gevaar dat de motivatie wordt afgestemd op het al gegeven oordeel.. De motivatie is dan het resultaat van het cijfer, in plaats van andersom. Het is beter om de beoordelingstaak gestructureerd aan te pakken en beide beoordelingstaken – namelijk het bepalen van de kwaliteit van het werk, en het geven van een normatief oordeel – zoveel mogelijk los te koppelen.. Dit kan heel goed met gestructureerde beoordelingsinstrumenten. Onderzoek heeft laten zien dat de kwaliteit van de beoordeling daarmee beter gewaarborgd kan worden (Sanders, 2017).

© Stichting Cito, Arnhem– versie mei 2020

5 Beoordelingsinstrumenten in de Praktijk

Wanneer je zelf aan de slag gaat met het maken van een beoordelingsmodel, of als je een bestaand instrument wilt selecteren, dan zul je een aantal keuzes moeten maken. Hieronder beschrijven we enkele van die keuzes.

5.1 Specifiek versus generiek

Een beoordelingsmodel kan specifiek gemaakt zijn voor een bepaalde opdracht. Dit zien we vaak bij praktische vaardigheidstoetsen waarbij de kwaliteit van het geleverde werk wordt bepaald door specifieke gedragingen. Neem bijvoorbeeld het toetsen van verkoopvaardigheden. Hierbij kijk je specifiek naar de manier waarop de leerling de klant aanspreekt, het product aanprijst, vragen beantwoord etc. Het nadeel van specifieke instrumenten is natuurlijk de beperkte toepasbaarheid.

Generieke beoordelingsschalen zijn minder toegespitst op de specifieke opdracht maar richten zich meer op algemene kenmerken en vaardigheden. Je zou bijvoorbeeld kunnen denken aan het beoordelen van onderzoeksverslagen bij exacte vakken. Met behulp van een generiek formulier ga je na of de leerling de onderzoeksvraag juist heeft geformuleerd, de juiste achtergrondinformatie heeft verzameld, de resultaten begrijpelijk heeft samengevat, etc. Het voordeel van generieke instrumenten is dat je ze bij meerdere vakken kunt gebruiken (Sanders, 2013). Leerlingen raken bekend met beoordelingsprocedure en dit verhoogt de transparantie. Generieke schalen hebben ook nadelen. De beoordelingscriteria sluiten minder goed aan bij de te beoordelen taak. Bij generieke schalen zie je dat de aandacht al snel op het normatieve oordeel gericht is, en hoe minder op de inhoudelijke beoordeling van de kwaliteit van het werk. Bij het gebruik van generieke schalen is het dan ook goed om extra aandacht te besteden aan de validiteit. Dus je moet je steeds blijven afvragen: wat meet ik en meet ik het goede?

5.2 Holistische versus analytisch

Bij een holistische beoordeling geeft de docent een algemeen oordeel. Dit kan een verbale beschrijving van het werk zijn (“terugkoppeling”) zonder dat er gebruik wordt gemaakt van een vaste structuur. Holistische beoordelingen zien we vaak bij creatieve opdrachten. Een bekende vorm van een holistische beoordeling is de recensie. Holistische beoordelingen geven veel vrijheid aan de beoordelaar. Dat is niet altijd persé verkeerd, maar over het algemeen wordt deze methode afgeraden. De holistische benadering is erg gevoelig voor allerlei vormen van bias zoals bijvoorbeeld bevestigingsbias en selectieve waarneming. Holistische beoordelingen zijn daarom maar in zeer beperkte mate toepasbaar bij schoolexamens.

Analytische schalen delen de te vaardigheidstoets op in verschillende deelvaardigheden of deeltaken die relevant zijn voor de kwaliteit van het werk. Per deelvaardigheid of taak wordt gekeken hoe de leerling presteert. Een analytische beoordelingsschaal bestaat dus uit een aantal subschalen op basis waarvan volgens een vooraf opgestelde procedure een eindoordeel wordt bepaald. Analytische schalen zijn bij uitstek geschikt voor beoordelen van prestaties op complexe vaardigheden zoals mondeling presenteren en taalgebruik (Sanders, 2017). Het draagt bij aan alle kwaliteitscriteria voor goede toetsing.

© Stichting Cito, Arnhem– versie mei 2020

5.3.Beoordelingsschalen: Normatief, beschrijvend, of productgericht

Uiteindelijk moet je als docent aangeven hoe de leerling gepresteerd heeft. Dit betekent dat je de leerling op een beheersingsschaal moet plaatsen. Bij analytische beoordelingen doe je dat per deelaspect. Als je een schaal wilt construeren moet je twee vragen beantwoorden: De eerste vraag is hoe verfijnd wil je een onderscheid kunnen maken? Dus wat is het aantal schaalpunten dat je wil hanteren. De tweede vraag betreft het duiden van de schaalpunten. We zullen beide vragen toelichten.

Keuze aantal schaalpunten. Wanneer je een tweepuntschaal hanteert, dan deel je de leerlingen in feite in twee vaardigheidsniveaus. Soms is dat voldoende, maar vaak wil je een verfijnder onderscheid maken. Wat het ideale aantal is, is echter lastig te zeggen. Als je weinig schaalpunten hanteert dan bestaat het gevaar dat je leerlingen niet meer goed onderscheidt. Heel veel schaalpunten is dan weer onpraktisch en veel gevoeliger voor toevalligheden. Hierdoor ontstaat zekere mate van schijnnauwkeurigheid. In het algemeen wordt aangeraden om twee tot zeven schaalpunten te gebruiken.

Schaalpunten benoemen. Stel je hanteert een tweepuntschaal, betekent dat dan "voldoende / onvoldoende", of is er sprake van "voldaan/niet voldaan." De wijze waarop de schaalpunten worden beschreven bepaalt het type schaal. Er zijn globaal drie typen schalen te onderscheiden: normatieve schalen, beschrijvende schalen, en productschalen. Om met het laatste te beginnen. Bij productschalen gebruik je concrete producten (bijv. een oude werkstukken van leerlingen) om de verschillende schaalpunten te duiden. De vraag is dan welke voorbeeldproduct het beste de kwaliteit van de leerling representeert. Dat bepaalt welke beoordeling gegeven wordt. Productschalen worden in de praktijk maar weinig gebruikt en laten we verder buiten beschouwing.

Bij normatieve schalen worden de schaalpunten beschreven in normatieve termen. Een voorbeeld van een normatieve schaal is een driepuntschaal met schaalpunten: "onvoldoende", "voldoende", en "goed". Een normatieve schaal geeft dus niet alleen hoe goed het werk is, maar ook of het voldoet aan de gestelde eisen. Normatieve schalen zijn met name handig bij het beoordelen van eenvoudigere taken die waarbij meteen duidelijk is of de taak wel of niet goed uitgevoerd wordt. Normatieve schalen zien we ook vaker bij meer generieke beoordelingsinstrumenten.

Een belangrijke overweging bij het construeren van normatieve schalen is de balans tussen positieve en negatieve kwalificaties. Bij een gebalanceerde schalen zijn er evenveel negatieve als positief geformuleerde schaalpunten. Het gebruik van balanceerde schalen heeft een aantal nadelen. Uit onderzoek weten we dat beoordelaars regelmatig bewust of onbewust de neiging hebben negatieve beoordelingen te vermijden. Er is sprake van een zeker mildheid. Het gevolg is dat bepaalde schaalpunten nooit gebruikt zullen worden. Je kunt er als docent ook voor kiezen om een niet-gebalanceerde schaal te gebruiken, zoals bijvoorbeeld: "onvoldoende, voldoende, ruim voldoende, goed, zeer goed, uitmuntend". Onderzoek heeft laten zien dat niet-gebalanceerde schalen minder snel leiden tot "mildheid". Bovendien kun je dan als docent veel beter een onderscheid maken tussen de "betere|" leerlingen. Dat zorgt voor meer spreiding in de scores (Kuhlemeier, Hemker, & Bergh, 2013).

© Stichting Cito, Arnhem– versie mei 2020

Beschrijvende schalen, geven bij elk schaalpunt, of bij een gericht gekozen selectie van schaalpunten een beschrijving van de kenmerken die passen bij het schaalpunt. Hieronder vind je een voorbeeld van een beschrijvende beoordelingsschaal voor het beoordelen van beheersing van de Nederlandse taal met drie schaalpunten.

