Beoordelingsruimte en cijfermanipulatie in eindexamens voortgezet onderwijs

EXECUTIVE SUMMARY

September 2018

Ilja Cornelisz
Chris van Klaveren

Amsterdam Center for Learning Analytics (VU-ACLA)
Vrije Universiteit Amsterdam

Inhoud

1. Aanleiding

Onlangs is er in de Verenigde Staten een onderzoek uitgevoerd over mogelijke cijfermanipulatie door docenten op gestandaardiseerde eindtoetsen in het voortgezet onderwijs (Dee et al., 2016). De bevindingen wezen uit (1) dat leerlingen stelselmatig werden bevoordeeld door docenten, zodat zij alsnog net aan een voldoende behalen, en (2) dat deze stelselmatige bevoordeling samenhangt met achtergrondkenmerken van leerlingen (etnische achtergrond en sociaaleconomische status). Zij merken tevens op dat deze bevindingen mogelijk maatschappelijk zeer onwenselijk zijn. Het eindexamen zou een valide en representatieve toets moeten zijn die niet aan cijfermanipulatie onderhevig is. Bewuste of onbewuste subjectiviteit in de beoordeling door docenten kan daarmee leiden tot een belangrijke mate van ongewenste kansenongelijkheid in het onderwijssysteem.

In Zweden is eveneens onderzoek gedaan naar de mogelijke invloed van docenten in het beoordelingsproces van gestandaardiseerde toetsen is uitgevoerd in Zweden (Diamond & Persson, 2016). Ook dit onderzoek, wat zich richt op de overgang van primair naar voortgezet onderwijs, vindt dat cijfermanipulatie door docenten wijdverspreid is. De resultaten laten zien dat vooral leerlingen die op de gestandaardiseerde toets lager scoren dan op de schooltoetsen bevoordeeld worden (zgn. "*bad test day effect*"). Ook hangt de kans om bevoordeeld te worden samen met het opleidingsniveau en inkomen van de ouders. Naast de genoemde samenhang met inkomen en opleidingsniveau van de ouders worden er ook grote verschillen tussen schooltypen en regio's gevonden. Dit alles draagt bij aan de conclusie dat dergelijke bevoordeling, ook in de Zweedse context, kansenongelijkheid verder in de hand werkt.

De onderzoekers van VU-ACLA hebben naar aanleiding van bovenstaande geanalyseerd of een dergelijke stelselmatige bevoordeling zich ook voordoet in het Nederlands voortgezet onderwijs (VO). Hierbij is gebruikgemaakt van de geregistreerde resultaten voor schoolexamens (SE) en centraal schriftelijk eindexamens (CSE) van alle eindexamenkandidaten VMBO, Havo en VWO in de periode 2007-2012. Het volledige onderzoek is beschikbaar als wetenschappelijk artikel (Cornelisz et al., 2018). Hieronder worden de belangrijkste kenmerken, resultaten en implicaties samengevat.

2. Onderzoeksopzet

Met bovenstaande resultaten als vertrekpunt hebben onderzoekers van VU-ACLA geanalyseerd of dergelijke stelselmatige bevooroordeling zich ook voordoet in het Nederlands voortgezet onderwijs (VO). Uniek aan de situatie in Nederland is de mogelijkheid om voor eindexamenkandidaten om voor één vak een herkansing te doen, wat veelal gebeurt op het moment dat een leerling nog niet geslaagd is op basis van de resultaten van de eerste poging. Daarnaast geldt de zogeheten N-term een belangrijke rol in dat het voor leraren pas mogelijk is om exact te bepalen wat het eindcijfer is van een leerling zodra deze normeringsterm bekend is (Figuur A). Dit onderzoek richt zich expliciet op deze twee eigenschappen van het Nederlandse eindexamensysteem.

Om correcte conclusies te kunnen ontlenen aan de hand van administratieve data is allereerst een model opgesteld waaruit kan worden afgeleid wanneer, en in welke mate, een leraar de beschikbare beoordelingsruimte in zal zetten om zodoende het behaalde cijfer te beïnvloeden. Uit dit model volgt een viertal conclusies:

