‘Gewogen loting 2.0: expliciet, eenvoudig en inzichtelijk’

Opinie | door Susan Niessen & Henk Kiers

1 mei 2024 | Opleidingen die de beste studenten willen selecteren maar tevens een mate van kansengelijkheid willen borgen, komen al snel uit bij gewogen loting. Het opzetten van zo’n procedure vraagt echter om een aantal beslissingen waarvan de invloed onduidelijk blijft. Onderzoekers Susan Niessen en Henk Kiers stellen daarom een ander model van gewogen loting voor: expliciet gewogen loting. Dat combineert voor elke kandidaat een selectiescore en een willekeurige score, en maakt vooraf voor opleidingen inzichtelijk hoe een verandering van het gewicht van beide scores invloed heeft op de validiteit en diversiteit van de selectieprocedure. In dit artikel laten Niessen en Kiers aan de hand van voorbeelden zien hoe dat werkt.

Nu gewogen loting weer tot de mogelijkheden behoort voor opleidingen met een numerus fixus, wordt veel bediscussieerd of en hoe gewogen loting moet worden vormgegeven.

Gewogen loting wordt vaak gekozen omdat men een compromis zoekt tussen twee conflicterende doelen: enerzijds de meest geschikte studenten aannemen (wat

‘geschikt’ dan ook is), anderzijds diversiteit en kansengelijkheid borgen. Met dat laatste bedoelt men meestal de kans op toelating tot de studie van keuze.

Het is erg lastig om beide doelen gelijktijdig te bereiken: de meest valide selectie-instrumenten (voor het voorspellen van cijfers en voortgang in de studie) laten tevens de grootste verschillen in scores op basis van sociaal-culturele achtergronden zien (Lievens et al., 2022). Als onderwijsinstellingen een middenweg willen zoeken, kan gewogen loting na inzet van valide selectie-instrumenten in theorie inderdaad een redelijk compromis bieden.

Loting via lotingscategorieën

Het is echter niet eenvoudig om een gewogen lotingsprocedure precies zo vorm te geven dat beide doelen in de gewenste mate worden geborgd. Gewogen loting wordt van oudsher uitgevoerd door kandidaten in verschillende lotingscategorieën te plaatsen. De categorie bepaalt de kans op toelating. Zo’n inrichting van de procedure vergt een hoop lastige beslissingen:

- Hoeveel categorieën worden er onderscheiden?
- Worden kandidaten uit de hoogste categorie altijd toegelaten (goed voor de validiteit, slecht voor de diversiteit), of niet?
- Worden kandidaten uit de laagste categorie altijd afgewezen (goed voor de validiteit, slecht voor de diversiteit), of niet?
- Hoe wordt de kans op toelating over de categorieën verdeeld? (relatief meer kans voor hogere categorieën is goed voor de validiteit, slecht voor de diversiteit)
- Hoe worden de grensscores voor de categorieën bepaald? Op basis van groepen van gelijke grootte (bijv. 25% – 25% – 25% – 25%)? Van vooraf bepaalde maar ongelijke grote (bijv. 25% – 50% – 25%)? Of op basis van grensscores (bijv. < 6.5, 6.6-7.5, 7.6+ op een selectie-instrument zoals een toets of een gemiddeld vo-cijfer)?
- Hoe moeten de kansen tot toelating precies verdeeld worden over de verschillende categorieën?

Idealiter worden deze keuzes gemaakt op basis van een afweging tussen het effect op enerzijds de validiteit en anderzijds de diversiteit.

Als er veel categorieën worden onderscheiden en de toelatingskans tussen categorieën sterk verschilt, is dat goed voor de validiteit, maar levert dat waarschijnlijk amper winst op wat betreft diversiteit. In sommige scenario’s had men het lotingsaspect dan misschien evengoed kunnen weglaten. Andersom is de diversiteit gebaat bij het geven van veel kans aan kandidaten in de lage(re) categorieën, maar kan dit de validiteit dermate negatief beïnvloeden dat de inhoudelijke selectietoets evengoed achterwege had kunnen blijven.

De effecten van alle mogelijke keuzes die hierboven staan zijn moeilijk concreter te maken dan hierboven beschreven. We stellen daarom een aanpak van gewogen loten voor die het veel makkelijker maakt om concrete keuzes te maken die passen bij de gewenste resultaten wat betreft validiteit en diversiteit: expliciet gewogen loten.

