whitepaper

# Bepalen van de Slaag-zakgrens en het omzetten van Scores naar Cijfers

Stichting Cito | Onderzoek, Kennis, & Innovatie

cito.nl

Dit is een uitgave in de reeks **Toetsen in de Praktijk voor VO en MBO**

december 2020

# voor wie?

**Over deze reeks ...**

Een toets is méér dan een willekeurige verzameling van opgaven. Er moet over allerlei dingen nagedacht worden. Bijvoorbeeld over de inhoud, de toetsingsvorm, de beoordelingscriteria, en de grens tussen onvoldoende en voldoende. Daarnaast heb je te maken met praktische zaken zoals het aantal herkansingen en de wijze van surveillance. Stichting Cito ondersteunt hier graag bij. Op basis van de ervaring die de afgelopen 50 jaar is opgebouwd, brengen we als Stichting Cito onder de noemer 'Toetsen in de praktijk' een speciale reeks whitepapers uit waarin we allerlei praktische handreikingen geven om het toetsen goed te laten verlopen. Deze whitepapers zijn speciaal ontwikkeld voor docenten in het voortgezet onderwijs en middelbaar beroepsonderwijs. Inmiddels zijn in deze reeks de volgende vier uitgaven verschenen:

- Toetsingsvormen voor Schoolexamens op Afstand
- Handreikingen voor Veilig Toetsen op Afstand
- Bepalen van de Slaag-zakgrens en het omzetten van Scores naar Cijfers
- Beoordelen van Complexe Productieve Vaardigheden

Deze whitepapers zijn eerder al in iets andere vorm verschenen in de serie Van schoolexamens naar Diploma.

**Meer info?**

Naast whitepapers zijn er ook verschillende infographics en tools beschikbaar. Ga daarvoor naar toetstoolkit.

Heb je vragen over één van onze uitgaves, of heb je een idee voor een onderwerp in deze reeks, neem dan gerust contact met ons op via citolab@cito.nl.

# Wat kun je in dit document lezen?

## Inhoud

# 1 Inleiding

In Nederland is het gebruikelijk om schoolprestaties uit te drukken in een (school)cijfer op een schaal van 1 tot 10. Met het cijfer geven we aan hoe goed de leerling heeft gepresteerd en of de geleverde prestatie voldoende is. Een 8 betekent dat de leerling ”goed” heeft gepresteerd, een 4 geeft aan dat de prestatie “onvoldoende” was, en bij een 5.5 zit de leerling precies op de grens van wat er van hem of haar verwacht wordt.

In dit whitepaper laten we zien hoe je de toetsscores op een betekenisvolle wijze kunt omzetten in schoolcijfers. Daarbij is de keuze voor de slaag-zakgrens cruciaal. Hiermee bepaal je welke leerlingen zullen zakken voor de toets en welke leerlingen zullen slagen. Maar daar blijft het niet bij. De specifieke keuze voor de slaag-zakgrens werkt meestal ook door op de verdere becijfering van de toets. Zou je dus een andere slaag-zakgrens kiezen, dan verandert niet alleen het percentage geslaagden maar, afhankelijk van de gebruikte omzettingsmethode, ook de cijfers van andere leerlingen. Voor het omzetten van scores naar cijfers zijn dan ook twee vragen van belang: bij welke score leg je de slaag-zakgrens? En als deze grens eenmaal is vastgesteld, hoe zet je de overige scores om in een schoolcijfer?

In de volgende drie paragrafen zullen we enkele praktische methoden en hulpmiddelen bespreken die je als docent, of als docententeam, kunt gebruiken om de slaag-zak grens te bepalen of te verfijnen en we bespreken verschillende manieren om scores om te rekenen naar schoolcijfers. Maar voor we dat doen zullen we eerst even kort stil staan bij het fenomeen *schoolcijfers* op zich.

# 2 Schoolcijfers

Het geven van schoolcijfers is inmiddels zo vanzelfsprekend geworden dat we nauwelijks meer stil staan bij hun oorsprong en hun betekenis.[1] Het is opvallend hoe weinig er eigenlijk bekend is over de precieze oorsprong van het geven van cijfers voor schoolprestaties (zie bijvoorbeeld, Dane, 2014). Een van de mogelijke overwegingen was destijds wellicht dat je met cijfers gemakkelijker de schoolresultaten van een leerling kunt communiceren naar ouders dan met geschreven feedback. We spreken nu over begin 20ste eeuw, waarin geletterdheid minder goed ontwikkeld was in de algemene bevolking. Cijfers stonden daarmee symbool voor de waardering die de leerkracht gaf aan de prestaties van leerlingen, inclusief inzet en vlijt. Naarmate de tijd vorderde kregen cijfers steeds meer een absoluut en onbetwistbaar karakter. Dit heeft bij A. D. de Groot geleid tot het schrijven van zijn beroemde boekje *Vijven en zessen* (de Groot, 1966). Zijn voornaamste kritiek ging over de subjectiviteit van docenten, die zich bij het cijfergeven in hoge mate lieten leiden door allerlei irrelevante kenmerken zoals de achtergrond van de leerling. Dat maakt de cijfers tamelijk willekeurig. De Groot pleitte dan ook voor objectieve studietoetsen, met als uitgangspunt nauwkeurig uitgewerkte leerdoelen zodat de subjectieve invloeden van de docent zoveel mogelijk buitengesloten worden. Zijn ideeën hebben een belangrijke invloed gehad op toetsing en hebben uiteindelijk geleid tot de oprichting van het Cito in 1968. Tegelijkertijd ontwikkelde zich sinds de jaren 60 een nieuw vakgebied, educational assessment, waarbij toetsing wordt gezien als een meetprobleem. Het idee achter deze meetkundige benadering is dat je leerlingen kunt ordenen op onderliggende vaardigheidsschalen. Aan de individuele vaardigheid kun je vervolgens een meetwaarde toekennen. Vanuit dit meetkundig perspectief kun je toetsen dus zien als meet-instrumenten voor de vaardigheid, en de cijfers als de meetwaarden. Je kunt het vergelijken met een weegschaal waarmee we de waarde van ons gewicht meten; de weegschaal is hierin de toets, het gewicht is de meetwaarde. Zoals we verderop zullen zien is het bijzondere van schoolcijfers dat ze zowel een belangrijke communicatieve functie hebben, als ook dienen als meetwaarden voor de onderliggende vaardigheid. Cijfers hebben dus beide elementen in zich.

