Handreiking DATAHARMONISATIE VAN CROSS-NATIONALE DATAFILES

Harry Ganzeboom

Versie 1: 31 mei 2016

Deze versie: 27 september 2016

Inhoud

Ter inleiding

In toenemende mate maken onderzoekers gebruik van cross-nationale databestanden, waarin een grote hoeveelheid gegevens over een groot aantal landen periodiek wordrn verzameld. Belangrijke voorbeelden zijn: ESS, ISSP, SHARE, EVS, WVS e.a. De genoemde databestanden groeien periodiek aan doordat periodiek nieuwe gegevens worden verzameld en worden toegevoegd. Er ligt hier een enorm rijkdom aan gegevens, waarvan de relevantie in de volgende punten kan worden gekarakteriseerd.

- Cross-nationale gegevens maken *samenlevingen* de natuurlijke eenheid van onderzoek.
- Doordat de genoemde databestanden periodiek worden aangevuld, laten zij tegelijkertijd historische vergelijkingen toe op basis van het onderzoeksjaar of cohortvergelijking: herhaalde cross-secties brengen hetzelfde cohort binnen een land meermaals in beeld, en dat maakt het mogelijk om cohorteffecten van leeftijdseffecten te scheiden.

- De betreffende databestanden zijn van groot van omvang en beogen de hoogst mogelijke kwaliteit van dataverzameling te bieden.

- De betreffende databestanden zijn voor iedereen (gratis en zonder restricties!) toegankelijk en leiden op die manier automatisch tot een stijl van wetenschapsbeoefening die competitief, controleerbaar en cumulatief is: iedereen kan erbij en een andere proberen te verberen.

Hoewel in achterliggende organisaties enorme inspanningen leveren om de data tijdig en panklaar af te leveren, is het gebruik van cross-nationale databestanden toch niet zonder puzzles en moeilijkheden, die met name de beginneling tegenkomt. De oorzaken hiervan zijn de volgende:

- Er zijn sociale verschijnselen die zich naar hun aard niet laten harmoniseren of alleen met grote schade. Sprekend voorbeeld is partijvoorkeur bij verkiezingen. De vraag “wat hebt u bij de laatste nationale verkiezingen gestemd?” is in elk land te stellen, maar een ex-ante harmonisatie van de antwoorden is niet mogelijk. De mogelijke antwoordcategorieën zijn politieke partijen en elke vorm van ex-ante harmonisatie (bv. indeling naar links-rechts) doet afbreuk aan de kwaliteit van de meting.

- De bestanden worden in elk land verzameld door nationaal georienteerde onderzoekers, die het onderzoek niet alleen inrichten naar de hun aangereikte cross-nationale standaard, maar daarin ook locale omstandigden, voorkeuren en gangbare praktijken meewegen. Ze hebben daar vaak goede redenen voor.

- Het genoemde onderzoek groeit periodiek aan. De verzamelde informatie wordt daardoor meer divers. Dat komt door sociale verandering, maar ook door voortschrijdend inzicht. Het opleidingsstelsel is van beide een goed voorbeeld: opleidingsstelsels worden periodiek hervormd en daardoor moeten de antwoordcategorieën meeveranderen, maar er zijn ook nieuwe inzichten dat het de moeite waard is om naar opleiding liever zo gedetailleerd mogelijk te vragen. In ESS vind je bv een tendens om steeds meer detail te vragen.

Daar komen dan nog eens praktische problemen bij. Er zijn te noemen:

- De databestanden zijn bijzonder omvangrijk, met duizenden variabelen en tienduizenden of honderdduizenden cases. Computers worden ook steeds sneller, maar het loont toch nog steeds de moete om analysestrategieen uit te denken die de processing time van mens en machine reduceren.

- Wanneer men erop gericht is gegevens uit meerdere databestanden (bv ESS Round of ISSP Waves) samen te voegen, kan men op allerlei praktische problemen stuiten.

Deze handreiking is bedoeld om de beginneling in deze taken op weg te helpen en van wat handige tips en trucs te voorzien. Ze gaan ervan uit dat je SPSS op een PC werkt. Hetzelfde kun je ook voor elkaar krijgen in Stata, maar als je in Stata werkt, weet je vermoedelijk ook wel hoe je de SPSS suggesties kunt aanpassen. En als je op een Apple werkt, geldt vermoedelijk hetzelfde.