	Schaalpunten		
	1	2	3
Beschrijving	De leerling maakt structureel d/t fouten. Samengestelde woorden worden vaak los van elkaar geschreven.	Leerling maakt alleen veelvoorkomende d/t fouten ("hij mag weg als hij de boete betaald"). Samengestelde woorden merendeel aaneen geschreven.	De leerling maakt geen d/t fouten. Alle samengestelde woorden worden aaneengeschreven.

Kenmerkend aan de beschrijvende schalen is dat er nog geen normatief oordeel wordt gegeven. Het enige wat je probeert te doen is de leerling aan de hand van observeerbare kenmerken van het ingeleverde werk op een schaal te plaatsen die aangeeft wat de inhoudelijke kwaliteit van het werk is. De vraag of dit ook "voldoende" is wordt pas later beantwoord. Er is dus een expliciet onderscheid tussen het inhoudelijke oordeel en het normatieve oordeel..

Het construeren van beschrijvende schalen is vaak erg bewerkelijk. Het is bovendien vaak moeilijk om de juiste balans te vinden in de mate van detail waarin je de schaalpunten beschrijft. Dit voor is veel minder het geval voor normatieve schalen. Hoewel normatieve schalen (nog) redelijk vaak voorkomen wordt het gebruik er van afgeraden (Sanders 2013). Het leidt tot subjectieve oordelen en het maakt de beoordelingsprocedure gevoelig voor beoordelaarseffecten. Beschrijvende schalen zijn daarmee minder ambigu, bieden betere ondersteuning bij de beoordelingstaak, en zijn minder gevoelig voor beoordelingsfouten dan algemene (normatieve) beoordelingsschalen (Kuhlemeier, Hemker & Bergh, 2013).

© Stichting Cito, Arnhem– versie mei 2020

6 Beoordelen met Rubrics

Beschrijvende schalen worden in de onderwijskundige literatuur ook wel rubrics genoemd. In deze paragraaf zullen we een stappenplan presenteren waarmee je gemakkelijk je eigen rubrics kunt samenstellen. Rubrics bevatten tenminste de volgende drie basiselementen (Ragupathi & Lee, 2020):

1. Een beschrijving van de factoren (vaardigheden, deelvaardigheden, leerdoelen) die relevant zijn voor de kwaliteit van het geleverde werk.
2. Een scoreschaal met schaalpunten die aangeven welke beheersingsniveaus te behalen zijn.
3. Een beschrijving de gewenste kennis of gedrag die bij de verschillende beheersingsniveaus horen.

Hieronder vind je de schematische lay-out van een rubric:

Schematische opzet van een rubric.

	Schaalpunten		
	(1)	(2)	(3)
Beschrijving deelvaardigheid + eventueel korte toelichting op bepalende factoren	Beschrijving van concrete waarneembare kenmerken voor een leerling functionerend op niveau 1	Beschrijving van concrete waarneembare kenmerken voor een leerling functionerend op niveau 2	Beschrijving van concrete waarneembare kenmerken voor een leerling functionerend op niveau 3

De uitwerking van elke van deze elementen kan op verschillende manieren en in verschillende mate van detail [7]. In de volgende paragraaf zullen we laten zien hoe je in drie stappen tot een beoordelingsinstrument met rubrics kunt komen.

Stap 1: Beschrijf de deelvaardigheden die mee moeten wegen in de beoordeling.

De eerste vraag die beantwoord moet worden is: welke (deel)vaardigheden en leerdoelen moeten meewegen in de beoordeling en waaraan kun je zien op welk niveau een leerling het leerdoel heeft behaald of de vaardigheid beheerst? Het uitgangspunt bij summatieve toetsing is dat er duidelijk omschreven eindtermen zijn op basis waarvan de toets of taken zijn samengesteld. De eindtermen kun je gebruiken om aan te geven welke factoren bepalend zijn voor de kwaliteit. Neem bijvoorbeeld het schrijven van een betoog voor Nederlands. Hierbij draait het om een aantal deelvaardigheden: Nederlandse taalvaardigheid, argumentatietechnieken, en tekstverzorging. Bij een mondelinge voordracht draait het naast taalvaardigheid ook om presentatie, begrijpelijkheid, etc. (Baack, 2008). Per deelvaardigheid geef je waaraan je kunt zien

[7] Sanders, 2013

© Stichting Cito, Arnhem– versie mei 2020

op welk niveau de leerling functioneert. Bij taalvaardigheid kun je bijvoorbeeld denken aan correcte spelling, correct gebruik van verwijswoorden, interpunctie, etc.

Bij complexe productieve vaardigheden is de neiging vaak groot om veel verschillende deelvaardigheden en factoren te onderscheiden. Het is echter aan te raden om je beperken tot maximaal vijf vaardigheden. Dat zorg ervoor dat het beoordelingsinstrument hanteerbaar blijft. Wanneer je veel deelvaardigheden onderscheidt is de kans ook groot dat er veel overlap is. Het afzonderlijk beoordelen van die factoren voegt dan weinig toe aan het eindoordeel.[8] Bovendien wil je voorkomen dat leerlingen dubbel gestraft worden. Soms is het handig om deze stap in twee rondes te doen. In de eerste rond (brainstormfase) beschrijf je alle deelvaardigheden en factoren waarvan je denkt dat die relevant zijn. In de tweede rond probeer je clusters te maken van samenhangende factoren die die samen te voegen tot een beperkt aantal.

Stap 2: Kies het aantal schaalpunten

In deze stap bepaal je het aantal niveaus waarop je de leerlingen op elke factor wilt onderscheiden. Dit vormt de basis voor de scoreschaal. De meest eenvoudige variant is een schaal met twee schaalpunten. Een dergelijke schaal kan bruikbaar zijn bij eenvoudige vaardigheden die een leerling wel of niet beheerst. Een dergelijke schaal is minder goed geschikt voor vaardigheden waarbij veel meer differentiatie in de beheersing mogelijk is. Als je dan toch kiest voor een tweepuntschaal dan ervaren leerlingen dat mogelijk als unfair omdat daarmee voor hun gevoel hun inzet niet altijd wordt beloond. Vaak is het beter om meerdere schaalpunten te kiezen. Het is wel belangrijk dat de schaalpunten ook beschreven kunnen worden. Naarmate je meer schaalpunten kiest wordt dat steeds lastiger.

Stap 3: Beschrijving van de beheersingsniveaus

In de laatste stap worden de schaalpunten gekoppeld aan inhoudelijke duiding van de prestaties. De vraag die in deze stap beantwoord moet worden is: hoe zou je de verschillende vaardigheidsniveaus typeren? (Sanders, 2013, toetsen op school). Het gaat hierbij om een zo concreet mogelijke beschrijving van herkenbare elementen in het werk van een leerling waaraan je kunt zien op welk niveau hij of zij de stof beheerst. In feite is dit de verdere concrete uitwerking van de factoren die je in de eerste stap hebt geformuleerd.

Een aantal punten waar je op moet letten zijn:

- Hanteer korte beschrijvingen en beschrijf alleen observeerbare eigenschappen. Neem het beoordelen van een mondelinge presentatie. Minder goed is bijvoorbeeld: “de leerling voelt de klas aan”; beter is “de leerling spreekt in begrijpelijk woorden” of “de leerling vraagt bij moeilijke begrippen of de uitleg duidelijk is”. Zorg ervoor dat de beschrijvingen niet overlappen.

[8] Bijvoorbeeld: “woordenschat” en “gebruik van synoniemen”.

© Stichting Cito, Arnhem– versie mei 2020

- Vermijd normatieve taal (bijv. “slecht”, “redelijk”), zeker voor lagere niveaus. Zoals gezegd laat onderzoek zien dat beoordelaars het soms lastig vinden om negatieve oordelen te geven. Er is dan sprake van mildheid. Dit kan verschillende oorzaken hebben, waaronder de natuurlijke neiging om positief te oordelen, maar ook kunnen externe factoren (beeldvorming op school, reputatie) een rol spelen. Het gevolg is dat beoordelingen misschien positiever uitpakken dan bedoeld.