Laten we eerst nog iets verder ingaan op de betekenis van toetsscores. In eerste instantie levert een toets een *ruwe score* op. Dat is het totaal aantal punten dat is toegekend aan de correcte elementen in het antwoord. Bijvoorbeeld, bij een meerkeuzetoets is de ruwe score het aantal goede antwoorden. De ruwe scores zou je kunnen zien als een *meting* van de vaardigheid. Over het algemeen gaan we er (stilzwijgend) vanuit dat leerlingen die de stof beter beheersen meer punten zullen halen, en ook andersom: hoe meer punten de leerling behaald heeft, hoe hoger zijn of haar vaardigheid. Ruwe scores hebben zonder contextuele informatie echter geen enkele inhoudelijke betekenis! Het feit dat Marieke 14 vragen goed heeft (= 14 punten) vertelt ons alleen dat Marieke 14 vragen goed heeft 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 al als “voldoende” beschouwde, dan pas weet je dat zij zeer goed gepresteerd heeft op de toets.

Contextuele informatie is dus essentieel om de prestatie van leerlingen, zoals die van Marieke, zinvol te kunnen interpreteren. Het plaatsen van prestaties in een inhoudelijke context is in wezen wat er gebeurt als je de ruwe toetsscores omzet naar schoolcijfers. Het toekennen van betekenis aan toetsscores wordt ook wel *normeren* genoemd (Drenth & Sijtsma, 2006).

1 Een kort historische overzicht is te vinden in onder andere Dane (2014) en De Rooy (2018).

Schoolcijfers zijn in feite *genormeerde* scores, waaraan we *zelf* een bepaalde betekenis toekennen. In Nederland is de volgende duiding gangbaar:[2]

| 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 |

De betekenis van cijfers staat grotendeels los van de specifieke toets en het vak. In het geval van Marieke zouden we het cijfer 9,5 geven, waarmee voor iedereen duidelijk is dat zij zeer goed tot uitmuntend gepresteerd heeft, ook al hebben we (waarschijnlijk) geen enkel idee hoe de toets zelf er uit zag. En wanneer een leerling een 8 heeft voor Frans en een 6 voor Nederlands, dan mogen we daaruit concluderen dat de beheersing van de Nederlands taal op de grens zit van wat het veld vindt dat de leerling zou moeten kunnen, terwijl voor Frans de taalbeheersing veel hoger is dan we minimaal van de leerling zouden verwachten. Kun je nu ook zeggen dat de leerling beter Frans spreekt dan Nederlands? Het antwoord is uiteraard *nee*. Dat zou het spreekwoordelijke appels met peren vergelijken zijn. Uiteindelijk gaat het bij schoolcijfers om een kwalitatief oordeel van de schoolprestatie, gegeven de vak- en opleidingsspecifieke onderwijsleerdoelen, uitgedrukt in cijfers.

Meestal worden de schoolcijfers direct berekend uit de ruwe scores via een wiskundige omzettingsformule. Er is dan sprake van een één-op-één relatie tussen de toetsscores (meetwaarden) en de cijfers. Hierdoor krijgen cijfers eveneens het karakter van metingen op een schaal. In dit geval een vaardigheidsschaal van 1 tot 10. Deze eigenschap maakt dat we probleemloos cijfers op tienden nauwkeurig rapporteren, hoewel je je kunt afvragen of deze mate van nauwkeurigheid wel zo zinvol is. Verderop in het whitepaper zullen we enkele wiskundige omzettingsmethoden de review laten passeren. Het is daarbij goed om te beseffen dat de keuze voor de specifieke omzettingsmethode bepaalt of een leerling een 8.5 of een 9 krijgt. De gebruikte omzettingsmethode is dus mede bepalend voor de *inhoudelijke* kwalificatie die we aan een geleverde prestatie geven. Het wiskundig omzetten van scores naar cijfers is dus meer dan een rekenkundig trucje. Het kiezen van een omzettingsmethode dient daarom met dezelfde zorgvuldigheid te gebeuren als het vaststellen van een slaag- zakgrens.

In de volgende paragraaf zullen we eerste enkele methoden bespreken waarmee je op systematische wijze een slaag- zakgrens kunt bepalen. Vervolgens zullen we enkele procedures bespreken waarmee je scores om kunt zetten in cijfers.

2 Bron: https://nl.wikipedia.org/wiki/Schoolcijfer

# 3 Methoden om Slaag-Zakgrenzen vast te Stellen

## 3.1 De “50% of meer goed is geslaagd” regel

Een simpele regel om de slaag-zakgrens vast te stellen is de “50%-goed of meer is geslaagd” regel. Volgens deze regel is de leerling geslaagd als hij of zij de helft of meer van het totaal aantal punten heeft behaald. Dit is een hele inzichtelijke regel en makkelijk toe te passen. Het is ook voor de leerlingen vaak ook een hele intuïtieve regel. Het roept weinig discussie op. Er zijn echter twee bezwaren tegen het blind gebruiken van deze regel. Ten eerste is de minimale prestatie-eis van “50% of meer goed” ook maar een arbitraire keuze. Je kunt je als docent afvragen: Is het echt zo dat leerlingen die de helft of meer goed hebben de stof op het gewenste niveau beheersen? Dat hoeft natuurlijk niet persé het geval te zijn. Neem een spellingstoets. Het juist kunnen spellen van woorden is ontzettend belangrijk. Dat zou je dan ook graag tot uitdrukking willen brengen in een strengere slaag-zakgrens. Op basis van inhoudelijke overwegingen zou je dan de slaag-zakgrens bijvoorbeeld bij 70% goed of meer willen leggen.

Ten tweede houd je met het blind toepassen van deze regel onvoldoende rekening met de moeilijkheid van de toets. Onderzoek heeft laten zien dat zelfs ervaren docenten de moeilijkheidsgraad van hun vragen moeilijk in kunnen schatten (Sanders, 2017). Daarnaast is het maken van goede toetsopgaven een lastige en tijdrovende klus, waarbij je ook rekening moet houden met een heel scala van inhoudelijke eisen. Dus je hebt als docent ook maar een beperkte bewegingsvrijheid bij het maken van nieuwe opgaven. Al deze praktische belemmeringen zorgen er voor dat ondanks een zorgvuldige constructie de moeilijkheid van een toets toch anders kan uitpakken dan waar je als docent van te voren op ingezet hebt. Zou je dan alsnog de “50% of meer regel” hanteren, dan verleg je in feite de prestatienorm. Bijvoorbeeld, stel dat de toets moelijker uitvalt dan je had gewild, en je zou dan toch de “50%” regel hanteren, dan ben je in feite strenger geworden in je beoordeling.