Bestanden, opslag en naamgeving

- Ga naar de website van het betreffende databestand en download de meest recente versie van de databestanden die je wil hebben.
- Databestanden worden periodiek herzien en aangevuld, vertrouw er niet op dat je de meeste recente versie al op je disk hebt. Ook is het safer om naar de bron terug te gaan, dan om onderhands een databestand te verkrijgen.
- Onthoud het versienummer van het databestand en zorg ervoor dat het voorkomt in de naamgeving van het bestand op je disk.
- Websites van de dataproducenten bevatten vaak handige overzichten van bekende mankementen aan de databestanden. De ESS website is op dit punt werkelijk voorbeeldig.
- Bij een bestand hoort een – vaak omvangrijke – datadocumentatie. Het meest cruciale onderdeel hiervan is in mijn ervaring de vragenlijst. Het loont vaak de moeite om die ook te downloaden.

Datacitatie

Bij het (her)gebruik van (andermans) data behoort een fatsoenlijk citatie. De regels van datacitatie zijn helaas niet universeel uitgekristalliseerd. Endnote heeft er een vast formaat voor, Mendeley moet op dit punt nog leren. Zorg ervoor dat je citatie de volgende gegevens bevat:

- Principal investigator: ingeval van cross-nationale databestand is het acceptabel niet de oorspronkelijke onderzoeker te noemen, maar een verzamelnaam, zoals ISSP Workgroup, of ESS-R1.
- Jaartal: dit is het jaartal waarin de betreffende versie van de data is gepubliceerd. Dit zal vaak afwijken van het jaartal van dataverzameling.
- Titel van het databestand met daarin het jaar van dataverzameling.
- Versie / editie.
- Data producer: dit is de instantie die de data in zijn uiteindelijke vorm beschikbaar stelt, meestal het data-archief waarvan je de data betrekt. Elementen kunnen zijn: plaats, nummer van het bestand in het archief.
- Ook wel toegevoegd worden: een explicatie [machine-readable data file] of [MRDF].
- Wees je ervan bewust dat verwijzen naar data iets anders is dan het verwijzen naar de documentatie over de data, of naar een oorspronkelijke publicatie over de data. Beiden kunnen hun nut hebben in je referentielijst, maar ze zijn niet een correctie datacitatie.

- Op de website van het databestand en in de datadocumentatie van het bestand is vaak een voorstel voor een citatie opgenomen. Je mag hierop variëren – met name de voorgestelde ESS citaties zijn onhandig lang.

Extract files

Als je gegevens uit meerdere databestanden wilt combineren, is het verstandig uit elk bronbestand alleen maar te werken met de variabelen die je nog hebt / denkt te hebben. In SPSS is computer processsing time evenredig met de omvang van een bestand en dit is de belangrijkste manier om die een beetje beperkt te houden (de andere manier is om het aantal cases te reduceren, via **sample**).

Je kiest de variabelen via een **/keep** statement, die zowel bij **get file** als bij **save outfile** gebruikt kan worden:

```
Get file = filename /keep = varlist.
Save outfile = filename /keep = varlist.
```

(je kunt ook een **/drop** statement gebruiken.)

Overlappende variabelennamen

Als je variabelen uit verschillende bronbestanden combineert, kun je tegen het probleem aanlopen dat verschillende variabelen in deze bronbestanden dezelfde naam hebben. Bij bv. ISSP wordt voor elke module de variabelennamen V1-V63 gebruikt. Als je hier niet op bedacht bent, zullen de grootst mogelijke ongelukken gebeuren: SPSS zal de betreffende gegevens opnemen volgens de labels van de betreffende variabelen in het eerst aangeroepen bestand.

