Toelichting Ankeronderzoek met Referentiesets

Saskia Wools & Anton Béguin, Cito 2014

Ankeronderzoek

Deze handleiding bevat een korte beschrijving van ankeronderzoeken. In het algemeen geldt dat meer informatie te vinden is in het boek *Test Equating* van Kolen en Brennan (2004), het hoofdstuk van Holland en Dorans uit het boek Educational Measurement (2006) en het hoofdstuk van Engelen en Glas uit het boek Psychometrie in de Praktijk (1993).

Zie voor de volledige referenties de referentielijst.

Beschrijving ankeronderzoek

Om toetsen te kunnen ankeren aan de referentiesets moet data verzameld worden waarbij leerlingen zowel (een deel van) de referentieset als (een deel van) de te ankeren toets maken. Het proces van ankering bestaat in grote lijn uit vier stappen:

1. keuze van de opgaven
2. ontwerpen afnamedesign
3. steekproeftrekking en dataverzameling
4. analyse

Keuze van de opgaven

Om een toets te ankeren aan de referentieset moet bepaald worden welke opgaven uit de referentieset worden afgenomen en hoe die worden gecombineerd met de opgaven uit de toets. In principe kunnen alle opgaven uit een referentieset worden gebruikt voor de ankering. Wel geldt dat sommige opgaven beter geschikt zijn doordat ze aansluiten bij de doelpopulatie van de toets en niet te gemakkelijk of te moeilijk zijn in die doelpopulatie. Verder moet het anker bestaan uit een voldoende groot aantal opgaven en moeten de opgaven inhoudelijk een representatieve dekking hebben. Om dit te realiseren kunnen de volgende richtlijnen worden gehanteerd:

- *Aantal opgaven in het anker*:
  Een anker bestaat uit minimaal 20 opgaven. En voor taal geldt aanvullend dat er minimaal 3 teksten voor ankering worden gebruikt.
- *Representativiteit.*
  Wat als representatief geldt hangt sterk af van de inhoud van de toets. In het algemeen geldt dat het goed is een zo divers mogelijk anker samen te stellen, waarin alle domeinen[1] en opgaventype en afnamesituaties[2] zijn vertegenwoordigd. In specifieke gevallen geldt echter dat de domeinen, opgaventypen en afnamesituaties moeten passen bij de te ankeren toets. Als het doel bijvoorbeeld is een toets te maken die zonder rekenmachine moet worden gemaakt, is het niet efficiënt te

[1] Bij rekenen: verbanden, getallen, verhoudingen en bewerkingen. Bij taal: begrijpen, interpreteren, evalueren, samenvatten en opzoeken.
[2] Bijvoorbeeld open of gesloten opgave en bij rekenen opgaven met en zonder rekenmachine en met of zonder context.

ankeren aan de opgaven met een rekenmachine. Inperking van het doel van de toets (en van het anker) leidt natuurlijk tot een inperking in de generaliseerbaarheid van de toets. Een toets met bijvoorbeeld alleen opgaven uit het domein *verbanden* kan niet zomaar worden gegeneraliseerd naar alle domeinen van rekenen.

- *Doelpopulatie.*
  De geschiktheid van opgaven voor ankering hangt samen met de geschiktheid van de opgave voor de groep leerlingen waar de toets voor bedoeld is (doelpopulatie). Als een opgave voor een bepaalde populatie relatief gemakkelijk of juist moeilijk is zal die minder informatie leveren over het vaardigheidsniveau van leerlingen in die populatie. Daarmee is zo'n opgaven minder geschikt voor ankering.

Ontwerp afnamedesign

- *Ankertoetsdesign met extern anker*
  Het geselecteerde anker kan worden afgenomen naast een afname van de toets. In dat geval moet er voor worden gezorgd dat de afnamecondities van het anker en de toets vergelijkbaar zijn. Dit type design wordt een *ankertoetsdesign met een extern anker* genoemd. Dit design is redelijk overzichtelijk en analyse hiervan kan zowel met regressiemethoden als met item response theorie (IRT).
- *Ankertoetsdesign met intern anker*
  Als alternatief voor het externe anker kan een toetsversie worden samengesteld waarbij de opgaven uit het anker worden gecombineerd met opgaven uit de toets. Dit type design heet een *ankertoetsdesign met een intern anker*. In dit design is gegarandeerd dat de afnamecondities van beide typen opgaven vergelijkbaar zijn. Wel moet er dan voor gezorgd worden dat de toetsversies serieus worden gemaakt door de leerlingen. Dit is vooral een aandachtspunt in VO en MBO, maar speelt ook in groep 7 en 8 van PO. Ook dit type design is via regressiemethoden en IRT analyseerbaar.
- *Gecombineerd ankertoetsdesign*
  In een laatste klasse designs worden meerdere toetsversies gemaakt waarbij de opgaven uit de referentiesets en de toets worden gecombineerd. Via deze designs is het mogelijk om langere toetsen te ankeren waarbij het niet mogelijk is om zowel het anker als de toets bij dezelfde leerling af te nemen. Van belang bij dit type design is dat er voldoende opgaven uit de toets en de referentieset in elke toetsversie zitten De data uit dit type design zijn minder goed via regressiemethoden te analyseren en er wordt daarom vrijwel uitsluitend gebruik gemaakt van IRT analysemethoden.