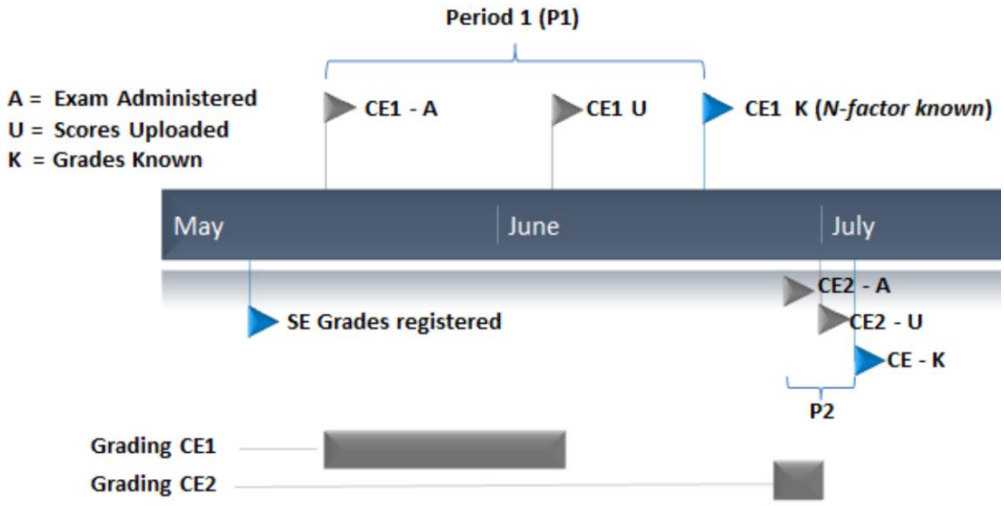
1. Een leraar zal enkel de beschikbare beoordelingsruimte benutten wanneer dit betekent dat een leerling als gevolg hiervan zal slagen in plaats van zakken.

2. Een leraar zal het cijfer niet verder beïnvloeden dan strikt noodzakelijk voor het laten slagen van een leerling.

3. Een leraar zal enkel de beschikbare beoordelingsruimte effectief kunnen inzetten wanneer er volledige informatie is met betrekking tot de normeringsterm (N-term).

4. De mate waarin de leraar het cijfer zal beïnvloeden neemt toe naarmate de beoordelingsruimte toeneemt.

Op basis van bovenstaande voorspellingen spitsen de analyses zich toe op eindexamenkandidaten die een herkansing nodig hebben om alsnog te slagen voor het eindexamen (conclusie 1 + 3), wordt gekeken of er voor deze leerlingen een discontinuïteit waarneembaar is bij het cijfer 5,5 in de uiteindelijke cijferverdeling (conclusie 2) en wordt onderzocht of de mate van overheveling van zakken naar slagen samenhangt met de mate van beoordelingsruimte op het herexamen (conclusie 4).



Figuur A: *Tijdlijn examenactiviteiten voortgezet onderwijs – tijdvak 1 + 2*

3. Data en onderzoeksmethode

In het onderzoek is gebruikgemaakt van geregistreerde resultaten voor schoolexamens (SE) en centraal schriftelijk eindexamens (CSE) van alle eindexamenkandidaten VMBO, Havo en VWO in de periode 2007-2012. Deze data is geanalyseerd op zogeheten Microdata-bestanden binnen de beveiligde remote-access (RA) omgeving van het Centraal Bureau voor de Statistiek (CBS). In totaal worden de resultaten van 1.12 miljoen leerlingen geanalyseerd (zie Tabel A).

Tabel A: *Beschrijvende statistieken examenpopulatie*

	Range	Mean	SD
Male	{0,1}	0.50	0.50
Age on October 1^{st}	{12,24}	16.08	0.96
Migrant Background	{0,1}	0.20	0.40
Impoverished neighborhood	{0,1}	0.13	0.34
SE	{0,10}	6.52	0.84
Pre-vocational education	{0,1}	0.55	0.50
Upper general education	{0,1}	0.26	0.44
Pre-university education	{0,1}	0.17	0.38
Number of students		1118650	

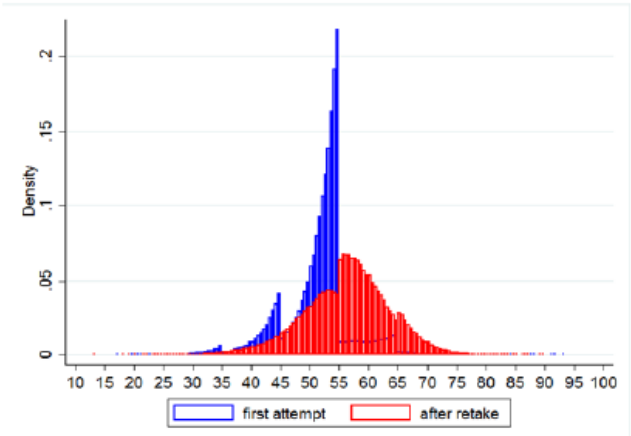
Voor het fenomeen cijferbeïnvloeding door leraren zal gekeken worden naar de subpopulatie herkansers (N=253.796). Binnen de groep herkansers is op basis van de slaagzak-regelingen van 2007-2012 voor iedere leerling een inschatting gemaakt of de herkansing nodig was om uiteindelijk alsnog te slagen voor het eindexamen VMBO, Havo of VWO. Dit heeft geresulteerd in een totaal van 136.698 studenten, waarvoor de volgende analyses zijn uitgevoerd:

i. Een vergelijking van de cijferverdeling voor en na de herkansing.
ii. Berekening van het aantal leerlingen dat geslaagd is als gevolg van het herexamen.
iii. De mate van overheveling en de beschikbare beoordelingsruimte voor de docent.
iv. Verschillen tussen en binnen scholen in de openbaring van cijferbeïnvloeding.

Om een geschikte maat te hebben voor de mate van beoordelingsruimte is voor alle herexamens 2007-2012 nagegaan hoeveel procent van de vragen geen multiple choice (MC) waren (bijv. open vragen, essay-vragen) en dus door de docent op de mate van correctheid beoordeeld diende te worden. Dit varieert sterk tussen herexamens en vakken, maar gemiddeld kent circa 70% enige mate van beoordelingsruimte voor de docent.

4. Resultaten

In Figuur B wordt de cijferverdeling voor en na de herkansing beschouwd. Hieruit blijkt dat een aanzienlijk deel van de leerlingen zich verbetert op het herexamen en dat er een scherpe discontinuïteit ontstaat bij het al dan niet voldoende afsluiten van een examenvak (i.e. bij een 5,5). Gegeven de details van de slaagzakregeling is het verschil tussen het wel/niet voldoende afsluiten van een examenvak vaak van doorslaggevend beland voor het al dan niet behalen van het diploma.



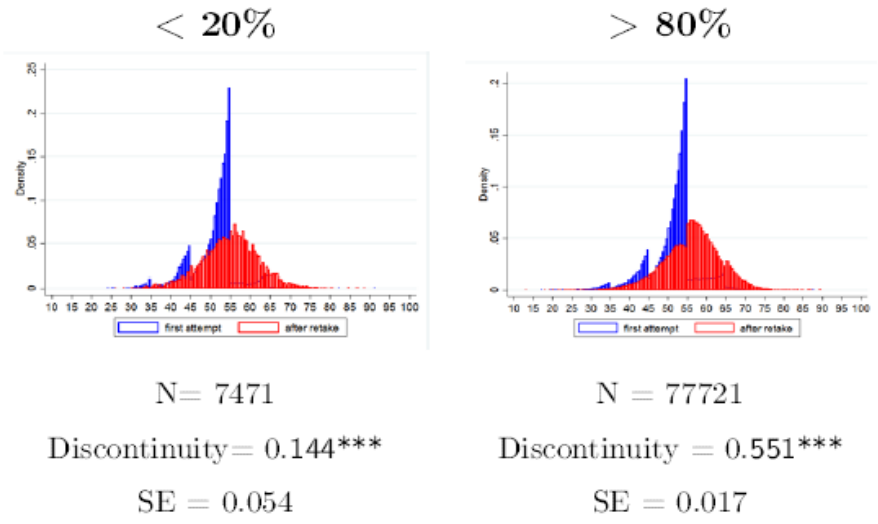
N= 136698

Discontinuity= 0.439***

SE = 0.017

Figuur B: *Cijferverdeling voor en na herkansing*

Uit de analyses blijkt dat van de herkansers voor wie de herkansing cruciaal was om alsnog te slagen ruim de helft het diploma behaalt als gevolg van het herexamen. Figuur C laat zien dat de overheveling van zak- naar slaag-status heel anders is voor herexamens die verschillen in de proportie niet-MC vragen (i.e. <20% versus >80%).



Figuur C: *Cijferverdeling voor en na herkansing – weinig versus veel beoordelingsruimte*

Aanvullende kwantitatieve analyses laten zien dat de geobserveerde prestatieverbetering samenhangt met de mate van beoordelingsruimte en niet verklaard kan worden door andere factoren (zie voor een uitgebreide discussie (Cornelisz et al., 2018).

Door de gevonden relatie tussen beoordelingsruimte en prestatieverbetering te modelleren blijkt dat een significant gedeelte van de herkansers slaagt als direct gevolg van de leraar die de beschikbare beoordelingsruimte benut om het uiteindelijke cijfer te beïnvloeden, dat dit fenomeen sterk verschilt tussen scholen én dat het in beperkte mate afhangt van achtergrondkenmerken van de student.