Overlappende variabelennamen kun je repareren via rename:

```
Rename vars (V1 V2 V3 = aV1 aV2 aV3).
```
Deze statement is ook toegestaan als specificatie op **get file** en **save outfile**:

```
  /rename vars (V1 V2 V3 = aV1 aV2 aV3).
```

Boekhouding van de extract files

Heb je meerdere extract files die je later gaat samenvoegen, dan is het verstandig aan elke extract file een identifier toe te voegen, die je naderhand de weg wijst. Dit kunnen zowel de waarden van een variabelen zijn als verschillende variabelen. Mijn favoriete methode is via een string variabele Study

```
Format Study (a8).
Compute Study = "net1978e".
```

Maar als je niet van string variabelen houdt, is een alternatief:

```
Compute Studnr = 1.
Add value labels Studnr (1) "net1978e".
```

(Er is ook bij **Add files** nog een mogelijkheid om deze boekhouding op orde te krijgen.)

Gewichten

Bronbestanden zijn vaak voorzien van een of meer gewichten. Het meest gebruikelijk zijn dit post-stratificatiegewichten, die op basis van weging het databestand “representatiever” maken wat betreft een aantal achtergrondkenmerken waarvan populatieverdelingen goed bekend zijn, bv. geslacht, leeftijd, opleiding, burgerlijke staat en regio. De zin van post-stratificatie gewichten is omstreden. De ESS website geeft in tamelijk dreigende bewoordingen aan dat je echt gebruik moet maken van gewogen bestanden, mij lijkt dat klinkklare onzin. In het bijzonder is het onzin wanneer de variabelen die gebruikt zijn om het gewicht te maken als voorspeller voorkomen in je model (dat is al snel het geval voor geslacht, leeftijd en opleiding), dan wel geen enkele of een zwakke relatie met de afhankelijke variabele hebben (dat is vaak het geval voor burgerlijke staat en regio). In die beide gevallen brengt gebruik van een gewicht alleen maar de efficiëntie van je schatting in gevaar.

Tips:

- Als in een spss databestand het gewicht aan staat, kom in je extract file automatisch de variabele CASWGT terecht. Zet bij het maken van de extract het gewicht uit. Wil je er naderhand over beschikken, doe het dan in een expliciet benoemde variabele.
- Overweeg je analyse van gewogen databestanden, kijk dan eerst eens goed naar het gewicht:
    - Is het gemiddelde van het gewicht in het ongewogen databestand gelijk aan 1.0? Zo niet, wat is dan de rechtvaardiging van de ophoging / afschatting? Let erop dat wat gemiddeld 1.0 in een ongewogen compleet databestand is, dat niet meer hoeft te zijn in een geselecteerd databestand (bv. een leeftijdselectie 25-64 jaar).
    - Wat is de spreiding van het gewicht en wat zijn de uitbijters? Een gangbare regel is dat de verhouding tussen het laagste en het hoogste gewicht niet groter moet zijn dan 10/1 (bv 3.0/0.3). Zijn de extremen groter, dan kan met name tabelanalyse heel bedriegelijk zijn.
    - Bekijk eens wat zo mogelijk de ingrediënten van het gewicht zijn geweest. Als het goed is staat het in de data-documentatie, maar je kunt ook veel te weten komen als je het gemiddeld gewicht per geslacht, leeftijdscategorie etc. eens bekijkt.
    - Let erop dat je de analyse van een gewicht alleen maar in een ongewogen datafile kunt doen: **`Weight Off`**.

De enig zinvolle toepassing van poststratificatie-gewichten lijkt mij in de volgende situaties:

- De te bestuderen grootheid zijn niet modelparameters, maar een enkele karakteristiek van de afhankelijke variabele, bv. een gemiddelde of een percentage.
- Voor het construeren van het gewicht zijn variabelen gebruikt, die niet vergelijkbaar zijn tussen databestanden.

Maar mijn basisadvies is: ***gebruik post-stratificatiegewichten niet.***

Er zijn ook andere typen gewichten in databestanden, in het bijzonder design- of efficiëntiegewichten. Deze berusten op karakteristieken van het steekproefproces (in het bijzonder: clustering, (pre-)-stratificatie en systematische random trekking, die niet de 'representativiteit' van het bestand, maar de 'efficientie' van het bestand corrigeren. Efficientiegewichten zouden typisch niet gemiddeld 1.0 moeten zijn, maar kleiner (bij clustering) of groter (bij gestratificeerde of systematische steekproeftrekking). Helaas heeft ESS er juist op dit punt een potje van gemaakt: ze leveren een variabele DWEIGHT, die aangeeft design-effecten te corrigeren, maar dit niet doet.

