Flexible regression-based norming of psychological tests

Voncken, Lieke

DOI:
10.33612/diss.124765653

Samenvatting

Introductie

In mijn proefschrift richt ik mij op methoden voor normering van psychologische tests. Psychologische tests, zoals intelligentietests, zijn veelgebruikte instrumenten voor het testen van individuen. De resultaten van deze tests worden gebruikt voor bijvoorbeeld diagnosticering en selectie. Aangezien belangrijke beslissingen voor individuen worden gebaseerd op deze testresultaten, is het belangrijk dat de tests een hoge kwaliteit hebben en dat de testresultaten betekenisvol geïnterpreteerd kunnen worden. De interpretatie van testscores is meestal gebaseerd op een referentiepunt, zoals een extern criterium (bijv. een vooraf gekozen percentage aan testitems dat correct moet zijn) of de testscores van andere testnemers voor dezelfde test (Mellenbergh, 2011, p. 346). In dit laatste geval wordt de testscore geïnterpreteerd in vergelijking met de verdeling van de scores in een referentiepopulatie. Bij veel psychologische tests wordt dit soort referentiepunt gebruikt. Daarom richt ik mij in mijn proefschrift op normen waarbij de testscores van anderen als referentiepunt worden gebruikt, bijvoorbeeld de algemene bevolking van een land.

In eerste instantie hangt de keuze van de referentiepopulatie af van de gewenste interpretatie van de testscores, omdat de interpretatie hier direct door wordt bepaald. Bijvoorbeeld, bij intelligentietests wil men normaal gesproken de behaalde testscore vergelijken met de testscores van anderen met dezelfde leeftijd. Daarnaast moet onderzocht worden of er überhaupt een relatie is tussen de ruwe testscores en de gekozen persoonlijke kenmerken. Als de testscore geen relatie heeft met de leeftijd, heeft het geen zin om normen afhankelijk van de leeftijd te hebben.

Om de normen te berekenen, heb je de scoreverdelingen in alle referentiepopulaties nodig. Deze scoreverdelingen worden geschat op basis van de scores in een normeringssteekproef. Idealiter worden de scores verzameld in een representatieve steekproef van de normpopulatie. Om vertekende normen te voorkomen, is het belangrijk dat de steekproef representatief is ten opzichte van kenmerken die samenhangen met de testscore. Dit is erg lastig in de praktijk, omdat van tevoren niet duidelijk is welke kenmerken samenhangen

met de testscores. Theoretisch gezien is de beste manier om representativiteit te verkrijgen een random steekproef uit de populatie trekken, maar praktische problemen zoals privacywetgeving en non-respons maken dit onmogelijk. Daarom wordt in de praktijk vaak data verzameld door middel van zogenaamde *judgmental sampling* (Mellenbergh, 2011, p. 351), waarbij steekproeven worden verzameld uit subpopulaties die zijn gebaseerd op gemakkelijk te meten persoonlijke karakteristieken zoals leeftijd, geslacht, opleidingsniveau, en regio, en waarvan de verdeling in de gewenste populatie bekend is (bijv. door gegevens van het Centraal Bureau voor de Statistiek). Op deze manier wordt geprobeerd om de steekproef representatief te maken op basis van deze subpopulaties. Het nadeel is dat nooit zeker is of genoeg subpopulaties worden meegenomen en dat niet wordt gecontroleerd of combinaties van kenmerken op de juiste manier worden vertegenwoordigd.

Als de referentiekarakteristieken worden gebaseerd op categorische variabelen, zoals geslacht, kan de referentiepopulatie worden gedefinieerd voor elke categorie van de variabele. Traditioneel werd hetzelfde gedaan voor continue variabelen, zoals leeftijd, door ze te discretiseren. Hierbij werd aangenomen dat de conditionele testscoreverdeling trapsgewijs verandert als een functie van de continue variabele, terwijl het theoretisch realistischer is dat deze relatie gelijkmatig verloopt (Van Breukelen & Vlaeyen, 2005; Zachary & Gorsuch, 1985). In continue testnormering (Zachary & Gorsuch, 1985; Zhu & Chen, 2011), ook wel *regression-based* normering genoemd, wordt de ruwe scoreverdeling als een continue functie van de referentiekarakteristieken geschat in een regressiemodel. Deze methode is efficiënter dan traditionele normering (Oosterhuis et al., 2016), omdat de informatie van alle observaties in de normeringssteekproef wordt gebruikt in plaats van alleen de informatie binnen een subgroep.

Er kunnen drie typen continue normeringsbenaderingen worden onderscheiden (Emons, 2019): inferentiële normering (Wechsler, 2008; Zachary & Gorsuch, 1985; Zhu & Chen, 2011), *moments regression-based* normering (Oosterhuis, 2017; Van Breukelen & Vlaeyen, 2005), en niet-parametrische normering (Lenhard et al., 2018; Tellegen & Laros, 2014).