- Zorg er voor dat de schaal en de bijbehorende beschrijvingen het hele spectrum van de vaardigheid dekken. Hiermee voorkom je zogenaamde *restriction of range*, waardoor alle scores dicht bij elkaar komen te liggen. De toets verliest daarmee het discriminerend vermogen en over het algemeen neemt de betrouwbaarheid af.

- Zorg er voor dat alle beschrijvingen dezelfde kwaliteit beschrijven. Bij een mondeling examen is het dus onjuist om lage niveaus in termen van woordgebruik en hoge niveaus in termen van consistentie van het verhaal te beschrijven.

Een belangrijke afweging is hoe gedetailleerd je de beschrijvingen maakt. Een gedetailleerd model geeft de ene docent houvast, maar wordt door een andere docent misschien als te beperkt en rigide ervaren. Een globale beschrijving geeft de docent meer ruimte, maar wordt door sommige docenten misschien weer als te vaag en ambigue ervaren. In de praktijk is het zoeken naar modellen met een juiste balans in gedetailleerdheid en praktische toepasbaarheid.

© Stichting Cito, Arnhem– versie mei 2020

7 Twee Voorbeelden van Beoordelingsschalen

Beoordelings-aspect	Beheersingsniveau									
	"Laag"			"Midden"				"Hoog"		
Score =	1	2	3		5	6	7	8	9	10
Grammatica en stijl	Veel korte zinnen; veelvuldig foutieve vervoegingen; verkeerd gebruik van verwijswoorden			Spreekt met zinnen van gemiddelde complexiteit; maakt geen basale taalfouten; hanteert over het algemeen de juiste vervoegingen				Correct samengestelde zinnen; correct gebruik van vervoegingen, verwijswoorden		
Aantekeningen Tijdens het gesprek										
Score =	1	2	3	4	5	6	7	8	9	10
Woordenschat	Gebruikt veel eenvoudige woorden; weinig gebruik van synoniemen en antoniemen (niveau < 2F).			Woordenschat in overeenstemming met niveau 2F-3F .				Rijk taalgebruik; correct gebruik van gezegden; (niveau 3F-4F)		
Aantekeningen Tijdens het gesprek										

Beoordelings-aspect	Beheersingsniveau									
	"Laag"			"Midden"				"Hoog"		
Score =	1	2	3	4	5	6	7	8	9	10
Samenhang Tekst	De alinea's zijn onlogische geordend; geen gebruik van structuurwoorden; belangrijke elementen in argumentatie ontbreken			Logische opbouw argumentatie; gemiddeld gebruik van structuurwoorden; correct gebruik van alinea's				Logische verhaal; veelvuldig gebruik van structuur woorden; elke alinea behandeld een deel onderwerp		
Aantekeningen										
Score =	1	2	3	4	5	6	7	8	9	10
Taalverzorging	Tekst bevat veel basale taal-type en/of spelfouten; foutief gebruik van interpunctie; slordige opmaak			Incidentele taalfouten; bevat voornamelijk veel voorkomende taalfouten; basisregels voor interpunctie zijn juist toegepast; nette opmaak				Tekst bevat nauwelijks taal, spel, of typefouten. Juist gebruik van interpunctie; overzichtelijke lay-out.		
Aantekeningen										

© Stichting Cito, Arnhem– versie mei 2020

Deel II

Het Omzetten van Scores naar Schoolcijfers

8 Inleiding

In Nederland is het gebruikelijk om schoolprestaties uit te drukken in een (school)cijfer op een schaal van 1 tot 10, waarbij men vaak op één decimaal nauwkeurig rapporteert. Het cijfer 5,5 markeert de grens tussen onvoldoende en voldoende. Hogere cijfers duiden op een hogere mate van kennis en vaardigheden waar de toets een beroep op doet. Schoolcijfers hebben een min-of-meer vaste betekenis, los van het vak, en ook los van de specifieke toets waarop het cijfer gebaseerd is. Wanneer leerling een 8 heeft gehaald voor Engels, dan wordt dat gezien als bewijs dat de leerling "goed" in Engels is. Cijfers zeggen dus niet alleen iets over de prestatie op de specifieke toets zelf, maar geven een beeld van iemands vaardigheidsniveau in het algemeen.[9] [10]

In het geval van school- en eindexamens is de keuze voor de slaag-zakgrens erg belangrijk. Ten eerste bepaalt deze keuze wie wel en wie niet slaagt. Ten tweede bepaalt de keuze voor de slaag-zakgrens voor een groot deel hoe de omzetting van scores naar een schoolcijfer. Dit betekent dat zelfs kleine verschuivingen in de slaag-zakgrens aanzienlijke gevolgen kunnen hebben voor een belangrijk deel van de leerlingen. Niet alleen of men zakt of slaagt, maar ook het uiteindelijke schoolcijfer dat men krijgt. De slaag-zakgrens dient dan ook systematisch, zorgvuldig en weloverwogen gekozen te worden. In dit deel van het whitepaper zullen we enkele praktische hulpmiddelen bespreken die je kunt gebruiken om de slaag-zak grens te bepalen of te verfijnen.

[9] Neem het voorbeeld van een rekentoets. Wanneer iemand relatief veel punten haalt, dan zeggen we "hij heeft de toets goed gemaakt, dus hij is goed in rekenen".
[10] Hoe cijfers, waaronder schoolcijfers, een prominente rol in onze maatschappij hebben wordt onder andere mooi beschreven in "Het best verkochte boek" van Sanne Blauw.

© Stichting Cito, Arnhem– versie mei 2020

9 Waarom Geven we Cijfers?

Hoewel het geven van schoolcijfers inmiddels zo vanzelfsprekend is geworden[11], is het toch interessant om even stil te staan bij de vraag waarom we werken met schoolcijfers. In eerste instantie levert een toets een *ruwe score* op. Dat zijn punten die zijn toegekend aan correcte elementen in het antwoord. Voorbeeld van een ruwe score is het aantal correcte antwoorden op een toets met meerkeuzevragen. Ruwe scores hebben zonder context echter geen enkele betekenis. Het feit dat Marieke 14 vragen goed had vertelt ons alleen dat Marieke 14 vragen goed had en verder niets. Maar als je weet dat de toets uit 15 vragen bestond, 90% van de andere leerlingen lager gescoord heeft dan Marieke, en de docent 6 vragen of meer goed als "voldoende" beschouwde, dan weet je dat zij zeer goed gepresteerd heeft op de toets, zowel relatief ten opzichte van andere leerlingen als ook ten opzichte van de inhoudelijke norm.

Contextuele informatie is dus essentieel om de prestatie van leerlingen, en zoals die van Marieke, zinvol te kunnen interpreteren. Het plaatsen van prestaties in een context is in feite wat er gebeurt als je de ruwe toets-scores omzet naar schoolcijfers. Het toekennen van betekenis aan scores wordt ook wel normeren genoemd. Schoolcijfers zijn dus genormeerde scores, waaraan we in Nederland een eenduidige inhoudelijke betekenis toekennen: [12]:

Cijfer	Omschrijving		Cijfer	Omschrijving
10	uitstekend		5	twijfelachtig / zwak
9	zeer goed		4	onvoldoende
8	goed		3	ruim onvoldoende
7	ruim voldoende		2	slecht
6	voldoende		1	zeer slecht

In het geval van Marieke zouden we het cijfer 9,5 geven, waarmee voor iedereen duidelijk is – ook al hebben zij geen idee van de toets zelf – dat zij zeer goed tot uitmuntend gepresteerd heeft. Schoolcijfers spelen dus niet alleen een belangrijke rol binnen het Nederlandse onderwijssysteem, maar in zekere zin ook veel breder in onze maatschappij.

[11] Zie Dane (2014) voor een beknopt historisch overzicht
[12] Bron: https://nl.wikipedia.org/wiki/Schoolcijfer

© Stichting Cito, Arnhem– versie mei 2020

10 Vaststellen van de Slaag-Zakgrens

Het vaststellen van de slaag-zakgrens begint met een inhoudelijke omschrijving van wat (inhoudelijke) experts vinden dat een leerling moet kunnen om een bepaald vak goed uit te kunnen oefenen of om een diploma toegekend te krijgen. In het VO zijn deze eisen bijvoorbeeld vastgelegd in eindtermen. De eindtermen geven een inhoudelijke beschrijving van de prestatiestandaard weer. De vervolgvraag is dan: hoe vertaal je de eisen zoals die worden omschreven in de eindtermen naar een slaag-zakgrens op een bepaalde specifieke toets?