De uitdaging bij het vaststellen van een adequate slaagzak-grens is het vinden van de juiste balans tussen de inhoudelijke eisen en de moeilijkheid van de toets. Dit betekent dat je bij een relatief moeilijke toets de slaag-zakgrens wat lager legt; een leerling heeft dan minder punten nodig voor een 5,5. Is de toets relatief gemakkelijk, dan leg je de slaag-zakgrens juist wat hoger.[3] Door de slaag-zakgrens af te stemmen op de moeilijkheid van de toets kun je er voor zorgen dat de strengheid waarmee je slaag-zakbeslissingen neemt recht doet aan de gestelde prestatie-eisen. Bovendien voorkom je dat leerlingen uit het ene jaar worden benadeeld ten opzichte van leerlingen uit andere jaren doordat zij toevallig een toets voorgelegd kregen die wat moeilijker bleek te zijn.[4]

---

3 Ten onrechte wordt wel eens beweerd dat wanneer bij een makkelijk tentamen de grensscore wordt opgehoogd er sprake is van een strengere norm. Dit is misleidend taalgebruik want het feit dat de grensscore bij een hoger aantal punten ligt betekent niet dat de *norm* strenger is.

4 Het basisprincipe “bij een gelijke prestatie, hoort een gelijke waardering” is de grondslag voor de normhandhaving bij de centrale eindexamens.

In de volgende paragraaf geven we enkele handvatten voor het vaststellen van een inhoudelijk gefundeerde slaag-zakgrens.

## 3.2 Geavanceerde Standaardbepalingsmethoden

De meest gebruikte standaardbepalingsmethodes kun je grofweg in twee verschillende varianten indelen. Methoden uit de eerste variant maken gebruik van een inhoudelijke inschatting van de moeilijkheid van de vragen door de docent. Op basis van de moeilijkheid wordt de slaag-zakgrens vastgesteld. In deze methode staan de items centraal. De tweede variant bestaat uit methoden waarbij gebruik wordt gemaakt van het oordeel van docenten over de vaardigheid van de eigen leerlingen. Dus bij deze methode staan de leerlingen zelf centraal. We zullen de methoden nu verder uitwerken.

### 3.2.1 De Angoff-methode

De Angoff methode is een populaire methode (Cizek & Bunch, 2007). De methode wordt vooral bij meerkeuzetoetsen toegepast. Centraal in de methode staat het begrip "grensleerling". Een grensleerling is een denkbeeldige leerling met een vaardigheid die precies op de grens van slagen zit. De grensleerling representeert dus de minimale vaardigheid die je verwacht van leerlingen leerling die zullen slagen voor de toets. De vraag is dan: welke score verwacht je voor grensleerlingen? De Angoff-methode probeert hier antwoord op te geven.

Om de Angoff methode te illustreren zullen we eerst uitgaan van meerkeuzevragen met één juist antwoord. De methode verloopt dan als volgt:

1 Neem een groep (hypothetische) leerlingen in gedachten die de stof nét voldoende beheersen. Dit noemen we *grensleerlingen*.[5]
2 Geef bij elke opgave in de toets antwoord op de vraag: *'Van alle grensleerling, hoeveel procent van de leerlingen denk je dat het item goed zal beantwoorden?'*. Het percentage is een inschatting van de moeilijkheid van de vraag.
3 Wanneer je alle vragen hebt beoordeeld, deel dan de percentages door 100 en tel de uitkomsten bij elkaar op. De uitkomst van deze som is het aantal punten waarvan je verwacht dat een leerling die de stof net voldoende beheerst zal behalen. Dit is tevens de beoogde slaag-zakgrens.

Wanneer je een toets hebt met opgaves waarbij leerlingen een X-aantal punten kunnen halen, dan verandert stap 2. Je kijkt dan niet langer naar het percentage van de leerlingen dat het item goed zal beantwoorden. In plaats daarvan geef je een inschatting van het gemiddelde percentage van het totaal aantal te behalen punten dat je verwacht dat grensleerlingen zullen behalen. De gemiddelde percentages kun je omrekenen naar een verwachte gemiddelde score per vraag en op basis daarvan kom je tot de beoogde slaag- zakgrens.

Onderstaande tabel geeft een voorbeeld van de procedure voor een denkbeeldige test met vijf opgaven, waarbij het aantal te behalen punten varieert tussen 1 en 6 (zie kolom 2).

5 Soms ook wel "zesjesleerlingen" genoemd.

**Tabel 1** | Voorbeeldtabel Angoff Procedure

| Vraag | Max Punten | Verwachte % van totaal punten door grenslln. | Verwachte Gemiddeld aantal punten door grenslln. |
|---|---|---|---|
| 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 inschattingen van de fictieve docent. Bijvoorbeeld, bij vraag 2 schat de docent in dat grensleerlingen gemiddeld genomen 50% van het totaal aantal punten halen. Omgerekend naar scorepunten komt dat neer op een verwachte gemiddelde score van 2 punten (zie kolom 4). Verder zien we dat vraag 1 als relatief gemakkelijk wordt ingeschat, terwijl vraag 5 volgens de docent juist relatief moeilijk was. Tellen we de verwachte scores in kolom 4 bij elkaar op dan komen we uit op een totaal van 6.7 punten. Dit is de gewenste slaag-zakgrens volgens de Angoff methode. Dit betekent dat volgens de Angoff-methode een leerling is geslaagd als hij of zij afgerond 42% (= 6.7 / 16 x 100%) van het totaal aantal punten behaalt.

De Angoff methode kan een handige manier zijn om beter grip te krijgen op het bepalen van een slaag-zakgrens. Een nadeel van de methode is dat het nogal subjectief is. Wanneer de methode wordt ingezet voor grootschalige high-stakes tests worden er daarom altijd meerdere beoordelaars (standaardsetters) ingezet. De beoordelaars gaan dan eerst uitgebreid met elkaar in gesprek om tot een goede afbakening van grensleerlingen te komen en daarover consensus te bereiken. Het gemiddelde van alle beoordelingen, eventueel na uitsluiting van de strengste en minst strengste beoordelaar, bepaalt dan de slaag-zak grens.

### 3.2.2 De directe consensusmethode

De directe-consensus-methode is vergelijkbaar met de Angoff methode, alleen wordt er nu van docenten gevraagd om een inschatting te maken van de prestaties van grensleerlingen op een cluster van vragen. Bovendien gaat deze methode er sowieso vanuit 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 de 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 beoogde slaag-zak grens.