Het voordeel van *moments regression-based* normering vergeleken met inferentiële normering en niet-parametrische normering is dat geen arbitraire subgroepen van de predictor(en) hoeven te worden gemaakt, dat de resulterende percentielcurves elkaar niet kunnen snijden – wat theoretisch gezien ook onmogelijk is –, en dat al statistische criteria

voor de selectie en beoordeling van modellen beschikbaar zijn. Bij dit type continue normering worden momenten (tot nu toe alleen het gemiddelde) van de ruwe scoreverdeling geschat als functie van de predictor(en). Tot nu toe werd binnen *moments regression-based* normering de onrealistische aanname gemaakt dat de conditionele scoreverdeling normaal verdeeld is met een constante variantie.

In dit proefschrift wordt onderzoek gedaan naar een flexibele vorm van *moments regression-based* normering, namelijk door middel van distributionele regressie (Rigby & Stasinopoulos, 2005; Umlauf et al., 2018). Hierbij kunnen meerdere kenmerken van de verdeling, zoals het gemiddelde, de variantie, de scheefheid en kurtosis, worden geschat als functie van de predictor(en). Deze benadering omvat de eerdere *moments regression-based* normeringsbenaderingen, maar maakt het mogelijk om veel meer verschillende verdelingen en typen functies te gebruiken. Hierdoor hoeft niet worden aangenomen dat de conditionele scoreverdelingen normaal verdeeld zijn met een constante variantie, terwijl de percentielcurves elkaar nog steeds niet kunnen snijden. We maken gebruik van zowel een frequentistische als Bayesiaanse aanpak, namelijk door middel van de *generalized additive models for location, scale, and shape* (GAMLSS; Rigby & Stasinopoulos, 2005) en de *Bayesian additive models for location, scale, and shape (and beyond)* (BAMLSS; Umlauf et al., 2018).

De grote beschikbaarheid van modellen maakt normering met distributionele regressie flexibel. Dit vergroot de kans op de beschikbaarheid van een passend model, maar dit heeft als nadelen dat de modelselectie lastig is en dat de gekozen modellen complex kunnen zijn. Hoe complexer het model is, hoe meer het model onderhevig is aan steekproeffluctuaties. Hierdoor is er meer onzekerheid in de geschatte genormeerde scores. Deze steekproeffluctuaties kunnen worden verlaagd door de steekproef te vergroten, maar dit is kostbaar en niet altijd mogelijk in de praktijk. In dit proefschrift onderzoeken we deze uitdagingen gerelateerd aan de modelselectie en de steekproeffluctuaties.

Hoofdstuk 2

In dit hoofdstuk wordt onderzoek gedaan naar de kwaliteit van de geschatte normen als gebruik wordt gemaakt van een geautomatiseerde modelselectieprocedure. Binnen GAMLSS kan de relatie tussen de kenmerken van de conditionele ruwe scoreverdeling en

de predictor(en) gemodelleerd worden door middel van polynomen. Hierbij moet gekozen worden welke ordes van de polynomen moeten worden meegenomen om een goede modelfit te krijgen zonder overfitting. Aangezien het aantal mogelijke modellen oneindig groot is, is het belangrijk om een goed presterende geautomatiseerde modelselectieprocedure te hebben. In een simulatiestudie vergelijken we een bestaande geautomatiseerde modelselectieprocedure uit het `gamlss` R package (Rigby & Stasinopoulos, 2005) met een door ons bedachte geautomatiseerde modelselectieprocedure voor modellen met één predictor. Hierbij kijken we naar verschillende modelselectiecriteria, namelijk kruisvalidatie en verschillende varianten van het *Generalized Akaike Information Criterion* (GAIC; Akaike, 1983). Daarnaast variëren we de complexiteit van de data, de steekproefmethode (gelijk verdeelde predictorwaardes, of meer predictorwaardes naarmate de relatie tussen de mediaan van de conditionele ruwe scoreverdeling en de predictor sterker is), en de steekproefgrootte ($N = 100, 500,$ of 1.000). Voor de geschatte modellen vergelijken we de ware percentielen vanuit het populatiemodel met de geschatte percentielen, en we kijken naar de bias en variantie in de percentielschattingen. De resultaten laten zien dat de normen het efficiëntst worden geschat met de nieuwe procedure in combinatie met één van de GAIC, ongeacht de steekproefmethode. Hoe groter de steekproefgrootte is en hoe minder complex het populatiemodel is, hoe beter de percentielen worden geschat. We laten in dit hoofdstuk ook zien hoe de twee vergeleken geautomatiseerde selectieprocedures kunnen worden gebruikt voor empirische data van de Snijders-Oomen niet-verbale intelligentietest (SON-R 6-40; Tellegen & Laros, 2014).