Het belang van efficiëntiegewichten is in beginsel vele malen groter dan van post-stratificatiegewichten, maar ook hier hangt het af van samenhang tussen de ingredienten van het gewicht en het proces dat je bestudeert. Als een kenmerk random verspreid is over je steekproefclusters, maakt de methode van steekproeftrekking niets uit. De verstandiger strategie hier is om de steekproefclusters en –strata als kenmerken mee te nemen in je extractfile en je schattingen uit de voeren met **`Complex Samples`**. (een bekende vorm hiervan in Stata is clustercorrectie en svy-estimation.)

Add Files

Als je meerdere extractfiles hebt, zul je die uiteindelijk willen samenvoegen.

```
Add files
  /file = "filenaam"
  /file = "filenaam"
  Etc.
```

Het kan erg handig zijn om naderhand te kunnen beschikken over een identifier per samengevoegde file, die je weer kunt werken tot een study-identifier:

```
Add files
  /file = "filenaam"  in=In1
  /file = "filenaam"  in=In2
  Etc.

Recode in1 (1=1) into Studnr
Recode in2 (1=2) into Studnr.
```

Je kunt per **`/file`** regel een **`/rename`** component invoegen. Dat is handig om overlappende variabelennamen te repareren en ook moeilijkheden met string formats uit de weg te gaan.

Als je gebruik maakt van alfanumerieke [string] variabelen in de extractfiles, loopt SPSS vaak vast bij **`Add Files`**, omdat je alleen maar stringvariabelen met dezelfde naam kunt samenvoegen die hetzelfde format hebben. Lastig hierbij is:

- SPSS vaak spontaan de lengte van strings verandert.
- SPSS geen overzicht van de locatie van de problemen (welke variabelen?) geeft wanneer het meer dan drie samen te voegen bestanden gaat (bij twee of drie bestanden krijg je een keurig overzicht).

Ik weet geen andere oplossing dan bij de bereiding van je extract files het string format expliciet te definiëren:

```
Alter Type cntry (A2).
```

Landnamen

Het is handig om voor landnamen een string te gebruiken, en deze in te vullen het de tweeletterige of drie-letterige ISO-coderingen. Deze is bv voor Nederland: NL en NET. De tweeletterige zijn erg handig in figuren en als afkorting in interactietermen en landspecifieke variabelen. Lichtend voorbeeld is hier de ESS waarin Cntry deze benaming voor het land bevat.

Stringvariabelen kunnen value labels hebben (bv. **`add value labels cntry 'NL' 'Netherlands'.`**), maar nodig en handig is het niet.

Het gebruik van een numerieke codering voor landen en andere nominale variabelen met veel categorieën leidt veel gemakkelijker tot vergissingen en werkt bij sortering niet goed.

Strings

Voor landnamen en studie-identificatie zijn strings handige dingen: anders dan bij een nummering kun je hierin inhoudelijk informatie kwijt en ontstaat er bij alfanumerieke sortering altijd een logische volgorde. Voor het aanmaken en bewerken van strings in SPSS gelden echter wel een paar speciale regels die je goed moet begrijpen.

- Het format van een string is geregeld via het statement: **`Format <varname> (a2)`**. Hierbij geeft (a2) de lengte van de string aan. Je kun het format van een bestaande string veranderen via **`Alter type <varname> (a2)`**.
- Anders dan bij numerieke variabelen is het format van een string van groot belang, vooral bij het samenvoegen van bestanden: alleen als de strings een identiek format hebben, lukt de samenvoeging.
- Strings zijn case-sensitive. Je kunt de case veranderen met **`Compute <varname> = Upper(<varname>`**, en **`Compute <varname> = lower(<varname>.`**
- Ook getallen kunnen in een stringformat staan. Die kun je dan converteren met: recode <string> (convert) /into varname.
- Stringvariabelen kun je gebruiken in frequenties en kruistabellen en ook UNIANOVA accepteert ze ze als categorische variabelen. Je kun er niet mee rekenen, ook al zien ze eruit al getallen.

