**‘Compensatoir toetsen komt kwaliteit hoger onderwijs niet ten goede’**

Opinie | door Niels Smits & Michael S. Merry

**23 januari 2024** | In de afgelopen tijd is compensatoir toetsen in het Nederlandse onderwijs een populaire toetsvorm geworden. Hierbij slaagt een student voor een reeks toetsen indien het gemiddelde op deze reeks voldoende is. In deze analyse betogen twee UvA-onderzoekers dat het compensatoire systeem echter zowel onverdedigbaar als moreel gezien onverantwoordelijk is. Het grote gevaar van compensatoire toetsregimes is namelijk dat het hiaten in kennis en vaardigheden tolereert, wat de validiteit van een diploma ondermijnt en bovendien medeburgers in gevaar brengt.

Beeld: Nguyen Dang Hoang Nhu

**Inleiding**

Lange tijd was *conjunctief* toetsen, waarbij voor elke toets een voldoende moet worden behaald, het standaard toetsregime. In het Nederlandse hoger onderwijs is echter in de afgelopen tijd *compensatoir* toetsen een populaire toetsvorm geworden [1–4], waarbij men voor een reeks toetsen slaagt indien het *gemiddelde* op deze reeks voldoende is (met soms randvoorwaarden zoals: voor alle toetsen tenminste een 4) [5].

Er zijn grote verschillen in de schaal waarop compensatoir toetsen wordt ingezet. Bij de Erasmus Universiteit is bijvoorbeeld het toetsbeleid ‘Nominaal is Normaal’ doorgevoerd, waarbij alle vakken binnen een heel studiejaar kunnen worden gecompenseerd [6]. Bij de Universiteit van Amsterdam, onze eigen werkgever, heeft het Kader Toetsbeleid [4] tot recent het advies gegeven om compensatoir toetsen toe te passen bínnen vakken, waarbij cijfers op deeltoetsen konden worden gecompenseerd.

Er is al eerder kritisch geschreven over compensatoir toetsen [7], maar de toenemende populariteit en het feit dat een aantal kwesties nog niet eerder tegen het licht is gehouden nopen ons om in de pen te klimmen. We zullen laten zien waarom de vier meest genoemde argumenten niet valide zijn, dat het compensatoire systeem dientengevolge onverdedigbaar is en, bovendien, betogen dat het moreel gezien onverantwoord is.

**Betrouwbaarheid**

Het meest gehoorde argument is dat compensatoir toetsen betrouwbaarder zou zijn dan conjunctief toetsen [1,8–10]. Betrouwbaarheid betreft de mate waarin een meting vrij is van meetfouten, en in de testtheorie (de verzameling statistische modellen die de basis vormt voor de constructie en analyse van toetsen) is deze gedefinieerd als de consistentie van een meting over replicaties [11].

Meetfouten worden bij studietoetsen aan twee bronnen toegeschreven. Ten eerste de *toets*: elke studietoets is een steekproef van een beperkt aantal vragen uit het kennis- of vaardigheidsdomein, waardoor de score van een gegeven student op verschillende toetsen over hetzelfde domein zou kunnen variëren [12]. Ten tweede de *student*: door toevallige factoren als vermoeidheid en gokken zou de toetsscore kunnen afwijken van de score die op een ander afnamemoment met dezelfde toets zou zijn behaald [13].

Het argument om compensatoir te toetsen is gericht op het beperken van de invloed van meetfouten [9], maar in de redenering zitten drie tekortkomingen. Ten eerste wordt in de testtheorie inderdaad gesteld dat de impact van meetfouten kan worden verkleind door verschillende metingen te combineren tot één score, maar daarbij is het wel noodzakelijk (maar overigens niet voldoende) dat bij de verschillende metingen *hetzelfde* wordt gemeten [bijv. 11, hoofdstuk 4]. Maar dit is nu precies het probleem, want zowel in het geval van deeltoetsen bij hetzelfde vak als dat van verschillende toetsen bij afzonderlijke vakken is er sprake van de examinering van *verschillende* stof waardoor het combineren van resultaten dus niet zal leiden tot de geclaimde verhoging van de betrouwbaarheid (maar wel tot een complicatie, zie argument 3).