### 3.2.3 Methode van contrasterende groepen

De methode van contrasterende groepen (Sanders & Verstralen, 2011) is gebaseerd op een indeling van de leerlingen op basis van externe criteria, onafhankelijke van de resultaten op de toets. Deze indeling vindt plaats door de docenten die de leerlingen goed kennen. In het kort werkt de methode als volgt:

- Geef voor elke leerling aan of je vindt dat hij of zij op voldoende niveau zit ('ja', 'nee', 'weet niet'). Het resultaat is een indeling van de leerlingen in twee contrasterende groepen (wel of niet op niveau), en een restgroep van wie je het niet weet.
- Neem de toets af.
- Kijk op basis van de behaalde scores door de leerlingen bij welke grensscore zoveel mogelijk leerlingen uit de groep die als "voldoende vaardig" worden beschouwd zullen slagen, en tegelijkertijd zoveel mogelijk leerlingen uit de groep "onvoldoende vaardig" zullen zakken. De grensscore die zorgt voor de beste balans is de beoogde slaag- zakgrens.

Ter illustratie zullen we een fictief voorbeeld bespreken waarbij we uitgaan van een toets met 10 meerkeuzevragen en 200 leerlingen. Van elke leerling heeft de docent een inschatting gemaakt of de leerling al dan niet op niveau zit, en dus of de leerling wel of niet de toets zou moeten kunnen halen. Tabel 2 geeft de fictieve resultaten weer. In dit voorbeeld schat de docent in dat 50 leerlingen over onvoldoende vaardigheid beschikken en zouden moeten zakken en dat 150 leerlingen wel voldoende vaardig zijn en dus zouden moeten slagen. In de kolommen 2 en 3 zie je hoe de leerlingen uit beide groepen hebben gescoord. Je ziet bijvoorbeeld dat van de 150 leerlingen die door de docent als "voldoende vaardig" zijn ingeschat, er 100 leerlingen zijn die 6 punten of meer hebben behaald.

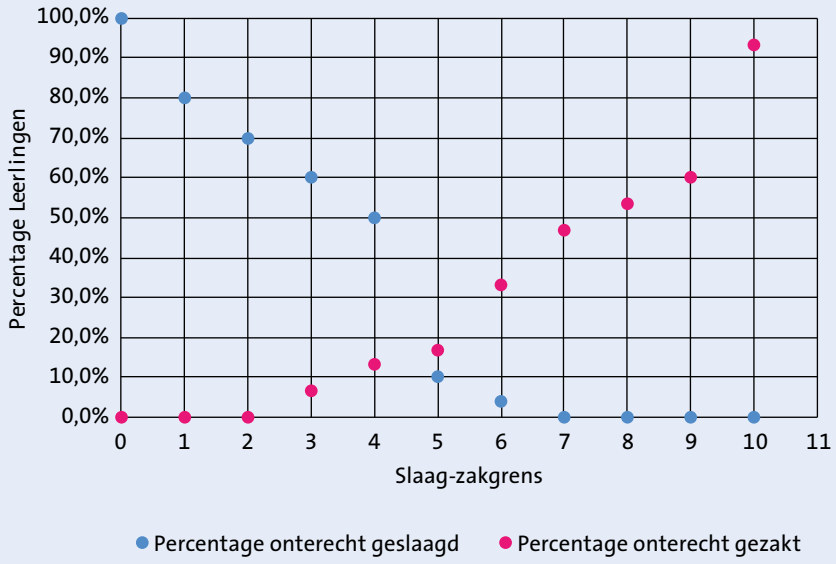
Op basis van deze gegevens kun je voor elke mogelijke slaag-zakgrens bekijken hoeveel procent van de leerlingen die hadden moeten zakken (volgens de docent) toch slaagt (= “onterecht geslaagd”). Evenzo kun je berekenen wie van de leerlingen die zouden moeten slagen (volgens de docent) alsnog zal zakken (= “onterecht gezakt”). Deze resultaten staan respectievelijk in de derde en vierde kolom. Stel bijvoorbeeld dat de slaag-zakgrens op 5 wordt gezet (zie vet-gedrukte tekst in de tabel), dan zouden volgens de inschatting van de docenten 5 van de 50 “zakkers” (= 10%) onterecht slagen. Van de 150 “slagers” zouden er bij deze norm 125 leerlingen ook daadwerkelijk slagen voor de toets. Dit betekent ook dat bij deze norm 25 van de 150 de “slagers” (= 16.67%) onterecht zouden zakken. Kijken we naar de groep als geheel dan zien we dat wanneer de docent voor een slaag-zakgrens van 5 kiest, dat naar schatting voor 30 van 200 (= 15%) van de leerlingen een onjuiste beslissing wordt genomen (= 15% misclassificaties).

**Tabel 2** | Denkbeeldige data voor contrasterende groepen

| Score | Aantal leerlingen met deze score of hoger | | Classificatie | | Misclassificaties |
|---|---|---|---|---|---|
| | “Zakkers” (50 lln) | “Slagers” (150 lln) | 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% |

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

**Figuur 1** | Illustratie Methode Contrasterende Groepen




In het figuur zien we dat de best passende grensscore rond “5 punten of meer is 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 ergst vindt.[6] In dit geval ligt de grensscore halverwege de schaal. Dus de toets sluit qua moeilijkheid prima aan bij de doelgroep.

### 3.2.4 Benchmarking met voorgaande jaren

De derde manier om de slaag-zak grens vast te stellen kijkt naar de resultaten uit voorgaande jaren. Volgens deze methode kies je de grens zodanig dat het percentage leerlingen dat slaagt ongeveer overeenkomt met het slagingspercentage op een vergelijkbare toets uit vorige jaren. De aanname hierbij is dat de gemiddelde vaardigheid van de leerlingen elk jaar ongeveer hetzelfde is, en dus dat elk jaar ongeveer een even groot percentage moeten slagen.

In hoeverre de aanname van gelijkblijvende vaardigheid houdbaar is hangt van verschillende factoren af. Ten eerste moet je rekening houden met natuurlijke fluctuaties. Dat wil zeggen dat puur door toeval de leerlingen uit het ene jaar gemiddeld genomen wat vaardiger zijn dan in andere jaren. Met name bij kleine groepen kunnen deze toevallige schommelingen aanzienlijk zijn. Deze methode is dan ook minder goed geschikt voor kleine groepen. Ten tweede is ook de onderwijscontext van belang. Een veranderde leermethode, onevenredige uitval door ziekte van een docent, of overstappen op onderwijs op afstand, kunnen invloed hebben op de gemiddelde

6 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”).

vaardigheid van de groep als geheel. Als dit soort zaken hebben gespeeld dan is de benchmarking over het algemeen geen valide methode.