Een alternatief: expliciet gewogen loten

In plaats van gewogen loting door kandidaten in lotinsgscategorieën in te delen, wordt aan iedere kandidaat – ongeacht diens score op de selectietoets – middels loting willekeurig een 'random score' toegekend. Die random score wordt getrokken uit een normale verdeling met hetzelfde gemiddelde en dezelfde spreiding als de verdeling van de selectietoetsscores. Vervolgens wordt voor iedere kandidaat de eindscore berekend als het gewogen gemiddelde van de selectietoetsscore en de random score, op basis van vooraf gekozen gewichten. De *ranking* van kandidaten op grond van deze eindscore wordt dan gebruikt voor de selectie.

Deze aanpak heeft twee voordelen. Ten eerste hoeft alleen te worden bepaald hoeveel gewicht er aan de selectietoetsscore en de random score gegeven wordt; keuzes over de hoeveelheid lotingsklassen, grensscores en de verdeling van relatieve kansen zijn niet meer nodig. Ten tweede kunnen op deze manier gewichten worden gekozen die tot de gewenste balans tussen validiteit en diversiteit leiden. Dit laatste kan simpelweg worden gedaan door in gewichten zelf het relatieve belang van de selectietoets tegenover de random score, die op het principe van gelijke kansen is gebaseerd, te laten reflecteren. Als er bijvoorbeeld twee keer zoveel belang wordt gehecht aan selectie op grond van de selectiescore dan aan selectie op grond van gelijke kansen, kiest men de gewichten 2/3 vs. 1/3. Het is echter ook mogelijk om deze afweging gerichter in te regelen door gebruik te maken van Pareto-optimalisatie.

Pareto-optimalisatie

Pareto-optimalisatie is een uit *operations research* afkomstige techniek om zo goed mogelijk aan conflicterende doelen tegemoet te komen. De techniek achter Pareto-optimalisatie is enigszins ingewikkeld, maar de uitwerking wordt makkelijk inzichtelijk door gebruik te maken van een gratis beschikbare online app (Song et al., 2017).

De app heeft alleen informatie nodig over de selectieratio (welk percentage van alle kandidaten wordt aangenomen), de proportie kandidaten uit de minderheidsgroep waarvan men de kans op toelating wil bevorderen, de correlaties tussen zowel de selectie-instrumenten onderling als met de maat voor geschiktheid die men hanteert, en de mate waarin scores verschillen tussen kandidaten uit de minderheids- en meerderheidsgroep voor ieder selectie-instrument.

Het feit dat één van de selectie-instrumenten bij deze vorm van gewogen loting gewoon een random score is, maakt het extra simpel. Dat leggen we hieronder uit.

Een voorbeeld

We gebruiken een inhoudelijke selectietoets en een random score, gegenereerd zoals hierboven voorgesteld. Om de app te gebruiken, hoeven we dan alleen informatie over de selectietoets te achterhalen*. Stel dat er een selectietoets wordt gehanteerd met een hoge predictieve validiteit (*R* = .57) maar ook met gemiddelde scoreverschillen tussen kandidaten uit een relevante meerderheids- en minderheidsgroep (*d* = 0.30**). De selectieratio van de opleiding is .35 (35 procent van de kandidaten wordt aangenomen) en 20 procent van de kandidaten behoort tot een minderheidsgroep waarvan men de kans tot toelating wil bevorderen. Op basis van deze informatie kan een Pareto-optimalisatie-analyse uitgevoerd worden, bijvoorbeeld in de online app. Het resultaat op basis van dit voorbeeld staat hieronder.

Resultaat Pareto-optimalisatie-analyse

Adverse impact ratio	Predictieve validiteit	Gewicht	
		Selectiescore	Random score
1.00	0.00	0.00	1.00
0.96	0.03	0.05	0.96
0.97	0.05	0.09	0.91
0.96	0.08	0.13	0.87
0.94	0.11	0.16	0.84
0.93	0.14	0.20	0.80
0.91	0.17	0.23	0.77
0.90	0.19	0.26	0.74
0.88	0.22	0.30	0.70
0.87	0.25	0.33	0.67
0.85	0.28	0.36	0.64
0.84	0.31	0.39	0.61
0.82	0.34	0.42	0.58
0.81	0.36	0.45	0.55
0.80	0.39	0.49	0.51
0.78	0.42	0.52	0.48
0.77	0.45	0.56	0.44
0.75	0.48	0.61	0.39
0.74	0.51	0.67	0.33
0.73	0.54	0.75	0.25
0.71	0.57	1.00	0.00

Notitie: De adverse impact ratio is de kans om toegelaten te worden voor iemand uit een minderheidsgroep, gedeeld door de kans voor iemand uit een meerderheidsgroep.