Harmonisatie van leeftijd en geboortejaar

Harmonisatie van leeftijd tussen verschillende bronbestand lijkt een eenvoudige zaak. Je neemt de bronvariabelen en copieert de geldige informatie in een nieuwe standaardnaam.

```
Recode V3 (1 thru 99=copy) into Age.
```

Maar wat doe je als in een van de bronbestanden leeftijd in categorieën is gemeten? Een goed idee is dan om de categoriemiddens te nemen:

```
Recode W3 (1=23)(2=28)(3=32) etc into Age.
```

Een simpel idee, maar het illustreert een belangrijk beginsel van harmonisatie: probeer zo weinig mogelijk informatie weg te gooien en kies in plaats van de grootste gemene deler voor een *schaling* van de informatie.

Wees je er wel van bewust dat de gegevens ook na schaling categorisch blijven: ga je bv geboortecohorten vormen, dan kan zo'n discrete leeftijdsvariabele tot onregelmatig gevulde groepen leiden.

Er zijn ook bronbestanden die ipv de leeftijd het geboortejaar als gegeven bij de respondenten hebben verzameld. Je dient dan de leeftijd te construeren als:

```
Compute Age = Surveyjaar – Geboortejaar.
```

In vergelijkende analyse heb je soms de leeftijd nodig en soms het geboortejaar. Deze kun je met grote precizie uit elkaar berekenen en het lijkt daarom niet nodig om beide in de harmonisatiefase al te construeren. Toch kan het waardevol zijn om geboortejaar zelf te harmoniseren, met name als dit gegeven bij dataverzameling onafhankelijk van de leeftijd is vastgesteld. Je leert dan iets over meetfouten.

Codeer geboortejaren (en alle andere jaren, zoals surveyjaar of entreejaar in de arbeidsmarkt) altijd met vier cijfers: 1980, 2009, etc. In oudere bestanden kun je aantreffen dat de eerste twee digits zijn weggelaten en dat leidt na 2000 tot allerlei verwarring.

Inkomen

Bij harmonisatie van inkomen kun je beste het model van leeftijd volgen. Codeer inkomenscategorieën aan de hand van de categoriemiddens tot reële getallen. Het doet er daarbij niet toe dat de categorieën en valuta tussen bronbestanden verschillen – zolang ze maar in dezelfde eenheid zijn gemeten (bv. euri per maand) kun je ze tot een schaal harmoniseren.

- Inkomens zijn veel ingewikkelder dan leeftijd. Belangrijke kwesties zijn:
    - Huishoudinkomen of persoonlijk inkomen?
    - Wat is de betaalperiode: jaar, maand, week? Sommige bronbestanden (EU-SILC, IALS) laten toe dat respondenten een betaalperiode kiezen en vervolgens daarop afgestemde hoeveelheden antwoorden. Dit is een recept voor veel ellende, doordat de informatie inconsistent kan zijn of missend.

- Valuta: een minder groot probleem dan je zou denken. Er wordt veel energie besteed aan het harmoniseren op een standaard valuta in PPP (purchasing power parities), maar deze energie is meest verspild: voor het afmeten van inkomensverschillen heb je deze stap niet nodig.

- Harmonisatie van gekwantificeerde inkomens gaat als volgt: **`Compute LNINK = ln(INK / xINK)`**, waarin xINK het (per bestand) gemiddelde inkomen is. Deze logaritmisering neutraliseert valutaverschillen, maakt van (lognormaal verdeelde) inkomens een symmetrische normale verdeling en levert bovendien een goede maat voor inkomensingelijkheid van een land: SD(LNINK).

- Met LNINK kiezen we in feite voor *within*-harmonisatie: we kunnen de gemiddelden van deze variabele niet zinvol tussen bronbestanden (landen) vergelijken. Dat is niet erg: als je geinteresseerd ben in effecten van het gemiddeld inkomen, kun je beter gebruik maken van macroeconomische grootheden als BNP of BNI.

- Een lastige bij inkomen wil ook nog wel eens zijn dat in het bronbestand het inkomen in verschillende porties is gevraagd, bv. arbeidsinkomen, overdrachtsinkomen, vermogensaanwas. Voor het samenvoegen van deze kwantiteiten weet ik geen andere richtlijn te geven dan uiterst omzichtig te werk te gaan.

Beroepen

In toenemende mate gebruiken cross-nationale databestand de International Standard Classification of Occupation [ISCO] als standaard categorisering. Dit is een viercijferige codering die er voor de begnner erg onhandelbaar uitziet en ook nog eens verschillende valkuilen bevat.