Maar de vraag is dan natuurlijk: Hoe weet je of de aanname van gelijke vaardigheid redelijk is? Je kunt dit helaas nooit met zekerheid vaststellen, maar een praktische uitweg in dit dilemma is door gebruik te maken van een geankerde benchmark. Dit houdt in dat je in de nieuwe toets ook enkele vragen opneemt uit een oude toets. Als je ziet dat de leerlingen die de nieuwe toets maken op de oude toetsvraag slechter scoren dan hun voorgangers, dan is dat een aanwijzing dat de gemiddelde vaardigheid van de huidige groep leerlingen wat lager ligt. Ook hier geldt weer dat je voorzichtig moet zijn met het trekken van sterke conclusies. Je hebt over het algemeen maar een beperkte hoeveelheid informatie (weinig leerlingen, weinig vragen). Dus je moet er rekening mee houden dat de resultaten omgeven zijn door substantiële onzekerheidsmarges. In de context van het toetsen in de dagelijkse praktijk zijn bovenstaande methodes dan ook vooral bedoeld om als docenten iets meer grip te krijgen op het vaststellen van slaagzakgrenzen. Als je als docent ziet dat de leerlingen uit een bepaald jaar het opeens veel slechter doen dan leerlingen in andere jaren, dan is dat zeker iets wat je mee moet nemen in de slaagzakgrensbepaling.

## 3.3 Voor- en Nadelen van de Standaardbepalingsmethoden

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 expliciet gelinkt aan inhoudelijke criteria via de denkbeeldige “grensleerling”. Hiermee is de methode goed toepasbaar in de klas. Een groot nadeel van de Angoff methode is de hoeveelheid werk dat het met zich meebrengt. De directe consensusmethode is minder bewerkelijk, maar nog altijd erg intensief. Beide methodes gaan er vanuit dat docenten een goed en vergelijkbaar beeld hebben van grensleerlingen. Dit hoeft niet persé 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.

Bij de methode van contrasterende groepen geeft de docent een deskundigenoordeel 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 vaak onvoorspelbaar en omstandigheden kunnen ervoor zorgen dat men zich minder goed voorbereid heeft (bijvoorbeeld wanneer een toets de laatste in een rij is) of juist extra goed. Bovendien moet je er rekening mee houden dat de toetsen niet 100% betrouwbaar zijn. Het gevolg is dat dat de berekende percentages foutenmarges hebben. Bij kleine aantallen leerlingen, en onbetrouwbare toetsen, kunnen deze foutenmarges aanzienlijk zijn. Dit maakt de methode kwetsbaar voor toevalligheden. De methode van contrasterende groepen werkt dan ook alleen bij grotere aantallen leerlingen en toetsen met voldoende betrouwbaarheid.

## 3.4 Tot Slot

Het bepalen van slaag- zakgrens is een belangrijk onderdeel van de toetsing. In dit deel van het whitepaper hebben we enkele mogelijkheden besproken die helpen om tot een weloverwogen en geïnformeerde keuze te komen. 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.

De besproken methoden vormen slechts een greep uit de vele – nog meer geavanceerde – methoden die in de literatuur over standaardbepaling beschreven worden (zie Cizek & Bunch, 2007; Sanders & Verstralen, 2011). De meeste methoden zijn echter ontwikkeld voor gebruik bij grootschalige *high-stakes* tests, waarbij veel leerlingen betrokken zijn en de individuele belangen groot. Denk bijvoorbeeld aan de centrale eindexamens. Dat maakt de meeste methodes minder geschikt voor gebruik in de klas, en daarom hebben we ons beperkt tot de meest praktische methoden.

# 4 Omzetten van Scores in Schoolcijfers

Toetsscores kunnen op verschillende manieren worden omgezet in schoolcijfers. In alle gevallen hanteert men de volgende uitgangspunten:

1 Alle cijfers liggen tussen 1 en 10.
2 Hoe meer punten een leerling heeft behaald, hoe hoger het cijfer, waarbij de mogelijkheid open wordt gehouden dat een aantal opeenvolgende scores tot hetzelfde cijfer leidt. Het cijfer mag in ieder geval *niet* lager worden als het aantal punten toeneemt.[7]
3 Er is een score die de grens markeert tussen leerlingen die "gezakt" zijn en leerlingen die "geslaagd" zijn. Dit is de slaag-zakgrens, ook wel de grensscore genoemd. De slaag-zakgrens correspondeert met het cijfer 5,5. Het vaststellen van een slaag-zakgrens is hierboven besproken.

Hieronder bespreken we aantal van de meest gebruikte omzettingsmethoden.

## 4.1 Cijfers op basis van standaardbepaling

Bij deze methode bepaal je als docent op basis van inhoudelijke overwegingen welk cijfer bij welk puntenaantal hoort. We zien dit bijvoorbeeld terug bij het gebruik van beoordelingsschalen. Stel: leerlingen zijn voor vijf inhoudelijk criteria beoordeeld op een 3-puntsschaal: 'onvoldoende' (1 punt), 'voldoende' (2 punten) en 'goed' (= 3 punten). Stel dat je op basis van een inhoudelijke analyse hebt bepaald dat je maximaal voor één criterium een onvoldoende mag hebben om te slagen. Met andere woorden, leerlingen met twee of meer onvoldoendes zijn gezakt. Dit is de slaag-zakgrens. Vervolgens kun je op basis van inhoudelijk overwegingen (bijvoorbeeld) tot de volgende omzettingstabel komen.

| Beschrijving | Cijfer | Interpretatie |
|---|---|---|
| 5 punten | 3,5 | Zeer sterk onvoldoende |
| 6 of 7 punten | 4 | Sterk onvoldoende |
| 8 punten. | 5 | 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 |

7 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.