Ten tweede gaat het bij de keuze voor een toetsregime niet om de betrouwbaarheid van de toetsscore zelf, maar om de precisie van zak-slaagbeslissingen [14, p. 224]. Door de meetfouten in studietoetsen zijn deze beslissingen niet onfeilbaar. Een student die de stof wel beheerst, maar een onvoldoende haalt is een fout-negatief en een student die de stof niet beheerst, maar wel een voldoende haalt is een fout-positief. De kansen op zulke fouten variëren met de eisen omtrent een voldoende toetsscore en vertonen onderling een afruil [15, p. 101]: bij lagere eisen neemt het aantal fout-negatieven af maar het aantal fout-positieven toe.

*Onder een compensatoir regime kan dus in het geval van een onterechte onvoldoende een hoog cijfer op een andere toets worden gebruikt om alsnog te slagen [1], maar het stelt óók studenten met een terechte onvoldoende in staat om te slagen.* De precisie van zak-slaagbeslissingen is dus niet anders dan bij conjunctief toetsen, maar de frequenties van fout-negatieven (minder vaak) en fout-positieven (vaker) wel.

Ten derde ontbreekt een besliskundige motivering waarom de aanwezigheid van meetfouten zou moeten leiden tot een beslisregel van andere *aard* [7]. Ter vergelijking: in de natuurkunde bepaalt het meetdoel het soort beslisregel en maken meetfouten expliciet onderdeel uit van zulke regels [bijv. 16]. In de epidemiologie past men voor betere virusdetectie niet de beslisregel aan, maar probeert men een betere meting te ontwikkelen [bijv. 17].

**Verband met andere variabelen**

Een tweede argument is dat een gemiddeld toetscijfer vaak een grotere samenhang vertoont met relevante variabelen dan een enkel cijfer, waardoor er met compensatoire regels betere beslissingen zouden worden genomen dan met conjunctieve [18]. Er zijn hierbij minimaal twee kanttekeningen te plaatsen. Ten eerste is dit argument triviaal omdat er bij een gemiddelde over een *reeks* cijfers uit meer databronnen over de student wordt geput dan bij een enkel cijfer [zie, 19]. *Belangrijker is de verwijzing naar ‘relevante’ variabelen, die niet de cijfers zelf betreffen, maar externe indicatoren van studievoortgang zoals het aantal behaalde studiepunten in een volgend studiejaar.*

In de testleer zijn vier typen beslissingssituaties te onderscheiden: classificatie, plaatsing, selectie en beheersing. Bij de eerste drie situaties staat de voorspelling van een *extern* criterium centraal terwijl bij beheersing het criterium *intern* is [20, p. 401]. Aangezien de keuze voor een toetsregime betrekking heeft op beslissingen over het al dan niet machtig zijn van studieonderdelen, dat wil zeggen op beheersingssituaties, is de nadruk op het verband met andere variabelen niet op zijn plaats [21, pp. 234–236].

**Compensatoire aard van toetsen**

Een derde argument voor compensatoir toetsen is dat er binnen studietoetsen altijd al sprake zou zijn van compensatie [10,22]. Bij studietoetsen wil men uitspraken doen over de prestatie op een gedefinieerd kennis- of vaardigheidsdomein. Om tot een score te komen, worden doorgaans alle vragen even zwaar gewogen en is het *resultaat* inderdaad dat eventuele subdomeinen elkaar kunnen compenseren. De reden om dit resultaat tot *vereiste* te verheffen en bredere domeinen te definiëren ontbreekt echter helaas.

(Het principe dat vragen elkaar kunnen compenseren is sommigen een doorn in het oog [23, p. 299] en het opvoeren ervan als argument *vóór* in plaats van *tegen* een compensatoir toetsregime (waarbij ‘het domein’ over *verschillende*soorten stof wordt gedefinieerd) is vanuit logisch oogpunt absurd en lijkt op een drogreden van het type ‘Tu Quoque’ [24].).

Top of Form

Het grote gevaar van compensatoire toetsregimes is juist dat het hiaten in kennis en vaardigheden tolereert, wat de validiteit van een diploma ondermijnt. Universiteiten en hogescholen geven diploma’s uit met het signaal dat afgestudeerden de betreffende discipline in al haar facetten beheersen en waarmee ze zich in de maatschappij mogen profileren als experts. Het goedkeuren van hiaten kan grote gevolgen hebben.