Aandachtspunten:

1) Als tijdnood optreedt zullen opgaven aan het einde van de toets slechter worden gemaakt. Probeer bij het design rekening te houden met mogelijke tijdnood en zorg eventueel voor toetsversies waarbij opgaven op verschillende plaatsen voorkomen.

Steekproeftrekking en dataverzameling

Data moeten worden verzameld in de doelgroep waarop de toets zich richt. Om opgaven te kunnen ankeren aan de centraal verzamelde data is het nodig dat er gegevens zijn van minimaal 200 leerlingen bij gebruik van een eenvoudig IRT model (Rasch model) en van 400 leerlingen voor een complexer model zoals het OPLM en 2 parameter model (Keuning,

2002). De data moeten worden verzameld in een steekproef die representatief is voor de doelpopulatie van de toets. Om aannemelijk te maken dat de verzamelde data tot een stabiele ankering leidt verdient het aanbeveling de meetfout van de ankering te bepalen. Na overbrenging van de cesuur voor het referentieniveau kan de onzekerheid van de ankering worden bepaald door het betrouwbaarheidsinterval van de equivalente cesuur te bepalen.

- Wanneer een adaptieve toets wordt geankerd aan de referentiesets zal dit gebeuren door een anker met items uit de referentieset samen af te nemen met de adaptieve toets. Vanuit de correlatie tussen de gemeten vaardigheidsscore (zie hieronder) en de score op het anker kan de cesuur op de vaardigheidsschaal van de adaptieve toets worden bepaald.
- De referentieset is gebaseerd op een afname van opgaven op papier. Als de te ankeren toets een digitale toets is zal een onderzoek moeten worden uitgevoerd binnen de doelpopulatie om te kijken of de items zich vergelijkbaar gedragen op papier en digitaal. Voor eventuele afwijkingen kan worden gecorrigeerd door aanpassing van het anker (zodat alleen opgaven zonder verschil worden meegenomen) of door een toepassing van een correctiefactor zodat het verschil tussen de beide afnamen wordt gecorrigeerd.