Deze omzetting voldoet aan de minimale eisen van schoolcijfers. We zien dat de cijfers tussen 3.5 en 9.5 liggen, en dus tussen 1 en 10. Ook geldt dat het cijfer toeneemt met het aantal punten. 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 oorspronkelijke communicatieve functie van schoolcijfers. Deze methode kun je in principe bij elke toets toe passen, ook bij meerkeuzetoetsen. Je zou dit zelfs op een Angoff-achtige manier kunnen aanpakken. Je maakt dan een inschatting van de verwachte score voor leerlingen die de stof “voldoende”, “goed”, etc. beheersen. Die verwachte scores kun je dan gebruiken om een omzettingstabel te maken. De meerwaarde van deze aanpak is dat er een directe link is tussen het professionele onderwijskundige oordeel van de docent en het toegekende cijfer. Nadeel is echter dat leerlingen deze procedure mogelijk als erg subjectief, en misschien zelfs oneerlijk, ervaren omdat je bijvoorbeeld minder extra punten nodig hebt om van een 6 een 7 te maken, dan om van een 7 een 8 te maken, Met andere woorden, er is niet persé een lineaire relatie tussen het aantal punten en het uiteindelijke cijfer.

## 4.2 Lineaire methode

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

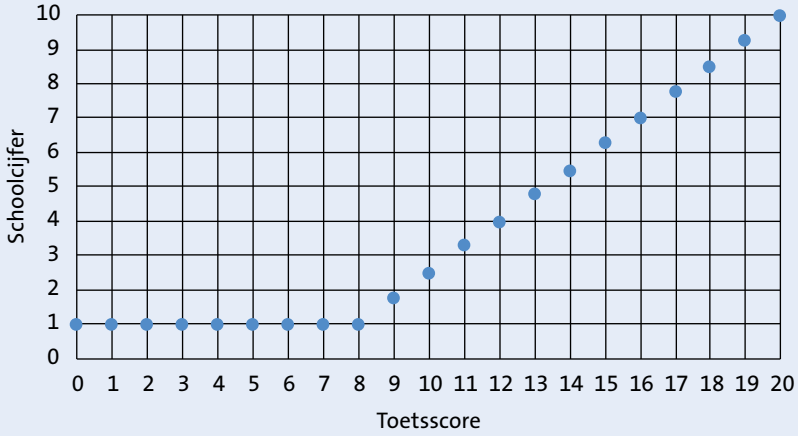
$$\text{cijfer} = 10 - (\text{maximale score} - \text{score}) \times \frac{4{,}5}{\text{maximale score} - \text{grensscore}}$$

Bijvoorbeeld, stel je hebt een meerkeuze toets met 20 vragen, dus de maximale score is 20 en de grensscore (= slaag- zakgrens) ligt bij 14 punten of meer. Het cijfer voor een leerling met 18 vragen goed (= 18 punten) wordt dan:

$$\text{cijfer} = 10 - (20 - 18) \times \frac{4{,}5}{20 - 14} = 8.5.$$

Deze berekening kunnen we voor alle scores uitvoeren. Dat levert onderstaande omzettingsgrafiek op:

**Figuur 2** | Omzetting volgens de lineaire methode




Op de horizontale as staan de toets-scores en op de verticale staat het bijbehorende cijfer. Intuïtief werkt de methode als volgt. Een leerling met een score gelijk aan het maximaal aantal punten krijgt een 10. Voor alle niet behaalde punten wordt het cijfer evenredig in mindering gebracht en wel zodanig dat een leerling met score op de grens precies op een 5,5 uitkomt.

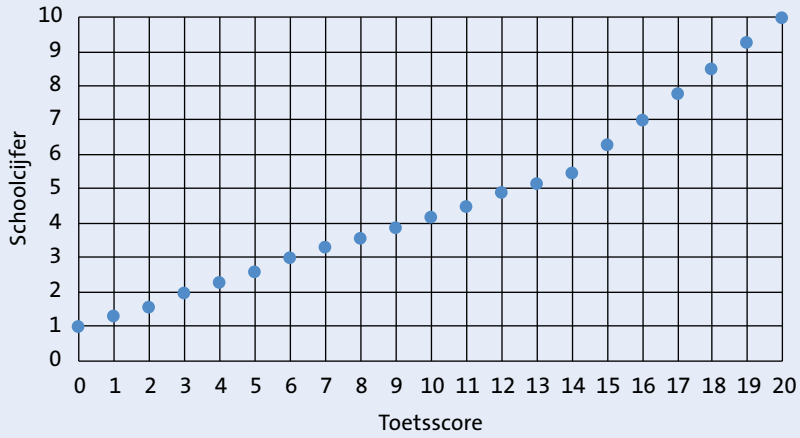
In het voorbeeld ligt de grensscore bij 14, Dit betekent dat je in 20 – 14 = 6 stappen van 10 naar 5,5 loopt. Daaruit volgt dat bij elk niet behaald punt 0.75 cijferpunten afgetrokken wordt. Bij een meerkeuzetoets komt deze methode dus neer op de regel “bij elk fout antwoord gaan er XYZ punten af”. Deze trend wordt doorgetrokken tot je bij het cijfer 1 uitkomt. Daarna wordt aan elke score die nog lager is het cijfer 1 toegekend omdat we geen cijfers onder de 1 willen geven.

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

## 4.3 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 per extra scorepunt 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 ook 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, en samen vormen die een functie met knik. De bijbehorende formules zijn:

Voor scores < grensscore:

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

Voor scores ≥ grensscore:

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

Als we de formules toepassen op ons eerder genoemde voorbeeld, dan zou een leerling met 18 scorepunten ook weer een 8.5 krijgen:

$$\text{cijfer} = 10 - (20 - 18) \times \frac{4{,}5}{20 - 14} = 8.5.$$

Echter een leerling met 6 punten zou nu een 2.9 krijgen:

$$\text{cijfer} = 1 + 6 \times \frac{4{,}5}{14} = 2.9$$

Net zoals de vorige methode voldoet ook deze methode aan de basisuitgangspunten.
De methode heeft als bijkomende voordeel dat in het algemeen geldt 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 punten-waardering hangt af van of je geslaagd bent of niet. Met name voor studenten rondom de grensscore kan dat als oneerlijk worden ervaren. Een methode dat aan dit bezwaar tegemoet komt is de methode die door het CvTE is ontwikkeld en 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.

