



Is een moeilijk examen hetzelfde als een té moeilijk examen? Een terugblik op het centraal eindexamen Engels gl/tl 2017

In 2017 deden in Nederland 61.871 leerlingen het examen Engels vmbo gl/tl. Er werd dit jaar veel geklaagd over dat examen: bij het LAKS, via de Examenlijn, in de media. Het examen was volgens velen veel te moeilijk. Wat bezielde de examenmakers bij het maken van deze toets!

Die examenmakers, een team van ervaren docenten, zijn zich zeer bewust van de impact van hun product op de leerlingen. De basis van het examen is de constructieopdracht van het College voor Toetsen en Examen (CvTE). Het examen moet op objectieve wijze toetsen of de eindexamenleerling het beoogde minimale leesvaardigheidsniveau heeft, voor Engels gl/tl vooral B1, maar ook deels B2 (zie <bit.ly/ltm-engsyl>).

De examens van de verschillende jaren zijn niet identiek. Zo werd in 2017 de formele briefopdracht afgeschaft. Daarnaast was het examen van 2017 moeilijker dan gemiddeld. Toch haalden de gl/tl-leerlingen dit jaar gemiddeld 28 van de 46 te behalen punten, waar 21 punten al genoeg was voor een 5,5. De verdeling over de scoreschaal was evenwichtig (zie figuur): er waren, ook in dit moeilijke exa-

men, genoeg opgaven die door het merendeel van de kandidaten goed beantwoord konden worden.

Scores worden omgezet naar cijfers middels de N-term. Die N-term kan en mag variëren van 0 tot 2. Er wordt geen ‘ophoging’ gegeven om eindexamenkandidaten aan een mooi cijfer te helpen of docenten tevreden te houden. De N-term moet ervoor zorgen dat de cijfers van examens van verschillende jaren eenzelfde waarde vertegenwoordigen. Hij wordt berekend met gebruik van de gegevens van statistische analyses van de scores, post-tests, standard-settingen en meerjarige vergelijkingen. Alleen kijken naar de gemiddelde score (of de gemiddelde p-waarde, die ook het verschil in scorepunten tussen de verschillende jaren meeneemt) is te simpel: de examens zijn niet elk jaar even moeilijk.

Onderzoek toont voor Engels een significante vaardigheidsstijging aan. Als we de N-term bepalen vanuit een vast percentage onvoldoendes, dan zou je deze generatie leerlingen dus afstraffen voor het feit dat ze gemiddeld beter zijn geworden. De hoge gemiddelde cijfers en het lage percentage onvoldoendes van de laatste jaren laten zien dat het examen dat niet doet: in 2017 gemiddeld 6,8 bij 18,2 procent onvoldoendes.

Een aantal jaren terug was het percentage onvoldoendes standaard hoger dan 25 procent en lag het gemiddelde cijfer rond 6,3.

Een vaardigheidsstijging heeft ook consequenties voor de toetsamenstelling. Als een examen te gemakkelijk is voor de doelgroep, maakt het geen onderscheid meer tussen vaardige en minder vaardige leerlingen, met name aan de bovenkant van de verdeling (plafondeffect). In de praktijk betekent dit dus dat er ook opgaven in moeten zitten die slechts door een klein deel van de kandidaten goed gemaakt worden. Als dit tevens de sterkste leerlingen zijn, dan doen deze opgaven precies wat ze moeten doen. Deze vragen zijn statistisch herkenbaar aan een hoge rit-waarde (item-totaalcorrelatie). Een rit van .20 is al ruim voldoende, een rit van boven de .30 geldt als hoog. Opgave 34, overigens niet de moeilijkste opgave in het examen, haalde een rit-waarde van .64 (zie voor de statistische informatie: <bit.ly/ltm-rit>).

De statistiek laat ons ook zien dat het bijvoorbeeld geen probleem hoeft te zijn als in de tekst een moeilijk woord zoals *conduct* staat: de vraag die bij de passage hoorde is door 85 procent van de leerlingen juist beantwoord. De vervanging

van *behaviour* door *conduct* was overigens ingegeven door de wens om het gemakkelijkere woord te kunnen gebruiken in een vraag. In de tekst stond genoeg context om te kunnen begrijpen wat *conduct* betekent.