- Er bestaan verschillende versies van ISCO. Het meest gebruikt worden ISCO-88 en (steeds meer) ISCO-08, maar ook ISCO-68 en zelfs ISCO-58 kun je aantreffen. Er bestaat ook nog een Europese versie van ISCO-88: ISCO-COM, die lange tijd in de ESS is gebruikt. Er bestaan conversies tussen de verschillende versies: zie www.harryganzeboom.nl/ismf/index.htm.

- Er worden soms tweecijferige of driecijferige versies van ISCO gebruikt. Deze kun je omzetten in viercijferige door met 100 dan wel 10 te vermenigvuldigen. Ook eencijferige ISCO's komen voor (bv. gebruikt door Zuid-Afrika in de ISSP. Je kunt dan met 1000 vermenigvuldigen om een 'viercijferige' categorisering te krijgen, maar het aan te bevelen na te gaan hoeveel schade is ontstaan door de grove classificatie. Neem de proef op de som door ook in andere landen de laatste drie digits weg te vegen: **`compute isco000=1000*trunc(isco/1000).`**

De viercijferige codes kunnen in internationale geldige statusschalen worden omgezet. Er zijn hier drie mogelijke keuzen:

- ISEI – een continue schaal voor de sociaal-economische status van beroepen

- SIOPS – een continue schaal voor het prestige van beroepen

- EGP / ESEC – categorische indeling van beroepen tot sociaal-economsche klassen, mede op basis van beroepskenmerken als zelfstandigheid en leidinggevendheid.

- Daarnaast zien we ook wel voorbeelden van een categorische indeling op basis van het eerste cijfer van ISCO.

Naast internationale beroepenclassificaties zijn er ook gedetailleerde nationale beroepenclassifficaties. Deze zul je aantreffen wanneer je bv. vergelijkingen maak met Amerikaanse of Australische data. Meestal zijn hiervoor conversies in ISCO beschikbaar. Zo niet, dan zul je die moeten maken. Dat is minder moeilijk dan het lijkt – beroepsstructuren en beroepenclassificaties hebben een sterke verwantschap.

Interessant is dat men in bronbestanden ook categorische metingen van beroep kunt aantreffen met een beperkt aantal categorieen. In ESS is dat bv het geval voor de beroepen van ouders, in ISSP is in 2009 eenzelfde classificatie van 10 categorieen gemeten voor alle beroepen als tweede meting. Het gemakkelijkst ga je met deze informatie om door er ook ISCO categorieen van te maken.

Opleiding

Opleidingshoogte wordt in surveys op twee verschillende manieren gemeten:

- Via kwalificaties – meestal van het hoogst bijgewoonde / laatst afgemaakte programma.

- Via duur – in de vorm van een inschatting van de totaal ondergane cursusduur of via de leeftijd van het verlaten van het onderwijs.

Sommige databestanden (ESS, ISSP) vragen naar beide, althans waar het de respondent aangaat. Er ziin ook databestanden (WVS), waarin in sommige landen het een en in andere landen het ander wordt gevraagd.

Kwalificatie- en duurmeting is diepgaand geanalyseerd door Schröder (2014). Haar belangrijkste conclusies zijn:

- Kwalificatiemetingen zijn valider en betrouwbaarder dan duurmetingen.

- Niettemin hebben duurmetingen een stukje unieke informatie, waarmee je kunt laten zien dat kwalificatiemetingen niet volmaakt zijn in termen van betrouwbaarheid en validiteit.

- Als je over beide beschikt is het nuttig om opleiding te modelleren via een multiple indicator model.

Deze laatste tip gaat de krachten van de gemiddelde gebruiker te boven. Die zal moeten kiezen tussen kwalificaties en duur. Dan geldt:

- Kwalificaties zijn vaak zeer moeilijk te harmoniseren tussen landen (en in feite ook: tussen historische episoden). Het is veel werk en vergt een brede kennis en documentatie.

- Duurmaten zijn gemakkelijk te hanteren, hebben een theoretisch en praktisch gemakkelijk te interpreteren eenheid.