## 4.4 De Methode van het College voor Toetsen en Examens (CvTE)

Deze methode is ontwikkeld door het college voor toetsen en examens (CvTE) en wordt onder gebruikt bij de centrale eindexamens. De methode is gebaseerd op de volgende 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:

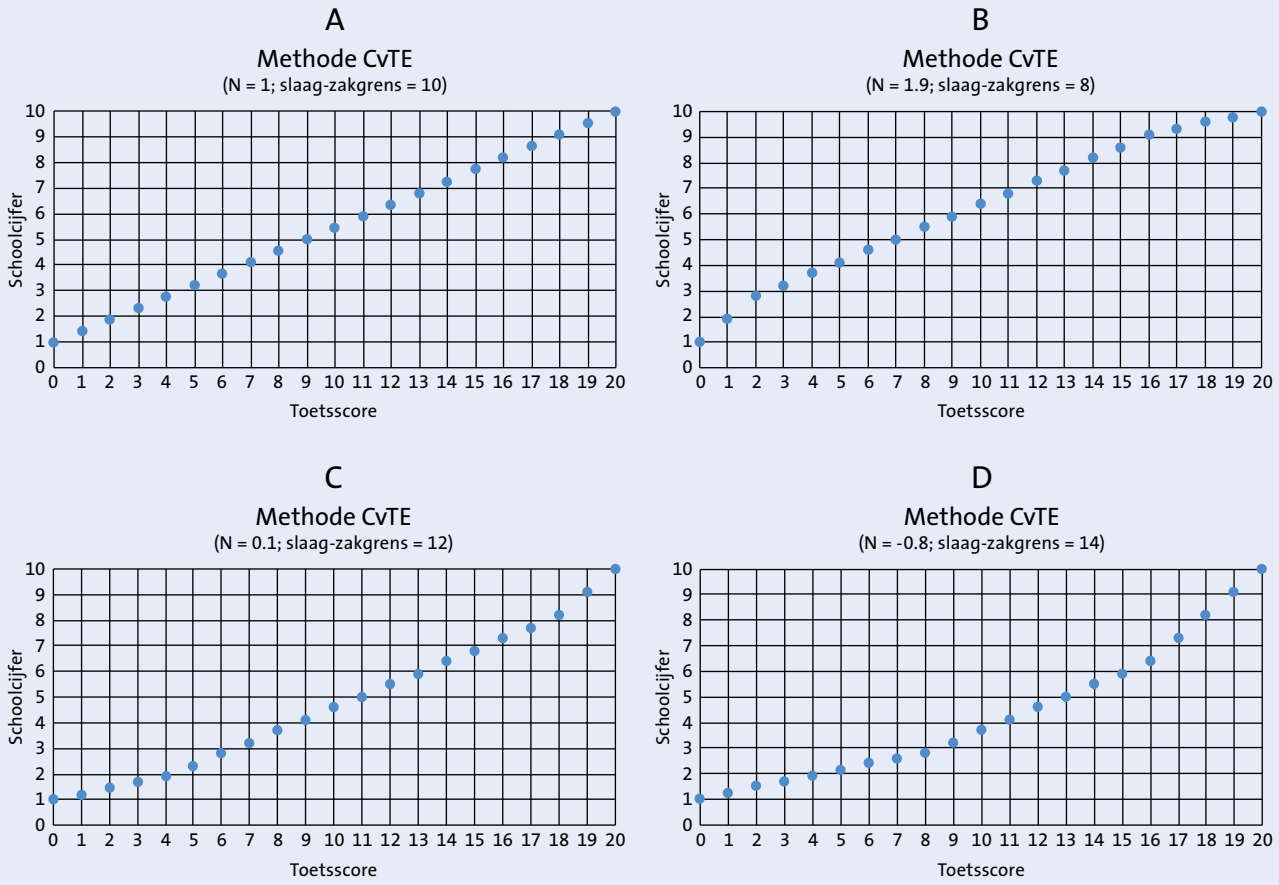
$$\text{Cijfer} = \frac{\text{behaalde score}}{\text{maximale score}} \times 9 + N.$$

Hierin is N de “N-term”. De keuze voor de N-term is aan de docent of de toetsende instantie (bijv. het CvTE). Deze N-term hangt af van de moeilijkheid van de toets, in combinatie met de gestelde onderwijskundige eindtermen. De N-term is bovenstaande formule de schakel tussen de slaag- zakgrens en het cijfer 5.5. Bijvoorbeeld, 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. Bij een moeilijke toets wordt doorgaans een hogere N term gekozen. Dit betekent dat je minder punten nodig hebt voor een 5,5. Een de toets die relatief gemakkelijk is krijgt een lagere N-term.

We zullen nu eerst voor een aantal N-termen de scoreomzetting bekijken, daarna zullen we uitgebreider stil staan bij het bepalen van de N-term zelf. In figuur 4 zie je voor verschillende N-termen de omzetting voor een fictieve toets waarbij maximaal 20 punten te behalen zijn. We beginnen met de situatie van N = 1 (Figuur 4a). In dit geval kun je de bovenstaande basisformule gebruiken om voor alle mogelijke scores het bijbehorende cijfer te berekenen. Alle cijfers liggen dan netjes tussen 1 en 10, en er is een rechtlijnig verband tussen de scores en cijfers.

Figuur 4B laat de omzetting zien voor N = 1,9. Dit betekent dat de toets relatief moeilijk is. Zou je nu de basisformule toepassen, dan zou een leerling die alles goed heeft een 10,9 halen. Dat is tegen de afgesproken regels in. Om er toch voor te zorgen dat alle cijfers tussen 1 en 10 liggen is er een speciale omrekeningsprocedure ontwikkeld. Het voert te ver om het in detail te bespreken.[8] We zullen ons daarom beperken tot enkele voorbeelden. Als we figuur 4b vergelijken met 4a dan zien we de cijfers wat hoger liggen dan wanneer N = 1. De lijn is als het ware wat naar boven geschoven. Om er voor te zorgen dat alle cijfers tussen 1 en 10 liggen, zien we twee knikjes aan de uiteindes. Deze knikjes zijn het resultaat van het achterliggende algoritme dat gebruikt is.

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




In Figuren 4C en 4D zien we twee verschillende omzettingsgrafieken 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

8 De geïnteresseerde lezer wordt verwezen naar de bijbehorende publicatie in de Staatscourant. https://zoek.officielebekendmakingen.nl/stcrt-2019-9325.html.

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.

De CvTE methode is ontwikkeld om het principe dat de cijfers evenredig mee moeten stijgen met het aantal punten zoveel mogelijk recht te doen. Als je kijkt naar de figuren dan zie je dat dat zeker het geval is rondom de slaag-zakgrens. Dit betekent dat rondom de slaag-zakgrens alle punten evenveel waard zijn. Dit zorgt er mede voor dat leerlingen die juist rondom de grensscore zitten niet onterecht het gevoel hebben dat ze benadeeld worden omdat dat ene punt tekort waardoor zij zakken minder waard zou zijn als dat ene puntje extra boven de grens.