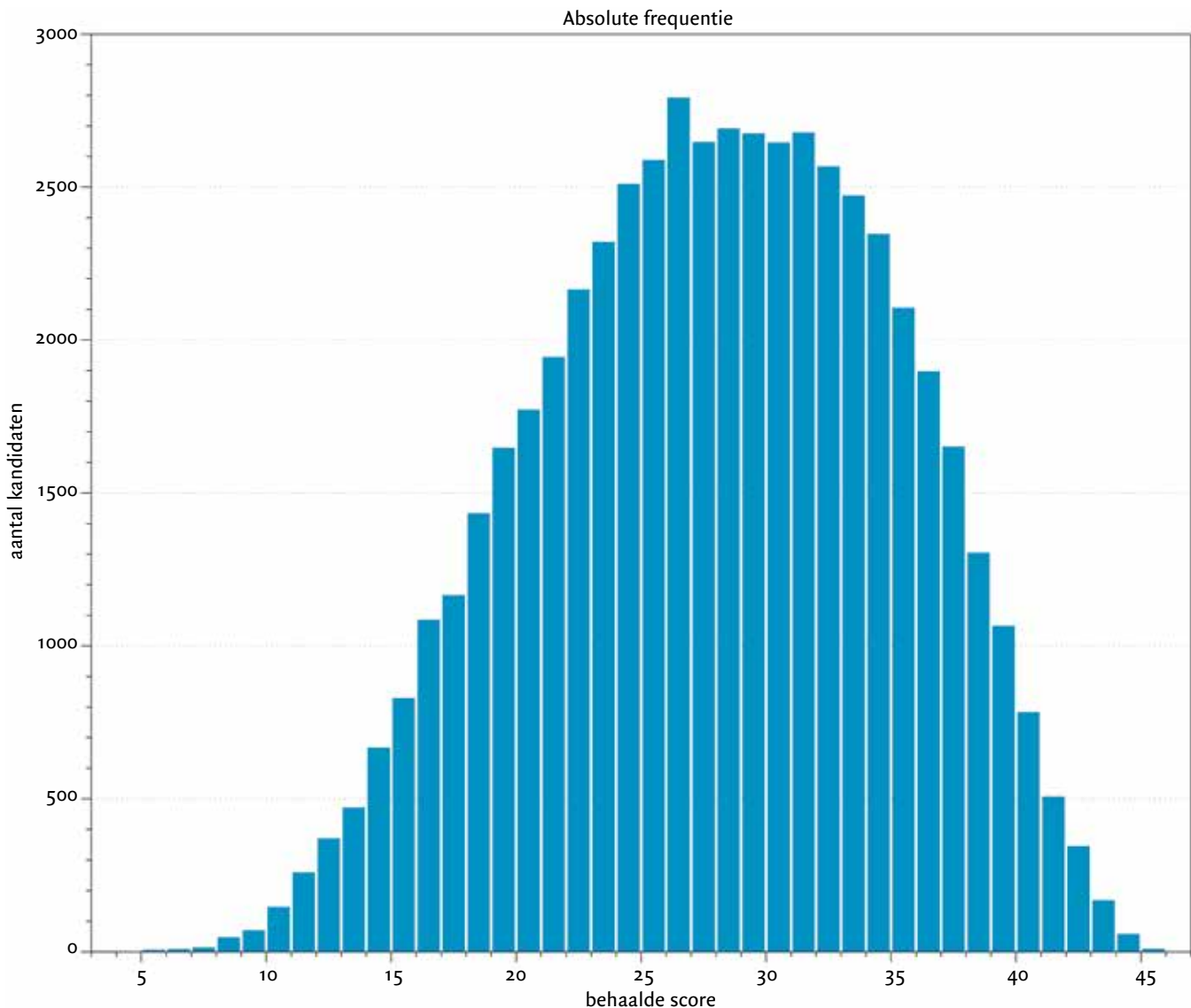
De huidige generatie leerlingen, opgegroeid met internet en games, is blootgesteld aan een ander en vooral ook uitgebreider idioom dan de populaties voor hen. Dit gegeven is, naast de door scholen getroffen maatregelen naar aanleiding van de kernvakkenregeling (strengere selectie en overgangseisen, gerichte

trainingen et cetera), waarschijnlijk ook de reden voor de stijging in de leesvaardigheid die uit alle meerjarige vergelijkingen en andere controles naar voren komt.

Het maken en samenstellen van een examen is een complex proces. Er hangt veel van de resultaten af, maar tegelijk is er amper de mogelijkheid om de teksten en opgaven van tevoren te testen. Niet iedere vraag werkt uiteindelijk in de praktijk zoals van tevoren bedacht en gehoopt. Tijdens de correctieperiode wordt dan ook iedere opmerking, kanttekening, vraag en klacht zorgvul-

dig bekeken en, indien nodig, wordt het correctievoorschrift uitgebreid of aangepast. Met mogelijk minder effectieve opgaven wordt in de systematiek van toetsanalyse, die onder andere de basis voor de formele normering is, rekening gehouden. Er wordt veel moeite gedaan om te waarborgen dat de kandidaten van de verschillende jaren bij vergelijkbare prestaties gelijke cijfers krijgen, ook al hebben de examens niet elk jaar exact dezelfde moeilijkheidsgraad. ■

Natalie Schols, toetsdeskundige Cito



De scoreverdeling van het centraal eindexamen Engels gl/tl 2017-1