In de praktijk probeer je de school- en eindexamens zo samen te stellen dat alle relevante leerdoelen worden afgedekt. Daarbij probeer je zoveel mogelijk vragen te stellen, en streef je er naar een toets van "gemiddelde" moeilijkheid is door zowel moeilijkere als makkelijkere vragen in de toets op te nemen. [13] Een simpele regel om de slaag-zakgrens vast te stellen is dan, bijvoorbeeld, "55% van totaal aantal punten = geslaagd". Dit is een inzichtelijke regel, makkelijk toe te passen en uit te leggen, en het sluit aan bij de intuïtie dat je bij een toets met een gemiddelde moeilijkheid iets meer dan de helft van het totaal aantal punten moet hebben om de toets te halen. [14] Feitelijk zeg je hiermee ook dat je verwacht dat leerlingen die juist over voldoende kennis en vaardigheden beschikken ongeveer 55% van de punten halen.

Deze regel houdt echter onvoldoende rekening met het feit dat de moeilijkheid van de toets heel lastig te beheersen is. Onderzoek heeft bijvoorbeeld laten zien dat docenten de moeilijkheidsgraad van vragen moeilijk in kunnen schatten (Sanders, 2013). Daarnaast is het maken van goede vragen een lastige en tijdrovende klus en moet je bij het maken van de vragen rekening houden met inhoudelijke eisen. Dus de bron van vragen waaruit geput kan worden is vaak ook beperkt. Dit alles zorgt er voor dat ondanks een zorgvuldige constructie de moeilijkheid van een toets toch anders uitpakt dan waar je van te voren op ingezet hebt. Bij het vaststellen van de uiteindelijke slaagzak-grens moet je hier rekening mee houden. Dit bekent dat je bij een moeilijke toets de slaag-zakgrens wat lager legt, wat betekent dat je minder punten nodig hebt voor een 5,5. Bij een makkelijke toets leg je de grens juist wat hoger, zodat je meer punten nodig hebt voor een 5,5. Door de slaag-zak grens af te stemmen op de moeilijkheid van de toets zorg je er voor dat *je altijd even streng bent* bij het nemen van slaag-zak beslissingen. [15] [16] Hiermee voorkom je dat leerlingen uit sommige leerjaren benadeeld worden ten opzichte van andere leerlingen omdat bij hen de toets toevallig moeilijker uitpakte.

[13] Psychometrisch onderzoek heeft laten zien dat vragen het meest informatief zijn als zij van gemiddelde moeilijkheid zijn. Een toets van gemiddelde moeilijkheid zal doorgaans betrouwbaarder zijn dan hele moeilijke of hele makkelijke toets.
[14] Er zijn natuurlijk ook situaties denkbaar waarbij je de lat veel hoger legt. Denk aan de rekentoetsen voor verpleegkundigen in opleiding. Goed kunnen rekenen is cruciaal. Voor de rekentoets wordt de lat dan hoog gelegd.
[15] Ten onrechte wordt wel eens beweerd dat wanneer bij een makkelijk tentamen de grensscore wordt opgehoogd er een strengere norm wordt gehanteerd. Dit is misleidend taalgebruik want het feit dat de grensscore bij een hoger aantal punten ligt betekent niet dat de norm strenger is.
[16] Het basisprincipe "bij een gelijke prestatie, hoort een gelijke waardering" is de grondslag voor de normhandhaving bij de centrale eindexamens.

© Stichting Cito, Arnhem– versie mei 2020

11 Verfijnen van de Slaag-Zak Grens

In deze volgende paragraaf zullen we enkele methoden beschrijven waarmee de grensscore kan worden gevalideerd en verfijnd. We gaan er vanuit dat de toets als zodanig is samengesteld dat er is toegewerkt naar een voorlopige slaag-zak grens die ongeveer in het midden van de schaal ligt. We presenteren de methoden dan ook vooral als praktische hulpmiddelen om de zak-slaaggrens te verfijnen en te monitoren. Op de website van Stichting Cito vind je een handige infographic die de procedures kort beschrijft. De achtergrondinformatie vind je hieronder.

Als je kijkt naar de meest gebruikte methoden dan kun je grofweg een onderscheid maken tussen methoden die gebruik maken van inhoudelijke inschatting van de moeilijkheid van de vragen door de docent, los van feitelijke prestaties, en methoden waarbij de docent inschatting maakt van de vaardigheid van leerlingen of die juist gebruik maken van de feitelijke prestaties van leerlingen. Een combinatie is uiteraard mogelijk.

11.1 De Angoff-methode

Deze methode is het meest geschikt voor een toets met enkel meerkeuzevragen die één punt waard zijn, maar kan ook worden toegepast bij andere typen vragen.

1) Neem een groep (hypothetische) leerlingen in gedachten die de stof nét voldoende beheersen. Dit noemen we *grensleerlingen*[17].

2) Geef antwoord op de vraag: '*Wat denkt u wat de gemiddelde score op dit item zal zijn als 100 grensleerlingen dit item beantwoorden*?'. Sommige beoordelaars vinden het handiger om te denken in termen van percentages. Bij meerkeuzevragen met 1 juist antwoord komt dit neer op het schatten van het percentage van de leerlingen die de vraag goed scoren. Bij meerkeuze vragen schat men in welk percentage van het totaal aantal punten de grensleerlingen gemiddeld zullen behalen. Door het percentage te vermenigvuldigen met de maximale score en te delen door 100 krijg je een inschatting van het gemiddeld aantal punten voor de grensleerlingen.

3) Tel de waarden uit de vorige stap bij elkaar op. De uitkomst is het aantal punten waarvan je verwacht dat een leerling die de stof net voldoende beheerst zal behalen. Dit is de grensscore en geeft aan waar de slaag-zakgrens komt te liggen.

[17] Soms ook wel "zesjesleerlingen genoemd".

© Stichting Cito, Arnhem– versie mei 2020

Tabel 1 hieronder geeft een voorbeeld van de procedure voor een fictieve test met vijf vragen. Het aantal te behalen punten varieert tussen 1 en 6 (Kolom 2).

Tabel 1. Voorbeeldtabel Angoff Procedure

Vraag	Max Punten	Percentage onder grensleerlingen	Gemiddeld aantal punten (max x perc/100)
1	1	70%	0.7
2	4	50%	2.0
3	1	40%	0.4
4	6	40%	2.4
5	4	30%	1.2
TOTAAL	**16**		**6.7**

In kolom 3 zien we de inschatting van de docent. Bijvoorbeeld, bij vraag 2 schat de docent in dat grensleerlingen gemiddeld genomen 50% van het totaal aantal punten halen, dat betekent gemiddeld genomen 2 punten. Verder zien we dat vraag 1 als relatief gemakkelijk wordt ingeschat, terwijl vraag 5 volgens de docent juist een hele moeilijk vraag was. Tellen we de verwachte scores bij elkaar op dan komen we uit op 6.7 punten. Dit is de gewenste slaag-zakgrens volgens de Angoff methode. Deze grens betekent dat een leerling geslaagd is als hij/zij afgerond 42% (= 6.7 / 16 x 100%) van het totaal aantal punten behaalt. Dit betekent dat de toets gemiddeld genomen aan de moeilijke kant was.

11.2 De Directe Consensusmethode

De “directe-consensus-methode” is vergelijkbaar met de Angoff methode, alleen wordt er nu niet van je gevraagd om een inschatting per vraag te maken, maar moet je inschatting maken van de moeilijkheid van een set van vragen die inhoudelijk bij elkaar horen. Bovendien gaat deze methode er van uit dat meerdere docenten betrokken zijn bij het bepalen van de slaag-zakgrens. Dus deze methode doe je samen met je collega’s.

De directe consensusmethode bestaat uit de volgende stappen:

1) Verdeel de opgaven in clusters van vragen, waarbij elke cluster van vragen een inhoudelijk domein bevraagt.
2) Neem een (hypothetische) groep leerlingen in gedachten die de stof nét voldoende beheersen. Dit zijn de zogenaamde grensleerlingen.
3) Schat in hoeveel punten grensleerlingen gemiddeld scoren op ieder domein.
4) Bespreek de resultaten met collega’s, met name de domeinen waar de beoordelingen erg uiteenlopen. Bereik hierover consensus met je collega’s.
5) Tel de domeinscores bij elkaar op. De uitkomst is de grensscore, en dus tevens de slaag-zak grens.