Het is handig om eerst naar de onderste regel te kijken. Daar zien we het resultaat als alleen de selectiescore gebruikt wordt (en de random score gewicht 0 heeft). De predictieve validiteit is .57 (dat hadden we zo opgegeven), en we zien dat de adverse impact ratio dan 0.71 is. Iemand uit de minderheidsgroep heeft dus, bij gebruik van

alleen de selectietoets, 0.71 keer zo veel kans om aangenomen te worden als iemand uit de meerderheidsgroep.

Stel nu dat we bereid zijn om de validiteit te verlagen tot R = .48 om zo de diversiteit te bevorderen. Op basis van Pareto-optimalisatie-analyse is te zien dat de selectiescore dan een gewicht van 0.61 moet krijgen, en de random score een gewicht van 0.39. Iemand uit de minderheidsgroep heeft dan 0.75 keer zo veel kans om aangenomen te worden als iemand uit de meerderheidsgroep.

Andersom kan ook als uitgangspunt genomen worden dat iemand uit de minderheidsgroep bijvoorbeeld 0.80 (een bekend uitgangspunt in de VS) keer zo veel kans moet hebben op toelating als iemand uit de meerderheidsgroep. De selectiescore moet dan een gewicht krijgen van 0.49 en de random score een gewicht van 0.51, resulterend in een validiteit van R = .39. Of, als men de kans voor iemand uit de minderheidsgroep wil ophogen tot 0.90 keer zo groot als iemand uit de meerderheidsgroep, dan moet de weging 0.26 zijn voor de selectiescore en 0.74 voor de random score, resulterend in een validiteit van R = .19.

Conclusie

Een duidelijk voordeel van de hier voorgestelde aanpak ligt erin dat de uitgangspunten en effecten van gemaakte keuzes over de invulling van selectie erg inzichtelijk en expliciet zijn. Een klein gewicht voor een random score blijkt bijvoorbeeld weinig effect te hebben; een weging van 75 procent vs. 25 procent levert slechts een adverse impact ratio op van 0.73 (t.o.v. 0.71 bij volledige selectie). Daartegenover staat dat een forse verbetering in diversiteit gepaard gaat met een forse daling van de validiteit. Dat geldt in zowel dit (op realistische uitgangspunten gebaseerde) voorbeeld als in de meeste andere gevallen waarin twee instrumenten gecombineerd worden die ieder aan één van twee conflicterende doelen (hier: validiteit en diversiteit) voldoen.

Als laatste willen we opmerken dat de volgende twee tegenwerpingen waarschijnlijk vaak voor zullen komen:

- *We beschikken niet over de benodigde informatie voor deze aanpak*
  De nodige informatie is inderdaad vast niet bij alle onderwijsinstellingen of voor alle selectie-instrumenten exact bekend. Als men uitgaat van *evidence-based* selectie, zou de nodige informatie in ieder geval geschat moeten kunnen worden op basis van interne gegevens en/of bestaand wetenschappelijk onderzoek. Voornamelijk over scoreverschillen tussen meerderheids- en minderheidsgroepen op specifieke selectie-instrumenten is echter weinig informatie beschikbaar op basis van onderzoek uit Nederland. In

dat geval bieden resultaten uit internationaal onderzoek de enige uitkomst. Ook kan men meerdere realistische waarden uitproberen om te zien hoeveel invloed dat heeft op de adverse impact ratio, om zo de bandbreedte daarvan weer te kunnen geven. Het gebrek aan concrete informatie over scoreverschillen op selectie-instrumenten is echter niet alleen een probleem als men Pareto-optimalisatie wil toepassen om selectieprocedures vorm te geven volgens bepaalde doelstellingen; het probleem is dan alleen saillanter.