- In duurmaten treedt vaak een praktisch probleem op, namelijk dat sommige respondenten extreme duren rapporteren – dit kan voortkomen door een verkeerd begrip van de vraag. Je kunt deze waarden aftoppen of helemaal missend maken. Als je in staat bent effcient met missing values om te gaan, vind ik het laatste te prefereren.

Comparatieve databestanden lossen de problemen met het harmoniseren van landspecifieke opleidingsmetingen op door hiervan een *common denominator* harmonisatie aan te bieden. Hierbij zijn de opleidingscategorieën teruggebracht tot een beperkt aantal 'gemeenschappelijke' categorieen. In ESS is dit de variabele EDULVLa en in ISSP de variabelen DEGREE. Beide hebben 5-7 categorieen en berusten op de meer gedetailleerde International Standard Classification of Education [ISCED], die een driecijferig stelsel biedt om kwalificaties in te coderen.

Schröder (2014) laat zien dat er door het gebruik van deze harmonisaties vrij veel informatie verloren gaat. Dat komt doordat de eerste digit van ISCED niet erg gevoelig is voor belangwekkende onderscheidingen in het secundair onderwijs. Op dit manier komen grote hoeveelheden respondenten (soms meer dan 70%!!) in één categorie terecht, hoewel zij naar locale maatstaven nogal verschillende diploma's hebben. In Nederland wordt volgens ISCED-97 bv HAVO, VWO en MBO samengerekend (en ook HBO en WO). Een verbetering zou zijn wanneer internationale databestanden gebruik zouden maken van een meer gedetailleerde vorm van ISCED – ESS doet dit in de laatste twee rondes.

Als er locale metingen voorhanden zijn en de harmoniserende variabelen leiden aan het grofheidseuvel, kun je winst behalen door de locale metingen toch te verwerken. Ganzeboom & Schroder (2016) stellen voor dat te doen door de locale metingen alsnog in ISCED-2011 om te zetten – eigenlijk dus het klusje uit te voeren dat de dataverzamelaars of dataproducenten hadden moeten uitvoeren. Een shortcut is hier als volgt:

- Bekijk de locale categorieen en zet ze zo goed mogelijk op de goede volgorde. Kom je daar op basis van labels en inhoudelijk kennis niet uit, dan is handzaam om te kijken wat de gemiddelde beroepstatus op opleiding van de partner is per opleidingscategorie.

- Als je ervan overtuigd bent dat je de kwalificaties in de juiste volgorde hebt staan, dan kun je als volgt vergelijkbaar maken tussen bestanden:

    - Door within-standaardisatie (in percentielen of Z-scores) verkrijg je een vergelijkbare meeteenheden voor de relatieve positie van personen binnen de opleidingshierarchie.

    - Door schaling verkrijg je (ook) een vergelijkbare eenheid voor de positie van personen die zowel binnen als tussen landen vergelijkbaar is.

Maar hoe doe je schaling van opleidingscategorieen?

- De meest gehanteerde methode is om de categorieen te schalen naar hun veronderstelde of gemiddelde duur. Voor veronderstelde duur moet je over documentatie beschikkken, voor gemiddelde duur dien je over een afzonderlijke duurmeting te beschikken (niet noodzakelijk in het bestand dat je aan het analyseren bent). Deze oplossing heeft als voordeel dat je de kwalificatiemeting gemakkelijk kunt vergelijken / intergreren met een duurmeting.

- Schröder & Ganzeboom (2015) stellen de International Standard Level of Education [ISLED] voor als meeteenheid. Deze loopt tussen 0 en 100 en zijn afgestemd met de duurmaat in de ESS. Bij ESS bestanden is dit gemakkelijke toegankelijk, bij andere bestanden is de conversie vooralsnog moeizaam.

Wat kun je doen als je over zowel een kwalificatiemaat als een duurmaat beschikt? Zoals ook bij attitudeschaling kan het een goed idee zijn om de beide metingen te middelen.

- Voorwaarde is dat ze beide ongeveer even goed meten. Volgens Schröder (2014) is de duurmeting minder betrouwbaar dan de kwalificatiemeting, maar haar model voor het gemiddelde laat zien dat je met middeling toch nog wat winst boekt.

- Om de twee metingen te kunnen middelen dienen ze in dezelfde eenheid te staan. Dit kun je krijgen door ze beiden als in duur uit te drukken, en ook door within-standaardisatie.