Analyse

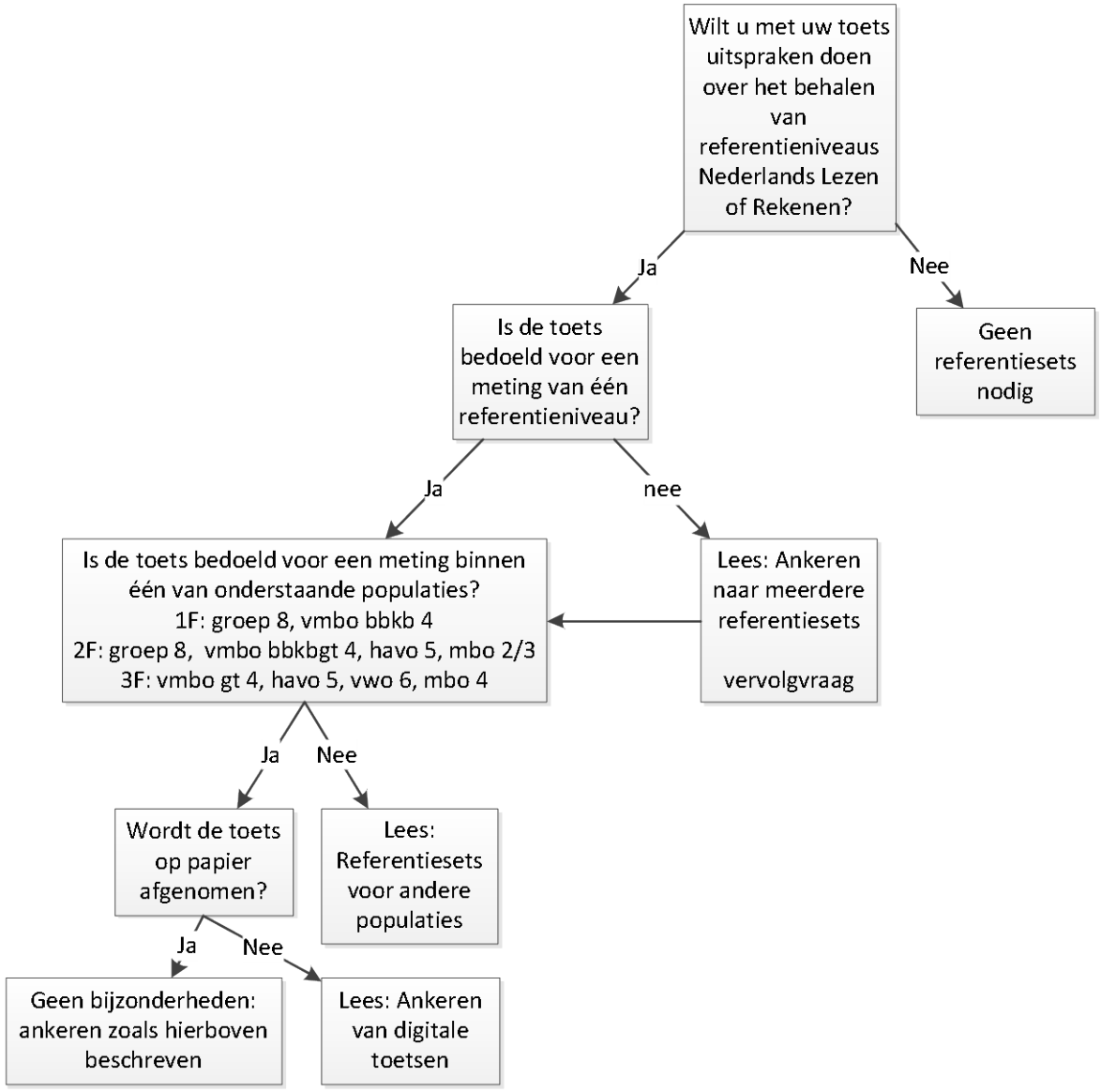
Er bestaan twee strategieën om de standaard van de referentieset over te brengen naar de te ankeren toets. De eerste strategie maakt gebruik van regressie om de scores op de toets en op het anker naar de referentieset met elkaar te vergelijken. Hierbij is de referentiecesuur gelijk aan een aantal goed op het anker uit de referentieset[3]. De andere procedure gaat uit van een vaardigheidsschaal of latente schaal (en gebruikt daarvoor IRT). De cesuur wordt omgezet in een cesuur op de latente schaal. De score op de te ankeren toets en de referentietoets worden beide afgebeeld op dezelfde latente schaal. De cesuur op de te ankeren toets kan worden bepaald als de score die een verwachte latente schaalscore heeft die het dichtst bij de latente cesuur ligt.

- Aandachtspunt bij de analyse is dat voor het bepalen van de latente schaal alleen de data op de referentieset gebruikt wordt die verzameld is bij de doelpopulatie. Voor elke doelpopulatie wordt dus een eigen latente schaal gemaakt.

[3] Bij oplevering van de data zal ook een tool worden geleverd waarmee voor selecties uit de referentieset de bijbehorende cesuur wordt geleverd. Voor bijvoorbeeld een subset van 20 opgaven wordt aangegeven bij hoeveel opgaven goed de cesuur ligt.

Specifieke situaties

In het vervolg van dit document wordt voor specifieke situaties aangegeven wat aandachtspunten bij een ankeronderzoek zijn. Deze situaties zijn via onderstaand stroomdiagram te vinden.



Ankeren naar meerdere toetsen

Situatie: u wilt met één toets uitspraken over meerdere referentieniveaus doen.

Voorbeeld Rekenen: u wilt met één rekentoets uitspraken doen over het wel of niet behalen van het 1F, 1S en 2F niveau.

Voorbeeld Taal: u wilt met één leestoets uitspraken doen over het wel of niet behalen van het 1F en 2F niveau.

Aandachtspunten:

1. *Keuze van de opgaven*

   Kies opgaven uit alle referentiesets waarop gerapporteerd moet worden voor de samenstelling van het anker

   Voor elk referentieniveau waarop gerapporteerd moet worden geldt:

   - Het anker bestaat uit minimaal 15 opgaven (bij 2 referentieniveaus dus 30 opgaven, bij 3 referentieniveaus 45 opgaven).
     - Voor taal geldt tevens: minimaal 15 opgaven verdeeld over minimaal 2 teksten per referentieniveau.
   - Het anker moet per referentieniveau representatief zijn over het domein. Het is dus niet de bedoeling om alle 1F opgaven uit één domein te kiezen en alle 2F opgaven uit de overige domeinen.
   - De gekozen opgaven moeten allen passen bij de doelpopulatie.

2. *Ontwerpen afnamedesign*

   Alle drie de beschreven designs (toetsdesign met extern anker, toetsdesign met intern anker, gecombineerd design) kunnen worden ingezet.

3. *Steekproeftrekking en dataverzameling*

   Er zijn geen specifieke aandachtspunten voor de dataverzameling.