© Stichting Cito, Arnhem– versie mei 2020

11.3 Methode van contrasterende groepen

Bij deze aanpak probeer je als docent van tevoren, dus zonder dat de toetsresultaten bekend zijn, de leerlingen in te delen in niveaus op basis van een inschatting van hun vaardigheid. Dit kunnen bijvoorbeeld twee contrasterende groepen zijn. Een groep van leerlingen van wie je op basis van hun eerdere prestaties in de klas vindt dat zij gezien hun vaardigheden de toets zouden moeten halen, en een groep van leerlingen die volgens jou zouden moeten zakken op de test. Vervolgens gebruik je de feitelijke resultaten op de toets om de uiteindelijke slaag-zakgrens vast te stellen. Dat wil zeggen, de slaag-zak grens is die score die de twee groepen het beste van elkaar onderscheidt.

Ter illustratie zullen we een hypothetisch voorbeeld bespreken waarbij 200 examenkandidaten een toets met 10 meerkeuzevragen heeft gemaakt. Van elke leerling heeft de docent een inschatting gemaakt of de leerling wel of niet de toets zou moeten halen. Tabel 2 geeft de resultaten weer. De docent schat dat 50 leerlingen zouden moeten zakken en dat 150 leerlingen zouden moeten slagen. Nadat de test is afgenomen kun je vervolgens voor elke mogelijke grensscore bekijken hoeveel procent van de leerlingen die hadden moeten zakken (volgens de docent) toch slaagt (= "onterecht geslaagd"). Evenzo kun je berekenen wie van de leerlingen die zouden moeten slagen (volgens de docent) alsnog zal zakken (= "onterecht gezakt"). Deze resultaten staan respectievelijk in de derde en vierde kolom. Stel bijvoorbeeld dat de grensscore op 5 wordt gezet, dan zou volgens de inschatting van de docenten 5 van de 50 "onder-grensleerlingen" (= 10%) onterecht slagen, en 25 van de 150 de boven-grensleerlingen (= 16.67%) onterecht zakken. Kijken we naar de groep als geheel dan zien dat wanneer de docent voor een grensscore 5 kiest, dat naar schatting voor 30 van 200 (= 15%) van de leerlingen een juiste beslissing genomen.

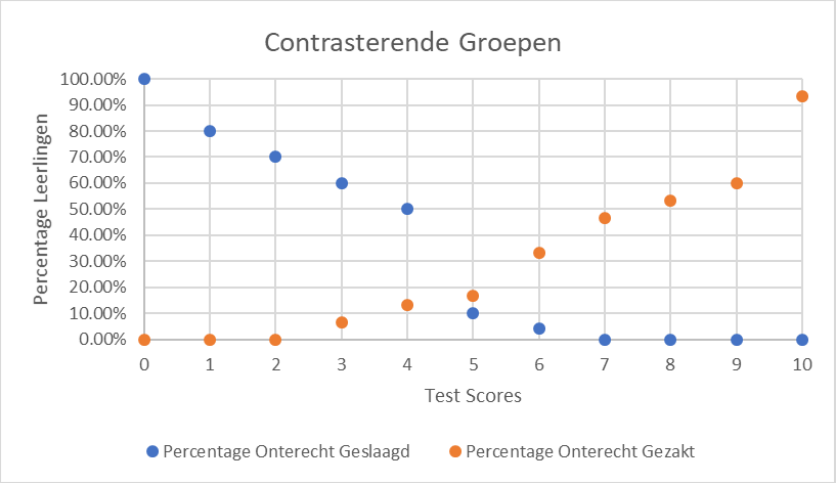
Tabel 2: Denkbeeldige data voor contrasterende groepen

Score	Aantal leerlingen met score of hoger		Classificatie		Misclassificaties
	Zakkers (50 leerlingen)	Slagers (150 leerlingen)	Onterecht geslaagd	Onterecht Gezakt	
0	50	150	100.00%	0.00%	25.00%
1	40	150	80.00%	0.00%	20.00%
2	35	150	70.00%	0.00%	17.50%
3	30	140	60.00%	6.67%	20.00%
4	25	130	50.00%	13.33%	22.50%
5	**5**	**125**	**10.00%**	**16.67%**	**15.00%**
6	2	100	4.00%	33.33%	26.00%
7	0	80	0.00%	46.67%	35.00%
8	0	70	0.00%	53.33%	40.00%
9	0	60	0.00%	60.00%	45.00%
10	0	10	0.00%	93.33%	70.00%

© Stichting Cito, Arnhem– versie mei 2020

De best passende grensscore zorgt er voor dat zo beide percentages zo laag mogelijk zijn. Om de best passende grensscore te vinden zetten we de percentages uit de vierde en vijfde kolom in een figuur.

Figuur 1: Illustratie Methode Contrasterende Groepen



In het figuur zien we dat de best passende grensscore rond 5 punten = geslaagd ligt. Bij die grensscore zijn de percentages onterecht gezakte leerlingen en geslaagden in balans. Zou je een andere grensscore kiezen, bijvoorbeeld 6, dan heb je minder onterecht geslaagden, maar dat gaat ten koste van het aantal leerlingen dat onterecht zakt. Uiteraard kan men kan eventueel rekening houden met welke fout men het vindt ergst. [18] In dit geval ligt de grensscore halverwege de schaal. Dus de toets sluit qua moeilijkheid prima aan bij de doelgroep.

11.4 Benchmarking met andere jaren / andere vakken.

Een derde manier om de slaag-zak grens vast te stellen is door de grens zo te kiezen dat het percentage leerlingen dat slaagt ongeveer overeenkomt met het percentage op een vergelijkbaar examen uit de vorige jaren. De aanname hierbij is dat de vaardigheid van de leerlingen gemiddeld genomen over de jaren heen vergelijkbaar zijn en dat daarom een ongeveer even groot percentage zou moeten slagen.

In hoeverre de aanname van vergelijkbare groepen houdbaar is hangt van verschillende factoren af. Ten eerste moet je rekening houden met natuurlijke fluctuaties. Dat wil zeggen dat puur door toeval het cohort uit het ene jaar gemiddeld genomen wat vaardiger is dan een ander cohort. Met name bij kleine groepen kunnen de schommelingen aanzienlijk zijn. De methode is dan ook niet geschikt voor kleine groepen. Ten tweede moet men ook goed naar de context van het onderwijs kijken. Een verandering in bijvoorbeeld de leermethode, onevenredige uitval door ziekte van een docent, of andere calamiteiten zoals het plotseling moet geven van een groot deel van het onderwijs op afstand, kunnen invloed hebben op de gemiddelde vaardigheid van de groep als

[18] Over het algemeen wordt een fout van de eerste soort (“onterecht slagen”) als minder problematisch gezien dan een fout van de tweede soort (“onterecht zakken”).

© Stichting Cito, Arnhem– versie mei 2020

geheel. Als je met de veranderende omstandigheden rekening mee wilt houden dan moet je deze methode niet gebruiken.

De vraag blijft: hoe weet je dat de aanname van gelijke vaardigheid redelijk is. Een praktische uitweg in dit dilemma is een geankerde benchmark. Door enkele vragen op te nemen uit oude toetsen is het mogelijk om beter inzicht te krijgen in of de verschillen in prestaties zijn toe te schrijven aan leerlingen of aan de toets zelf. Immers, als groep A gemiddelde genomen slechter presteert en je ziet dezelfde trend terug op de ankervragen, dan geeft dat sterkere aanwijzing dat het verschil is toe te schrijven is aan de groep.