Hoofdstuk 3

In dit hoofdstuk wordt onderzoek gedaan naar de sensitiviteit van normschattingen voor modelflexibiliteit en steekproefgrootte. Flexibele modellen met goede modelfit in de populatie hebben een kleinere bias dan meer strikte modellen met minder goede fit, maar hebben ook een hogere mate van steekproeffluctuatie. In een simulatiestudie onderzoeken we deze *bias-variance trade-off*. We variëren systematisch de aard en mate van geschonden modelassumpties (een kleine of grote schending van lineariteit, homoscedasticiteit en/of normaliteit), de steekproefgrootte ($N = 500, 1.000,$ of 2.000), en de flexibiliteit van het schattingsmodel. Net als in Hoofdstuk 2 kijken we naar het algemene verschil tussen

de ware percentielen vanuit het populatiemodel en de geschatte percentielen, en naar de variantie en bias in percentielschattingen afzonderlijk. De resultaten laten zien dat de nadelen door het gebruik van een te strikt model (d.w.z., de toename in bias) groter waren dan de nadelen door het gebruik van een te flexibel model (d.w.z., de toename in variantie) bij data uit een niet-normaal verdeelde populatie. Het was problematisch om een model met de *skew Student t* verdeling te schatten voor data uit een normaal verdeelde populatie. We denken dat dit komt doordat distributieparameter τ theoretisch gelijk is aan ∞ voor normaliteit, wat niet mogelijk is in de praktijk. Daarom raden we aan om flexibele modellen te gebruiken, maar om een normale verdeling te gebruiken als het waarschijnlijk is dat de data uit een normaal verdeelde populatie komen.

Hoofdstuk 4

In dit hoofdstuk wordt onderzocht hoe de onzekerheid in genormeerde testscores ten gevolge van steekproeffluctuaties uitgedrukt kan worden in betrouwbaarheidsintervallen. Testuitgevers rapporteren soms al betrouwbaarheidsintervallen die de onzekerheid in normen ten gevolge van testonbetrouwbaarheid uitdrukken, maar de onzekerheid in normen door steekproeffluctuaties wordt in de praktijk genegeeerd. In een simulatiestudie beoordelen we de kwaliteit van de betrouwbaarheidsintervallen die we opstellen met de zogenaamde *posterior simulation* methode (Wood, 2006). In deze methode simuleren we sets van modelparameters aan de hand van de geschatte regressiecoëfficiënten en de bijbehorende variantie-covariantiematrix. Voor elke set van modelparameters berekenen we de bijbehorende normen en uit de verdeling van deze normen bepalen we de betrouwbaarheidsintervallen. In de beoordeling van de betrouwbaarheidsintervallen kijken we onder andere naar het percentage van de betrouwbaarheidsintervallen dat de ware genormeerde score bevat. We variëren het gebruikte populatiemodel (gebaseerd op empirische normeringsdata van de SON-R 6-40 intelligentietest (Tellegen & Laros, 2014) of de FEEST emotieherkenningstest (Voncken et al., 2018)), de methode voor het bepalen van de betrouwbaarheidsintervallen aan de hand van de gesimuleerde normverdelingen, de grootte van de betrouwbaarheidsintervallen (90% of 95%), steekproefgrootte ($N = 501, 1.001,$ of 2.001), predictorwaarde, testscore, en het type variantie-covariantiematrix. De resultaten laten zien dat dat de kwaliteit van de betrouwbaarheids-

intervallen in de meeste gevallen goed is. We illustreren de methode aan de hand van normeringsdata van de SON-R 6-40 test.

Hoofdstuk 5

In dit hoofdstuk wordt onderzocht of het meenemen van bestaande normeringsinformatie van een test (bijv. de normen van dezelfde test voor een ander land) ervoor kan zorgen dat nieuwe normen efficiënter kunnen worden geschat. Het precies schatten van de normen vereist normaal gesproken een grote normeringssteekproef. We onderzoeken door middel van Bayesiaanse Gaussische distributionele regressie in een simulatiestudie of we de vereiste steekproefgrootte voor dezelfde normprecisie kleiner kunnen maken, en hoe robuust deze methode is voor verschillen tussen de populatiemodellen van de eerdere en nieuwe normering. In een simulatiestudie variëren we het type a priori verdeling, de misspecificatie van de a priori verdeling, en de steekproefgrootte in een compleet gekruist onderzoeksontwerp. We vergelijken voor twee soorten informatieve *priors* met één zwak-informatieve *prior* in welke mate de ware percentielen afwijken van de geschatte percentielen.