4. *Analyse*

   Wanneer gebruik gemaakt wordt van IRT is het mogelijk om alle opgaven (ongeacht het referentieniveau) op één schaal weer te geven.

Referentiesets voor andere populaties

Situatie: u wilt met een toets uitspraken doen over een populatie die niet in het onderzoek Referentiesets is meegenomen.

Voorbeeld Rekenen: u wilt met een rekentoets uitspraken doen over het wel of niet behalen van het referentieniveau 1F door leerlingen in groep 6. Of uw wilt met een rekentoets een uitspraak doen over het wel of niet behalen van het referentieniveau 3F in vwo leerjaar 3.

Voorbeeld Taal: u wilt met een leestoets uitspraken doen over het wel of niet behalen van het referentieniveau 2F door leerlingen in havo 2. Of u wilt met een leestoets een uitspraak doen over het wel of niet behalen va het referentieniveau 2F in mbo 4.

Aandachtspunten:

1. *Keuze van de opgaven*

   U wilt graag rapporteren over referentieniveaus voor een andere populatie dan waarvoor centraal data voor de referentiesets zijn verzameld. Hierdoor is het niet mogelijk te ankeren naar de juiste populatieverdeling. Het gaat hier bijvoorbeeld om MBO niveau 1, vmbo leerjaar 1 en 2, havo leerjaar 1,2 en 3 en vwo leerjaar 1,2,3 en 4.

   Om een goede ankering te creëren naar een andere populatie is het noodzakelijk om de totale referentieset af te nemen bij die betreffende populatie.

2. *Ontwerpen afnamedesign*

   Gezien de omvang van de gehele referentieset ligt het voor de hand te kiezen voor een gecombineerd afnamedesign.

3. *Steekproeftrekking en dataverzameling*

   De data moeten verzameld worden bij minimaal 400 kandidaten per opgave waarbij de algemeen geldende regels voor een representatieve steekproef gehanteerd dienen te worden.

4. *Analyse*

   De cesuur kan worden overgebracht zoals eerder beschreven. In dit scenario is het niet nodig om de bijgeleverde tool voor het berekenen van de cesuur bij een selectie van opgaven te gebruiken. Hier geldt de referentiecesuur zoals gehanteerd op de volledige referentieset.

Ankeren van digitale toetsen

Situatie: u wilt rapporteren op referentieniveau bij een digitale toets.

Voorbeeld Rekenen: u wilt met een digitale rekentoets uitspraken doen over het wel of niet behalen van een referentieniveau, dit kan zowel een lineaire als een adaptieve toets zijn.

Voorbeeld Taal: u wilt met een digitale leestoets uitspraken doen over het wel of niet behalen van een referentieniveau, dit kan zowel een lineaire als een adaptieve toets zijn.

Aandachtspunten:

1. *Keuze van de opgaven*

   In principe is het noodzakelijk alleen opgaven te kiezen die ook in digitale vorm afgenomen kunnen worden. Voor het aantal opgaven in het anker geldt dat er rekening gehouden moet worden met uitval doordat opgaven zich te afwijkend gedragen in de verschillende afnamecondities. We raden daarom aan minimaal 25 opgaven per referentieniveau te selecteren als anker.

2. *Ontwerpen afnamedesign*

   Alle ankeropgaven dienen zowel op papier als digitaal afgenomen te worden. Dit kan alleen door middel van een gecombineerd ankertoetsdesign.

3. *Steekproeftrekking en dataverzameling*

   Zoals gezegd worden alle opgaven zowel op papier als digitaal afgenomen. Het is niet de bedoeling dat dezelfde leerling een opgave twee keer maakt. Dit betekent dat er meerdere toetsversies nodig zijn.

4. *Analyse*

   Tijdens de analyse fase kan onderzocht worden in hoeverre de items zich vergelijkbaar gedragen op papier en digitaal. Voor eventuele afwijkingen kan worden gecorrigeerd door aanpassing van het anker (zodat alleen opgaven zonder verschil worden meegenomen) of door een toepassing van een correctiefactor zodat het verschil tussen de beide afnamen wordt gecorrigeerd.

   Door de opgaven op een latente schaal te plaatsen kan vervolgens de cesuur overgebracht worden zoals eerder beschreven.

Referenties

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking. New York, NY: Springer.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 187{220). Westport, CT: Greenwood.

Engelen, R. J. H., & Eggen, T. H. J. M. (1993) Equivaleren. In T.H.J.M. Eggen & P.F. Sanders (Eds.) Psychometrie in de praktijk. Arnhem: Cito.
Ook te downloaden via:
http://www.cito.nl/nl/Onderzoek%20en%20wetenschap/achtergrondinformatie/publicaties/psychometrie_praktijk.aspx