Wil je de CvTE methode gebruiken, dan heb je de N-term nodig. Voor de centrale eindexamens worden de N-termen vastgesteld door het CvTE, waarbij in de wet is vastgelegd de gekozen N-term altijd tussen 0 en 2 moet liggen.[9] Aan de keuze van N liggen allerlei inhoudelijke overwegingen en ingewikkelde psychometrische analyses van de toetsresultaten en proef-afnames ten grondslag, Het doel is om de N-term zo te kiezen dat in opeenvolgende jaren aan kandidaten dezelfde eisen worden gesteld. Een leerling die bijvoorbeeld in 2017 eindexamen doet heeft dan evenveel kans om te zakken of te slagen als een even vaardige leerling die in een ander jaar examen doet. Deze manier van het bepalen van N wordt ook normhandhavings-onderzoek genoemd (Eggen en Sanders, 2013). Op de website van Stichting Cito vind je een Excelbestand om scores om te zetten naar een schoolcijfer als de N-term bekend is.

Voor veel schooltoetsen, inclusief de schoolexamens, is er vaak helemaal geen N-term bekend. Wanneer je toch de CvTE omzettingsmethode wil toepassen dan kan dat ook via de slaag-zakgrens. Op de website van Stichting Cito staat een handig hulpmiddel waarmee je een omzettingstabel kunt genereren op basis van een slaag- zakgrens. Je hebt dan geen N-term nodig. Er is wel één beperking. De slag- zakgrens moet namelijk 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 slaag- zakgrens 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, want als de slaag- zakgrens buiten deze range valt dan is er sprake van een wel hele makkelijke of hele moeilijke toets. Mocht je toch een slaag-zakgrens willen hanteren die buiten de range valt, dan kun je het beste de lineaire methode met knik gebruiken.

9 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.

# 5 Geraadpleegde literatuur


Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.

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

Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie: Inleiding in de theorie van de psychologische test en zijn toepassingen*. Houten: Bohn Stafleu & Van Loghum.

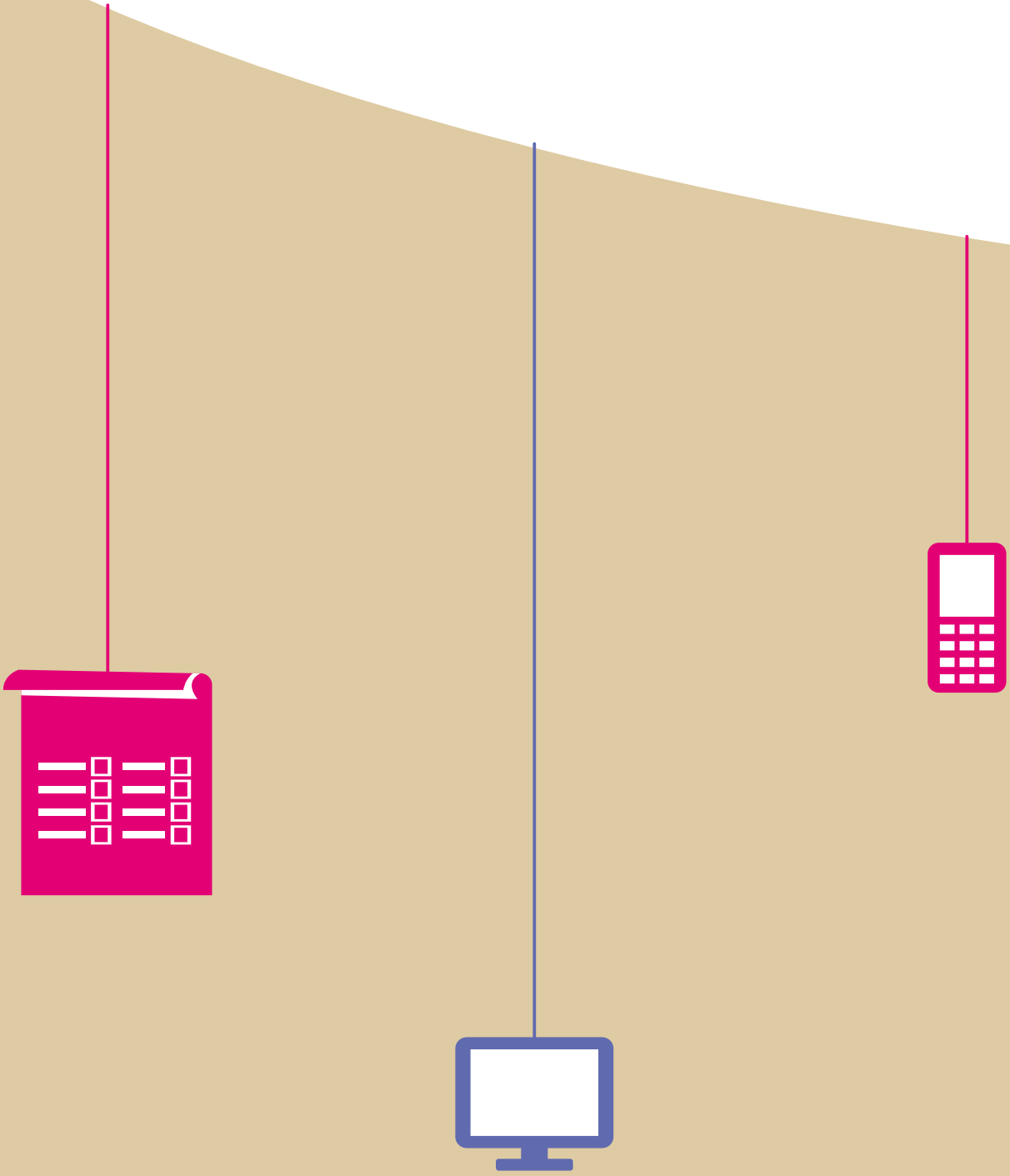
Eggen, T. J. H. M., & Sanders, P. (1993). *Psychometrie in de praktijk*. Arnhem, NL: Stichting Cito.

Groot, A. D. de (1966). *Vijven en zessen*. Groningen: Wolters.

Rooy, P. de (2018). *Een geschiedenis van het onderwijs in het Nederland*. Amsterdam: Wereldbibliotheek.

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 ]

Sanders, P.F. & Verstralen, H.F.M (2017). *Het beoordelen van toetsscores*. In: P.F. Sanders (Red.) Toetsen op School. Arnhem: Cito.



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