11.5 Voor- en nadelen van de methoden

Bij alle bovengenoemde methoden speelt het oordeel van de docent een belangrijke rol. In de Angoff methode wordt er geoordeeld op vraagniveau en bij de direct consensusmethode per inhoudsdomein. Voordeel van deze methoden is dat er geen afnamegegevens nodig zijn. Bovendien wordt de grensscore direct gelinkt aan inhoudelijke criteria via de denkbeeldige "grensleerling". Hiermee is de methode goed toepasbaar in de klas. Nadelen van de Angoff methode is dat het veel werk is. De directe consensusmethode is minder bewerkelijk, maar nogal altijd erg intensief. Beide methodes gaan er vanuit dat docenten een goed en vergelijkbaar beeld hebben van grensleerlingen. Dit hoeft niet perse het geval te zijn. Bij de directe consensusmethode wordt dit probleem deels ondervangen door de procedure samen met collega's uit te voeren en te zoeken naar consensus. Dit kun je natuurlijk ook bij de Angoff procedure doen. [19]

In de contrasterende groepen methode geeft de docent een deskundigen oordeel over leerlingen. De methode gaat er van uit dat docenten de leerlingen voldoende goed kennen en dat zij in staat zijn om een objectief oordeel te geven. Het is de vraag of deze aanname terecht is. Leerlingen zijn niet altijd even voorspelbaar en omstandigheden kunnen ervoor zorgen dat men zich minder goed voorbereid heeft (bijvoorbeeld wanneer een toets de laatste in een rij is). Bovendien, toetsen zijn niet 100% betrouwbaar; dus er moet altijd rekening worden gehouden met een zekere mate van onzekerheid. Dit maakt de methode kwetsbaar. De methode van contrasterende groepen werkt dan ook alleen bij grotere aantallen leerlingen.

Bovengenoemde methode zijn slechts enkele mogelijkheden die in de literatuur beschreven worden.[20] De methoden dienen met de nodige terughoudendheid toegepast worden en kunnen het beste als aanvullend worden beschouwd. Welke methode het meest geschikt is hangt af van de uitvoerbaarheid binnen de specifieke context waarin de toets wordt afgenomen. Soms is het handig om verschillende methoden naast elkaar toe te passen en te kijken naar opvallende verschillen, extreme scores en uitbijters te detecteren. Als er duidelijke discrepanties zijn, dan kan het leiden tot aanpassing.

[19] Bij de eindexamens worden bij een aantal vakken Angoff-achtige procedures gebruikt voor het bepalen van normen. Hierbij worden altijd meerdere beoordelaars gevraagd. De uiteindelijke normering is het gemiddelde van de beoordelingen. Dit geeft een betrouwbaarder beeld omdat individuele verschillen worden uitgemiddeld.
[20] Voor een uitgebreid overzicht zie bijvoorbeeld Cizek, C. J. & Putham, M. B. (2007). Standard Setting: A guide to establishing performance standards on tests. New York, NY: Thousand Oaks.

© Stichting Cito, Arnhem– versie mei 2020

12 Omzetten van Scores in Cijfers

Er zijn veel verschillende manier om de toetsresultaten (scores) om te zetten in schoolcijfers, maar men gaat altijd uit van de volgende kenmerken:

1. Alle cijfers liggen tussen 1 en 10.
2. Hoe meer punten hoe hoger het cijfer, waarbij de mogelijkheid wordt open gehouden dat een aantal opeenvolgende scores tot hetzelfde cijfer leidt. Het cijfer mag in ieder geval niet lager worden als het aantal punten toeneemt. Een leerling met veel punten mag dus nooit een lager cijfer krijgen dan andere leerlingen met minder punten. [21]
3. Er is een grensscore die de grens markeert tussen leerlingen die "gezakt" zijn en leerlingen die "geslaagd" zijn. Dit is de slaag-zak grens, ook wel de grensscore genoemd. De zak-slaaggrens correspondeert met het cijfer 5,5.

Hieronder wordt een aantal methoden beschreven waarmee scores kunnen worden omgezet in cijfers.

12.1 De intuïtieve methode

Bij deze methode bepaalt de docent op basis van inhoudelijke overwegingen welke cijfer bij welk puntenaantal hoort. We zien dit bijvoorbeeld terug bij het omzetten van scores naar schoolcijfers op beoordelingsschalen. Stel: leerlingen zijn beoordeeld op vijf inhoudscriteria op een 3-puntschaal: onvoldoende (= 1 punt) / voldoende (= 2 punten) / goed (= 3 punten). De slaag-zakgrens ligt bij hoogstens één onvoldoende; leerlingen met twee of meer onvoldoendes zijn gezakt. Vervolgens wordt de volgende becijfering als volgt uitgewerkt.

Beschrijving	Cijfer	Interpretatie
5 punten	3.5	Zeer sterk onvoldoende
6 of 7 punten	4	Sterk onvoldoende
8 punten.	5	Onvoldoende
Ten hoogste een onvoldoende:		
9 of 10 punten	6	Voldoende
11 of 12 punten	7	Ruim voldoende
13 punten	8	Goed
14 punten	9	Zeer goed
15 punten (alles goed)	9.5	Uitmuntend

Deze omzetting voldoet aan de minimale eisten van schoolcijfers. We zien dat cijfers tussen 3.5 en 9.5 worden gegeven en dat voornamelijk hele cijfers worden gegeven. Het interessante van deze methode is dat er in feite een directe relatie wordt gelegd tussen de prestatie en de betekenis van de cijfers. Deze methode sluit dicht aan bij de (communicatie)-functie van schoolcijfers.

[21] Formeel luidt de regel als volgt: Voor elk willekeurig paar van leerlingen geldt dat het cijfer dat wordt toegekend aan de leerling met de meeste punten altijd minstens even groot is als het cijfer dat wordt toegekend aan de leerling met de minste punten.

© Stichting Cito, Arnhem– versie mei 2020

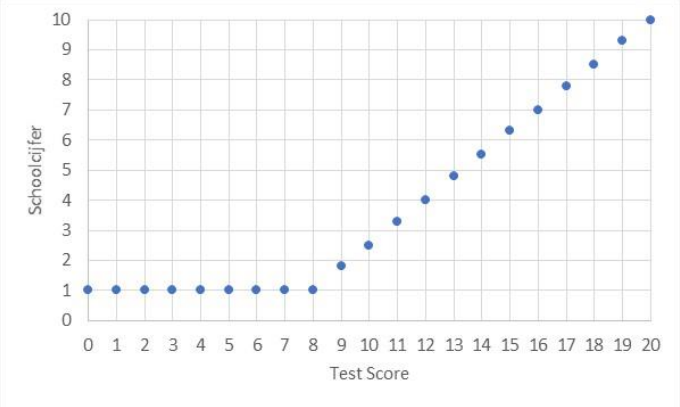
12.2 Lineaire methode

Deze methode zet de scores om in cijfers via de volgende lineaire functie:

$$\text{cijfer} = 10 - (\text{max. score} - \text{score}) \times \frac{4{,}5}{\text{max. score} - \text{grensscore}}$$

Hieronder zien we de methode toegepast op een meerkeuzetoets met 20 vragen. Elk goed antwoord levert één punt op. De grensscore is vastgesteld op 14 vragen is goed. De lineaire methode geeft de volgende omzettingsgrafiek:

Figuur 2: Omzetting volgens de lineaire methode



Intuïtief werkt de methode als volgt. Een leerling met een score gelijk aan het maximaal aantal punten krijgt een 10. Bij elk punt minder vindt wordt het cijfer in mindering gebracht zodanig dat een leerling met score op de grens precies op een 5,5 uitkomt. In het voorbeeld ligt de grensscore bij 14, Dit betekent dat je in 20-14 = 6 stappen van 10 naar 5,5 loopt en daaruit volgt dat bij elk punt minder wordt 0.75 cijferpunten afgetrokken. Bij een meerkeuzetoets komt deze methode dus neer op de regel “bij elke foute antwoord gaan er XYZ punten af”. Deze trend wordt doorgetrokken tot je bij de 1 uitkomt. Daarna wordt aan elke score die nog lager is het cijfer 1 toegekend.