- *Er is geen consensus over de gewenste balans tussen diversiteit en validiteit*
  Het kan lastig zijn om hiervoor breed gedragen concrete keuzes en uitgangspunten te formuleren. Dat neemt niet weg dat een effectieve realisatie van een balans vereist dat de gewenste balans eerst bepaald wordt. Daarnaast is effectiviteit niet te evalueren als die niet eerst wordt gedefinieerd. Dit geldt voor iedere mogelijke keuze voor een selectieprocedure – niet alleen bij het model van gewogen loting dat wij voorstellen. Het enige alternatief is afgaan op onderbuikgevoel, wat de kans op effectief beleid niet ten goede komt.

** De informatie over de random score is direct voorhanden, want die correleert nul met het selectie-instrument en met studiesucces, en minderheid- en meerderheidsgroepen scoren hier gemiddeld even hoog op.*

*** De d staat voor Cohen's d, een algemeen gebruikte maat voor het verschil van gemiddelden op gestandaardiseerde schaal. De informatie voor het hier gegeven voorbeeld is gebaseerd op schattingen voor een selectieprocedure waarin de score op een proefstudeertoets en het gemiddelde middelbare schoolcijfer meegenomen. Als maatstaf voor validiteit is het verband met het gemiddelde cijfer in de opleiding als uitgangspunt gekozen. We nemen aan dat de proefstudeertoets en het gemiddelde middelbare schoolcijfer ieder ongeveer een correlatie van r = .50 met het gemiddelde cijfer in de opleiding hebben, en dat de scores op de proefstudeertoets en het middelbare schoolcijfers sterk samenhangen (r = .55, zie Niessen et al., 2018). Op basis van de formule voor de validiteit van samengestelde voorspellers (Murphy, 2019) geeft een optimale weging van deze twee onderdelen (hier 50/50) R = .57. Verder nemen we op basis van onderzoek aan dat het verschil in middelbare schoolcijfers tussen kandidaten uit de minderheids- en meerderheidsgroep, uitgedrukt in standaarddeviatie-eenheden, gelijk is aan d = 0.27 en voor de proefstudeertoets gelijk is aan d = 0.25 (gebaseerd op Fikrat-Wevers et al., 2023). Op basis van een formule voor Cohen's d voor samengestelde voorspellers (Sackett & Ellingson, 1997) resulteert dat in d = .30 als beide onderdelen 50/50 worden gewogen.*

Referenties

- Bobko, P., & Roth, P. L. (2004). The four-fifths rule for assessing adverse impact: An arithmetic, intuitive, and logical analysis of the rule and implications for future research and practice. In J. J. Martocchio (Ed.), *Research in personnel and human resources management, Vol. 23*, pp. 177–198). Elsevier Science/JAI Press. https://doi.org/10.1016/S0742-7301(04)23004-3
- Fikrat-Wevers, S., Stegers-Jager, K.M., Afonso, P.M. et al. (2023). Selection tools and student diversity in health professions education: A multi-site study. *Advances in Health Science Education, 28*, 1027–1052. https://doi.org/10.1007/s10459-022-10204-9

- Murphy, K. R. (2019). Understanding how and why adding valid predictors can decrease the validity of selection composites: A generalization of Sackett, Dahlke, Shewach, and Kuncel (2017). *International Journal of Selection and Assessment, 27*, 249–255. https://doi.org/10.1111/ijsa.12253
- Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2018). Admission testing for higher education: A multi-cohort study on the validity of high-fidelity curriculum-sampling tests. *PLoS ONE, 13*(6). https://doi.org/10.1371/journal.pone.0198746
- Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50*, 707–721. https://doi.org/10.1111/j.1744-6570.1997.tb00711.xSong, Q. C., Wee, S., & Newman, D. A. (2017). Diversity shrinkage: Cross-validating pareto-optimal weights to enhance diversity via hiring practices. *The Journal of applied psychology, 102*, 1636–1657. https://doi.org/10.1037/apl0000240

Susan Niessen is universitair docent bij de Basiseenheid Psychometrie & Statistiek van de faculteit Gedrags- en Maatschappijwetenschappen aan de Rijksuniversiteit Groningen.

Henk Kiers is hoogleraar Methoden en Technieken van gegevensverwerking bij de Basiseenheid Psychometrie & Statistiek van de faculteit Gedrags- en Maatschappijwetenschappen aan de Rijksuniversiteit Groningen.