Stel dat een verpleger, ingenieur of orthopedagoog in hun opleiding compensatoir zijn getoetst, waardoor sommige competenties nog in onvoldoende mate aanwezig zijn. Er is niet veel fantasie nodig om te bedenken wat voor drama’s zich dan bij, respectievelijk, het uitdelen van medicijnen aan patiënten, het ontwerpen van een brug of het uitvoeren van een interventie bij een familiecrisis zouden kunnen afspelen.

Evenzo, als we het perspectief verleggen van de maatschappij naar de student verschijnt er nog een moreel probleem. Hoe rechtvaardig is het dat van twee studenten met hetzelfde onvoldoende toetscijfer er eentje toch slaagt omdat er op een toets over andere studiestof een hoog cijfer is gehaald [25]?

**Empirisch bewijs**

Een ander gevaar is dat compensatoire toetsregimes opportunistisch gedrag uitlokken, waarmee studenten zélf actief hiaten in kennis en vaardigheden creëren [7]. Indien de student bijvoorbeeld het cijfer op een (deel)toets reeds kent, zou ze de voorbereiding op een volgende (deel)toets kunnen optimaliseren om gemiddeld te voldoen (na een laag eerste cijfer is meer inspanning nodig en na een hoog cijfer juist minder). Gegeven wat de economische wetenschap aan kennis heeft voortgebracht, is het een gerechtvaardigde bewering dat zulk gedrag vóórkomt.

Opleidingen die compensatoir toetsen hebben geïntroduceerd voerden empirische analyses uit en stelden dat er geen sprake was van kennishiaten of strategisch gedrag [26–28]. Helaas is deze claim niet gerechtigd omdat een basale methodologische fout werd gemaakt: de afwezigheid van bewijs werd als bewijs voor de afwezigheid van zulke effecten opgevoerd, terwijl er nauwelijks statistisch onderscheidingsvermogen was om ze te detecteren [29].

Ander onderzoek toont overigens wél nadelige gevolgen. Bijvoorbeeld: voor eindexamenkandidaten op het vwo werden discrepanties tussen het schoolexamen en centraal examen gekoppeld aan berekenend gedrag [30], en onder psychologiestudenten bleek dat compensatie op een eerstejaarsvak gepaard ging met slechtere prestaties op een vervolgvak [31].

**Conclusie**

In ons betoog hebben we laten zien dat de vier meest gebruikte argumenten voor compensatoir toetsen niet valide zijn, waardoor het toetsregime simpelweg niet verdedigbaar is. Het gaat echter niet alleen om academische eerlijkheid, maar ook om morele verantwoordelijkheid. Zoals we bij het derde argument reeds schetsten, kan het voor de maatschappij gevaarlijke situaties opleveren. Waarom houden opleidingen zelfgenoegzaam vast aan dit systeem terwijl het tolereren van hiaten in kennis en vaardigheden kan leiden tot falende experts?