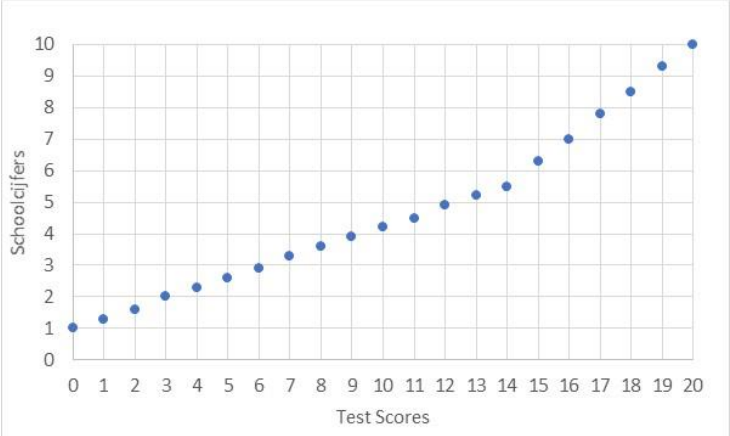
Dit is een eenvoudig uit te leggen regel. De methode voldoet aan de basiseisen van schoolcijfers. Echter zien we dat we niet de volle schaallengte benutten omdat een leerling met 7 punten of lager een 1 krijgt. Dit is niet persé onjuist, maar kan demotiverend werken omdat leerlingen die verschillend presteren toch hetzelfde cijfer krijgen, ook al is het dan onvoldoende. Je verliest op deze manier ook informatie. Om leerlingen beter inzichtelijk te maken waar ze staan zou je eigenlijk punten willen waarderen over de hele schaallengte. Een mogelijke oplossing is een lineaire functie met knik, die we hierna bespreken.

© Stichting Cito, Arnhem– versie mei 2020

12.3 Lineair omzetting met knik

Onderstaand figuur 3 geeft omzetting volgens de methode met knik:

Figuur 3: Lineair met knik



De methode werkt als volgt. Een leerling met een score gelijk aan de grensscore krijgt het cijfer 5.5. Vervolgens wordt bij elk punt extra het cijfer opgehoogd zodat de maximale score op 10. Je verdeelt de cijferschaal voor scores boven de grensscore dus op in gelijke stappen. Hetzelfde doe je voor de scores onder de grens. Je begint bij de grensscore en voor elk punt minder wordt het cijfer verlaagd zodanig dat je op 1 uitkomt als een leerling 0 punten heeft. Dus je verdeelt de cijferschaal voor scores onder de grensscore weer op in gelijke stapjes. Je hebt dus in feite twee lineaire functies: eentje voor het bepalen van de cijfers boven de grensscore, en een voor cijfers onder de grensscore. De bijbehorende formules zijn:

Voor scores < grensscore:

$$\text{cijfer} = 1 + \text{score} \times \frac{4{,}5}{\text{grensscore}}$$

Voor scores ≥ grensscore:

$$\text{cijfer} = 10 - (\text{max. score} - \text{score}) \times \frac{4{,}5}{\text{max. score} - \text{grensscore}}$$

De lineaire omzetting met knik resulteert in cijfers tussen 1 en 10. De methode heeft als bijkomende voordeel dat meer punten leidt tot hogere cijfers (afrondingen daargelaten). Echter zien we wel dat de waardering per scorepunt verschillend is voor scores boven de grensscore en scores onder de grensscore; dus de puntenwaardering hangt af van of je geslaagd bent of niet. Met name voor studenten rondom de grensscore kan dat oneerlijk overkomen. Een methode dat aan dit bezwaar tegemoet komt is de methode die bij de centrale eindexamens wordt gehanteerd. Hierbij probeert men om over een zo breed mogelijk interval van de scoreschaal de cijfers evenredig te laten stijgen met het aantal scorepunten ongeacht de normering. Op de Toetswijzer vind je een hulpmiddel waarmee scores volgens deze methode om kunt zetten.

© Stichting Cito, Arnhem– versie mei 2020

12.4 De centrale-eindexamen methode

Dit is de methode die bij de centrale eindexamens wordt gebruikt[22] en wordt ook wel de methode van Deus genoemd (Verstralen, 2010). De methode is gebaseerd op uitgangspunten:

1. Elk gescoorde punt draagt altijd bij tot een hoger examencijfer.
2. Het maximum aantal te behalen punten correspondeert met een examencijfer van 10,0.
3. Het minimum aantal te behalen punten correspondeert met een examencijfer 1,0.
4. Over een zo breed mogelijk interval van de scoreschaal is er sprake van een evenredige stijging van het cijfer met de scorepunten ongeacht de normering.

De basisformule bij deze methode is:

$$\text{Cijfer} = \frac{\text{behaalde score}}{\text{maximale score}} \times 9 + N.$$

Hierin is N de "N-term". De N-term hangt af van de moeilijkheid van de toets. Wanneer N = 1 dan volgt daar uit dat een leerling 50% van het totaal aantal punten moet behalen voor een 5,5. Het omgekeerde geldt dus ook: als de grensscore is vastgesteld op 50% van het totaal aantal punten, dan hoort daar een N-term van 1 bij. Als de toets moeilijker uitvalt, dan wordt een hogere *N* term gekozen. Dit betekent dat je minder punten nodig hebt voor een 5,5. Een de toets die gemakkelijker uitvalt krijgt een lagere *N*-term.

Zodra de N-term is vastgesteld, kunnen de cijfers worden berekend. Wanneer N = 1 kun je de bovenstaande basisformule gebruiken. Wanneer N afwijkt van 1 is het omrekenen ingewikkeld. Je ziet dat als N ongelijk is aan 1 de cijfers die uit de basisformule niet meer voldoen aan de uitgangspunten. Stel N = 1,5, dan zou een leerling die alles goed heeft een 10,5 halen. Dat is niet toegestaan. Daarom is er een ingewikkelde omrekeningsprocedure ontwikkeld, waarmee scores omgezet kunnen worden naar cijfers die voldoen aan de uitgangspunten. Het voert te ver om het in detail te bespreken.[23] We zullen ons beperken tot enkele voorbeelden van omzettingen om de methode te illustreren. Op de website van Stichting Cito vind je een Excelbestand om scores om te zetten naar een schoolcijfer als de N-term bekend is.

In figuur 4 op de op volgende pagina zie je voor een fictieve toets waarbij maximaal 20 punten kunt halen de omzetting voor verschillende N-termen en bijbehorende grensscores. Op de horizontale as staan de toets-scores, op de verticale as kun je het bijbehorende cijfer aflezen.

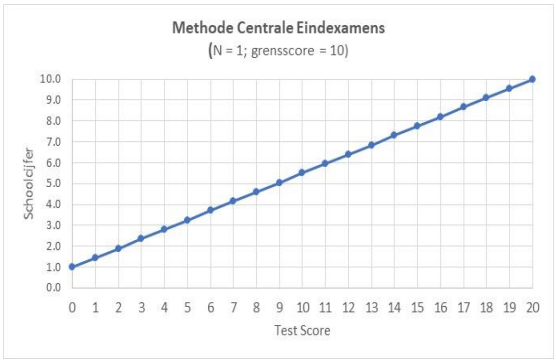
In Figuur 4A is N = 1, en we zien een rechte lijn tussen scores en cijfers. In figuur 4B is N > 1, wat betekent dat de toets bovengemiddelde moeilijk is. We dat in dit geval de cijfers wat hoger liggen dan bij vergelijkbare score voor N = 1. De lijn is als het ware wat naar boven-geschoven. Om er voor te zorgen dat cijfers tussen 1 en 10 liggen, zien we twee knikjes aan het uiteinde. Deze knikjes zijn door het algoritme optimaal gekozen

[22] Deze methode is vastgelegd door het College voor Toetsing en Examens.
[23] De geïnteresseerde lezer wordt verwezen naar de bijbehorende publicatie in de Staatscourant. https://zoek.officielebekendmakingen.nl/stcrt-2019-9325.html..