De resultaten laten zien dat met één van de informatieve *priors*, de zogenaamde *fixed effects prior*, de normen efficiënter worden geschat dan met de zwak-informatieve *prior*, zolang de misspecificatie niet leeftijdsafhankelijk is. Dit laat zien dat het kan lonen om bestaande normeringsinformatie te gebruiken in normering. We illustreren de methode met Duitse (Grob & Hagmann-von Arx, 2018) en Nederlandse (Grob et al., 2018) normeringsdata van de Intelligentie- en ontwikkelingsschalen voor kinderen en jongeren (IDS-2). Toekomstig onderzoek is nodig om deze methode te onderzoeken voor empirisch realistischere modellen dan Gaussische modellen.

Hoofdstuk 6

In dit hoofdstuk worden op basis van de bevindingen in dit proefschrift aanbevelingen gegeven voor toekomstig onderzoek en voor testuitgevers.

We benadrukken dat goede modelselectie cruciaal is. We illustreren dat het belangrijk is om niet zomaar een verdeling te kiezen, maar ook te kijken naar de aard van de

ruwe test scores (bijv. continu/discreet) en de testcontext (bijv. wel/geen variabele itemmoeilijkheid). Verder concluderen we dat zowel modellen met polynomen als modellen met *P-splines* (Eilers & Marx, 1996) een goede fit kunnen hebben, zolang de modelselectie goed is uitgevoerd. Toekomstig onderzoek is nodig om te onderzoeken wat de optimale selectie is voor modellen met meerdere predictoren en modellen met (monotone) *P-splines*. Ook is het interessant om in de toekomst verschillende continue normeringsbenaderingen (bijv. GAMLSS en niet-parametrische normering) met elkaar te vergelijken in verschillende normsituaties.

Een belangrijke praktische vraag is hoe de groot de normeringssteekproef minimaal moet zijn voor een bepaalde minimale normprecisie. Er zijn al richtlijnen voor *regression-based* normering met het standaard lineaire regressiemodel (Oosterhuis et al., 2016), waarbij homoscedasticiteit wordt aangenomen. Er zijn nog geen duidelijke richtlijnen voor de steekproefgrootte voor modellen met niet-lineariteit, heteroscedasticiteit en/of niet-normaliteit. In de studies in dit proefschrift vonden we nog amper toename in precisie als de normeringssteekproef groter werd dan ongeveer 1.000 observaties, maar het is lastig om dit resultaat te generaliseren naar andere normsituaties. Het is belangrijk dat in de toekomst meer onderzoek gedaan wordt naar de minimale vereiste steekproefgrootte voor een groot aantal normsituaties.

We bespreken ook een aantal aanbevelingen voor testuitgevers wat betreft het rapporteren van de genormeerde scores. In het algemeen bevelen we ze aan om in de testhandleiding meer informatie te geven over de gebruikte normeringsmethode, omdat op dit moment meestal erg weinig informatie hierover wordt gegeven. Daarnaast bevelen we aan om zowel de onzekerheid in genormeerde scores ten gevolge van testonbetrouwbaarheid als de onzekerheid ten gevolge van steekproeffluctuaties te rapporteren in de vorm van betrouwbaarheidsintervallen om de genormeerde scores.

Continue normering resulteert in accuratere en meer efficiënte normschattingen dan traditionele normering, maar een praktisch nadeel van continue normering is de complexiteit. Voor een uitgebreide beschrijving van hoe GAMLSS kan worden gebruikt om genormeerdere scores te bepalen – inclusief R code en voorbeelddata – verwijzen we naar Timmerman, Voncken, en Albers (2019). Verder bevelen we testuitgevers aan om visualisaties zoals centielcurves te presenteren, zodat zelfs complexe normmodellen gemakkelijk

te begrijpen zijn voor de testgebruiker. De normtabellen zijn ook uitgebreider dan bij traditionele normering, omdat de normen kunnen worden bepaald voor elke (combinatie van) exacte predictorwaarde(s). Daarom bevelen we ook aan om een digitaal scoringsprogramma te gebruiken, met de uitgebreide normtabellen daarin verwerkt.

Tot slot, continue normering met distributionele regressie is flexibel, waardoor normen nauwkeurig kunnen worden geschat. Deze flexibiliteit gaat gepaard met uitdagingen, zoals ingewikkelde modelselectie en mogelijk complexe modellen die moeilijk te begrijpen zijn en grote steekproeffluctuaties hebben. Zoals we hebben laten zien in dit proefschrift kunnen we deze uitdagingen het hoofd bieden door goede modelselectie, visualisaties, en efficiënte normering.