***Literatuur***

* 1] Van Lankveld, T., & Draaijer, S. (2010). *Compensatorisch toetsen* [Interne publicatie]. Onderwijscentrum Vrije Universiteit.
* 2] Task Force Studiesucces. (2009). *Studiesucces: Rapport van de task force studiesucces* [Interne publicatie]. Task force Studiesucces, Universiteit Leiden.
* 3] Werkgroep Studiesucces. (2009). *Studiesucces aan de Universiteit van Amsterdam* [Interne publicatie]. Universitaire Commissie Onderwijs, Universiteit van Amsterdam.
* 4] *Toetsbeleid*. (2019). Universiteit van Amsterdam. <https://www.uva.nl/over-de-uva/beleid-en-regelingen/beleid/onderwijsbeleid/kader-toetsbeleid/kader-toetsbeleid.html>
* 5] Rekveld, I. J., & Starren, J. (1994). Een examenregeling zonder compensatie in het Nederlandse hoger onderwijs? Een vergelijking tussen compensatie en conjunctie. *Tijdschrift Voor Hoger Onderwijs*, *12*(4), 210–219.
* 6] Kickert, R. (2020). *Raising the bar: Higher education students’ sensitivity to the assessment policy* [Proefschrift]. Erasmus Universiteit.
* 7] Smits, N., Kelderman, H., & Hoeksma, J. B. (2015). Een vergelijking van compensatoir en conjunctief toetsen in het hoger onderwijs. *Pedagogische Studiën*, *92*(4), 150–160.
* 8] Instituut voor Psychologie, FSW. (2013). *Bacheloropleiding Psychologie.* [Studiegids 2013/2014]. Erasmus Universiteit Rotterdam.
* 9] Yocarini, I. E., Bouwmeester, S., Smeets, G., & Arends, L. R. (2018). Systematic comparison of decision accuracy of complex compensatory decision rules combining multiple tests in a higher education context. *Educational Measurement: Issues and Practice*, *37*(3), 24–39.
* 10] Ramaekers, S. (2015). *Zonder compensatie studeert er niemand af* [Blog post in Kennisdossier Hoger Onderwijs, Universiteit Utrecht]. Verkregen 20 mei 2022 op https://www.uu.nl/onderwijs/onderwijsadvies-training/publicaties/blogs/zonder-compensatie-studeert-er-niemand-af.
* 11] Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
* 12] van den Brink, W. P. (1982). *Binomiale modellen in de testleer* [Proefschrift]. Dissertatiereeks, Faculteit der Psychologie, Universiteit van Amsterdam.
* 13] Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart; Winston.
* 14] Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall.
* 15] *Standards for educational and psychological tests*. (2014). [American Psychological Association and American Educational Research Association and National Council on Measurement in Education]. American Psychological Association.
* 16] Mimbs, S. M. (2010). Conformance testing: Measurement decision rules. *NCSL International Workshop and Symposium*.
* 17] Rong, G., Zheng, Y., Chen, Y., Zhang, Y., Zhu, P., & Sawan, M. (2023). COVID-19 diagnostic methods and detection techniques. *Encyclopedia of Sensors and Biosensors*, 17.
* 18] Schmidt, H. G. (2005). *Wat is er mis met het beoordelen van studenten in het hoger onderwijs?* Presentatie tijdens het Psy-colloquium, Instituut voor Psychologie, Erasmus Universiteit Rotterdam.
* 19] Wainer, H. (1976). Estimating coefficients in linear models: It don’t make no nevermind. *Psychological Bulletin*, *83*(2), 213–217.
* 20] van den Brink, W. P., & Mellenbergh, G. J. (1998). *Testleer en testconstructie*. Boom.
* 21] De Groot, A. D. (1970). *Vijven en zessen* (Zesde editie). Wolters-Noordhoff.
* 22] Baaijens, A. M. T. (2014). *Invitational conference compensatoir toetsen*. Notulair verslag, Vrije Universiteit Amsterdam.
* 23] Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2006). A plea for new psychometric models in educational assessment. *Medical Education*, *40*(4), 296–300.
* 24] Borrelli, K. (2019). Between show-trials and utopia: A study of the tu quoque defence. *Leiden Journal of International Law*, *32*(2), 315–331.
* 25] Close, D. (2009). Fair grades. *Teaching Philosophy*, *32*(4), 361–398.
* 26] Arnold, I. J. M. (2011). Compensatorische toetsing en kwaliteit. *Tijdschrift Voor Hoger Onderwijs*, *29*(1), 31–40.
* 27] Arnold, I. (2017). Resitting or compensating a failed examination: Does it affect subsequent results? *Assessment & Evaluation in Higher Education*, *42*(7), 1103–1117.
* 28] Cohen-Schotanus, J. (1995). De praktijk van de compensatie. *Onderzoek van Onderwijs*, *24*(4), 60–62.
* 29] Dorresteijn, C. van, Kan, K. J., & Smits, N. (2023). Absence of evidence is not evidence of absence: On the limited use of regression discontinuity analysis in higher education. *Assessment & Evaluation in Higher Education*, *48*(1), 16–26. <https://doi.org/10.1080/02602938.2021.2016606>
* 30] Leferink, J. (2015). *Strategic behavior by students: A supplementary explanation for differences between marks on school examinations and central examinations in Dutch pre-university education* [Master’s thesis, Universiteit Twente; Faculteit Behavioural, Management; Social Sciences]. <http://essay.utwente.nl/68752/>
* 31] Yocarini, I. E., Bouwmeester, S., Smeets, G., & Arends, L. R. (2020). Allowing course compensation in higher education: A latent class regression analysis to evaluate performance on a follow-up course. *Assessment & Evaluation in Higher Education*, *45*(5), 728–740. <https://doi.org/10.1080/02602938.2019.1693494>

*Michael S. Merry is hoogleraar Child Development and Education bij de Universiteit van Amsterdam. Niels Smits is universitair hoofddocent Methoden en Technieken bij de Universiteit van Amsterdam.*