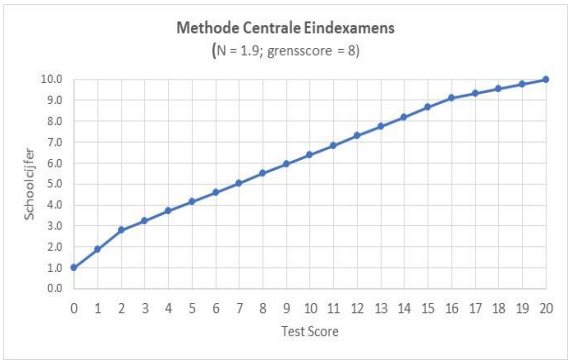
© Stichting Cito, Arnhem– versie mei 2020

Figuur 4: Omzetting van scores naar cijfers volgens Centrale Eindexamens Methode voor verschillende N-termen

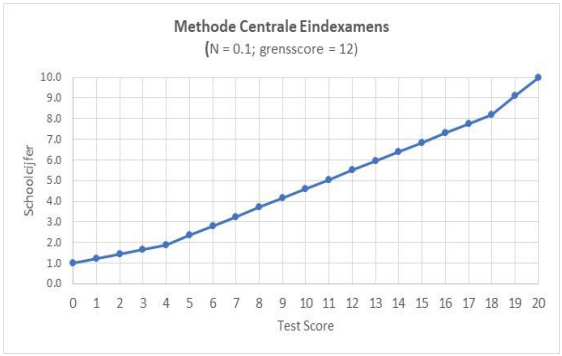
A



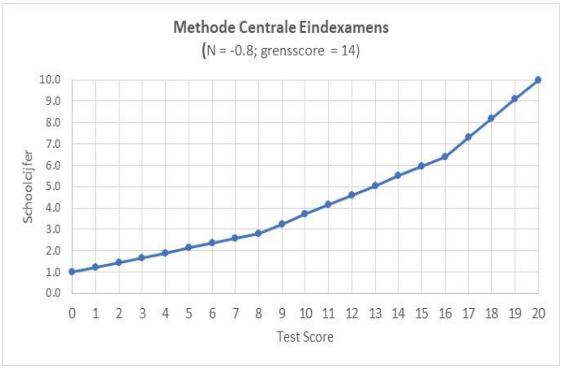
B



C



D



In Figuren 4C en 4D zien we omzettingstabellen voor N < 1. Dit zijn toetsen die gemiddelde genomen aan de gemakkelijke kant waren. Hier zien we het omgekeerde effect dan bij N > 1. De lijn is als het ware wat naar beneden geschoven, met aan de uiteindes de knikjes om er voor te zorgen dat de cijfers tussen 1 en 10 liggen. Als je C met D vergelijkt, dan zie je dat de plaats van knikjes ook afhangt van de N-term zelf. Het voordeel van deze methode is dat alle punten rondom de slaag-zakgrens evenveel waard zijn. Dit voorkomt dat leerlingen die juist rondom de grensscore het gevoel hebben dat ze benadeeld worden omdat dat ene punt net minder waard is omdat ze net onder de grens zitten.

Door de N-term slim te kiezen kun je er voor zorgen dat de gehanteerde norm over toetsen constant blijft. Een leerling die bijvoorbeeld in 2017 eindexamen doet heeft dan evenveel kans om te zakken of te slagen als een vergelijkbare leerling die in een ander jaar examen doet. Voor de centrale eindexamens worden de N-termen vastgesteld door de College voor Toetsen en Examens (CvTE), waarbij in de wet is vastgelegd de gekozen *N*-term altijd tussen 0 en 2 moet liggen.[24] Aan de keuze van N liggen ingewikkelde psychometrische analyses van de toetsresultaten en proefafnames ten grondslag, ook wel normhandhavingsonderzoek genoemd (Eggen en Sanders, 2013). Voor dit onderzoek heb je genoeg afnamegegevens nodig, en het vereist geavanceerde psychometrische kennis. Voor de schoolpraktijk van alledag is dat niet

[24] Naast een praktische grens zit er ook theoretisch grenzen aan de N-termen. Alleen *N*-termen tussen -2 en 4 zijn zinvol. Dit betekent dat N-termen onder de -2 (bijv. -3) dezelfde cijfers oplevert als een N van -2, en boven de 4 (bijvoorbeeld 5.3) dezelfde cijfers oplevert als een N-term van 4.

© Stichting Cito, Arnhem– versie mei 2020

mogelijk, maar met de paar eenvoudigere procedures, waaronder Angoff, kun je toch er voor zorgen dat de norm over de jaren heen constant blijft.

Wanneer je toch deze omzettingsmethode wil toepassen voor je eigen toetsen dan kun je ook via de grensscore te werk gaan. Wanneer je de grensscore hebt vastgesteld, dan kun je daar in principe de bijbehorende *N*-term bij berekenen en de scores omzetten in cijfers. Op de website van Stichting Cito staat een handige hulpmiddel waarmee je zelf volgens deze methode omzettingstabellen kunt generen op basis van de gekozen schaallengte en grensscore. Er is wel één beperking. De grensscore moet tussen de 25% en 75% van het maximaal aantal te behalen punten liggen. Dus stel de scores lopen van 0 tot 40, dan moet de grensscore op of tussen 10 en 30 liggen. Valt de grensscore buiten deze range dan werkt de procedure niet. Dit komt door de manier waarop het algoritme de scores omzet in cijfers. In de meeste gevallen zal dit geen probleem zijn. Als de grensscore buiten deze range valt dan is er sprake van een wel hele makkelijke of hele moeilijke toets. Mocht je toch een grensscore willen hanteren die buiten de range valt, dan kun je het beste de lineaire methode met knik gebruiken. Ook hiervoor zijn handige hulpmiddelen beschikbaar. Een daarvan is het programma Scores2cijfer die beschikbaar is via Toetswijzer: Toetsen op School.

13 Geraadpleegde literatuur

Baack, J. (2008). Praktische beoordelingsformulieren voor spreekvaardigheid. *Levende Talen Magazine*, 8, 13-16. http://www.lt-tijdschriften.nl/ojs/index.php/ltm/article/view/371/364

Berkel, H., & Bax A (2006). *Toetsen in het hoger onderwijs* (tweede druk. Houten, NL: Bohm, Safleu & Loghum.

Blauw, S. (2018). *Het best verkochte boek ooit*. Amsterdam, NL: De Correspondent.

Cotan (2015). *Aanvulling op het COTAN Beoordelingssysteem m.b.t. Fairness.* Utrecht, NL: Cotan / Nederlands Instituut voor Psychologen (NIP).

Dane J. (2014). *Een 5 voor vlijt*. In *Toets*. Magazine uitgegeven door Bureau Ice.

Ekens, T., & Meestringa, T. (2013). *Beoordeling van en feedback op schrijfvaardigheid. Een handreiking voor de tweede fase voortgezet onderwijs.* Enschede: Stichting Leerplanontwikkeling (SLO).

Eggen, T.J.H.M., & Sanders, P. (1993). *Psychometrie in de praktijk*. Arnhem, NL: Stichting Cito.

Jonsson, A. & Svingby, G. (2007). The Use of Scoring Rubrics: Reliability, Validity and EducationalConsequences. *Education Research Review*, 130-144..

Kuhlemeijer, H., Hemker, B., & Bergh, H. (2013). Impact of verbal labels on the Elevation and Spread of Performance ratings. *Applied Measurement in Education*, 26, 16-33.

Ragupathi, K., & Lee, A. (2020). Beyond Fairness and Consistency in Grading: The Role of Rubrics in Higher Education. In Sanger C. S. & Gleason (Eds). *Diversity and inclusion in global higher education.* Singapore: Palgrave Macmillan. https://doi.org/10.1007/978-981-15-1628-3_3

Sanders, P. (2017). Toetsen op school (Herziene versie). Arnhem, NL: Stichting Cito. https://www.cito.nl/-/media/files/kennis-en-innovatie-onderzoek/toetsen-op-school/cito_toetsen_op_school.pdf?la=nl-NL

Sluijsmans, D. (2013). Verankerd leren. Vijf bouwstenen voor professioneel beoordelen in het hoger onderwijs. Rede. Zuyd Hogeschool. https://www.vereniginghogescholen.nl/system/knowledge_base/attachments/files/000/000/450/original/Verankerd_in_Leren_Vijf_bouwstenen_voor_professioneelbeoordelen_in_het_hoger_beroepsonderwijs-_Dominique_Sluijsmans_-april2013.pdf?1443615013

Stichting Cito (2020). Handreiking RV-Toets Deel C – Cijfergeving: Handreiking cijfergeving. Arnhem, NL: Stichting Cito https://www.cito.nl/onderwijs/voortgezet-onderwijs/centrale-examens-voortgezet-onderwijs/schoolexamens-online-hulp-docenten-rv-toets

Verstralen, H. (2010). De methode omzetting examens VO (Dues). Online addendum bij: Sanders (2017). Toetsen op school. https://www.toetswijzer.nl/berichten/toetsspecial-toetsen-op-school

© Stichting Cito, Arnhem– versie mei 2020