De belangrijkste resultaten zijn als volgt samen te vatten:

- Leraren zetten beoordelingsruimte gericht in om leerlingen alsnog te laten slagen.
- Leerlingen worden vaker overgeheveld als de leraar meer beoordelingsruimte heeft.
- Van de herkansers die alsnog slagen is 12% geslaagd dankzij cijferbeïnvloeding door de leraar.
- 56% van de jaarlijkse verschillen in overhevelingspercentages zijn structurele verschillen tussen schoolvestigingen.
- Leerlingen die eerste poging doen in tijdvak 2 is een selectieve groep (0.6%) die relatief meer risico-leerlingen bevat (leerlingen met een relatief lagere slaagkans).
- 51% van de herkansers slagen alsnog als gevolg van de herkansing.

5. Implicaties en Aanbevelingen

De resultaten van het onderzoek laten zien dat leraren structureel gebruik maken van de beschikbare beoordelingsruimte om leerlingen te laten slagen. Deze beoordelingsruimte wordt zeer gericht ingezet bij het nakijken van herkansingen voor leerlingen die zich qua prestatie rondom de slaag/zak-grens begeven. Het gevolg is dat een significant aantal leerlingen alsnog slaagt als gevolg van deze beïnvloeding door de leraar. Concreet slagen per jaar zo'n 1400 leerlingen (12%) voor het examendoor benutting van de beoordelingsruimte van docenten. Los van de mogelijk goede bedoelingen van de leraar die hieraan ten grondslag ligt is het in termen van mogelijke kansenongelijkheid een zorgelijke bevinding dat de beoordelingsruimte verschilt tussen herexamens.

Daarnaast is gevonden dat schoolvestigingen structureel veel/weinig gebruik maken van deze ruimte tot cijferbeïnvloeding én dat binnen scholen de relatie tussen beoordelingsruimte en geobserveerde overheveling –zij het zeer marginaal- samenhangt met wel of niet-Nederlandse achtergrond van de leerling. Aldus ontstaat er een situatie dat leerlingen van gelijke prestatie een afwijkende kans op slagen hebben die afhankelijk is van de mate van beoordelingsruimte op het herexamen, van de gemaakte schoolkeuze en mogelijk ook met achtergrondkenmerken van de leerling.

Vervolgonderzoek zou zich dus kunnen richten op de onderliggende determinanten van de gevonden schoolverschillen én de mogelijke samenhang van cijferbeïnvloeding met meer gedetailleerde leerlingkenmerken of -gedrag (bijv. absenteïsme). Daarnaast verdient het vervolgonderzoek om nauwkeurig uit te zoeken of leerlingen die als gevolg van de –waarschijnlijk goedbedoelde- hulp van leraren geslaagd zijn hier ook bij gebaat zijn. Daarom zal in een vervolgonderzoek worden nagegaan wat de lange termijn gevolgen zijn van bijvoorbeeld door te onderzoeken wat de gevolgen zijn van de kansenongelijkheid op uitval en prestaties in het vervolgonderwijs.

Een mogelijke beleidsimplicatie van deze bevindingen is dat het laten nakijken van examens door de eigen vakdocent - die weet welke uitslag leidt tot slagen - leidt tot ongewenste kansengelijkheid. Daar komt bij dat dit onderzoek zich vooral heeft gericht op cijferbeïnvloeding ten tijden van het herexamen. Echter de totale omvang van mogelijke docentafhankelijke diplomering zal in werkelijkheid nog groter zijn dan de resultaten in dit onderzoek, wanneer wordt erkend dat een vergelijkbaar fenomeen zich voor kan doen bij de schoolexamens, dat een selectieve groep van leerlingen de eerste poging van het eindexamen maakt in tijdvak 2 (wanneer de normeringsterm al bekend is) én dat er ook een groep leerlingen zal zijn die het diploma net aan heeft behaald als gevolg van de mogelijke benutting van beoordelingsruimte van centraal schriftelijk eindexamens gemaakt in tijdvak 1.

Referenties

Cornelisz, I. Meeter, M., & van Klaveren, C. (2018) "*Teacher Discretion in Grading High-Stakes Exams: The case of Dutch Secondary Education*". Manuscript

Dee, T. S., Dobbie, W., Jacob, B. A., & Rockoff, J. (2016). *The causes and consequences of test score manipulation: Evidence from the new york regents examinations* (No. w22165). National Bureau of Economic Research.

Diamond, R., & Persson, P. (2016). *The long-term consequences of teacher discretion in grading of high-stakes tests* (No. w22207). National Bureau of Economic Research.