Within- en between-standaardisatie

Standaardisatie betekent dat je variabelen tot een standaard meeteenheid terugbrengt. Het meest gebruikt is Z-standaardisatie, waarin je variabelen in eenheden standaarddeviatie uitdrukt. In SPSS is hiervoor een standaardfunctie: **`desc varname /save`**, hetgeen de nieuw varibele **`Zvarname`** oplevert.

Een andere manier van standaardiseren is naar proportie of percentielscores. Standaardfunctie in SPSS: **`Rank varname /proportion.`** Dit levert de nieuwe variabelen **`Pvarname`** op. Je kunt deze functie ook gebruiken om varianten als decielen of kwartielen te construeren.

Ook het dichotomiseren van variabelen is een manier om ze een gelijke 'meeteenheid' te geven. Lang niet altijd aan te raden (er gaat veel informatie verloren), vaak wel inzichtelijk.

Z-scores en P-scores kunnen we zowel over de gehele file als over de afzonderlijke deelbestanden berekenen. In beide gevallen ontstaat er een vaste meeteenheid, die het toelaat effecten tussen variabelen te vergelijken. Bij between-standaardisatie kunnen we gemiddelden en standaarddeviaties tussen deelbestanden vergelijken, bij within-standaardisatie is dat nu juist onmogelijk.

Within-standaardizatie is nuttig wanneer we het samengevoegde bestand als een geheel wilen beschouwen en hierop een gepoolde analyse willen loslaten. Dit is bij uitstek van toepassing als je bv de betrouwbaarheid van een attitudeschaal wil rapporteren, of een gepoolde regressie wilt tonen.

Between-standaardisatie is vereist als je variabelen wilt gebruiken als controle voor compositie-effecten. Wil je weten wat het effect van X is op Y rekening houdend met het verschillende opleidingsniveau van de bevolkingen van verschillende landen? Na within-standaardisatie kom je dit niet te weten.

En zo doe je het:

Between-standaardisatie

```
Desc varname /save.
Rank varname /proportion
```

Within standaardisatie:

```
Sort cases by Country.
Split file by country.
Desc varname /save.
Rank varname /proportion
Split file off.
```

Attitudeschalen

Vaak zal je doel zijn om multiple-indicator attitudeschalen te vergelijken en daarin trends en/of cross-nationale verschillen op te sporen. Je komt hierin dezelfde problemen als bij sociaal-structurele variabelen tegen:

- Variabelen kunnen in licht varierende formuleringen, of bv. met andere antwoordcategorieën voorkomen.
- Een schaal kan in het ene bestand met meer indicatoren aanwezig zijn dan in een andere.
- Ook indien indicatoren wel letterlijk hetzelfde zijn, zie je verschillen in gemiddelde en spreiding.

Ook hier geldt dat harmonisatie volgens de grootste gemene deler uiteindelijk het minste oplevert. Naarmate je meer bestanden samenvoegt is er minder gemeenschappelijk en je eindigt tenslotte met lege handen. In plaats daarvan moet je proberen de beschikbare informatie zo compleet mogelijk te houden en indien noodzakelijk een correctie maken voor het feit dat informatie incompleet of in een ander format aanwezig is.

Je werk wordt hier weer heel gemakkelijk als het voldoende is om te beschikken over within-gestandaardiseerde gegevens. In dat geval doet de oorspronkelijke eenheid er niet zoveel toe, en zou je ook kunnen beargumenteren dat met een (licht) verschillende indicatorenset nog steeds vergelijkbaar (‘functioneel equivalent’) meet. Het recept wordt dan:

- Verzamel de relevante indicatoren per databestand. Bekijk goed of de meeteenheid als om en nabij metrisch (interval) kan worden beschouwd.

- Standaardiseer de indicatoren per databestand.
- Voer dimensionaliteit- en betrouwbaarheidsanalyse uit om de optimale set indicatoren te vinden. Het is geen ramp als deze set niet helemaal hetzelfde is tussen deelbestanden, al heeft het de voorkeur als dat wel het geval is.
- Rapporteer dimensionaliteits- en betrouwbaarheidsanalyses in een within-gestandaardiseerde file.
- Voeg dummies in de gepoolde regressie toe als de procedure niet helemaal vertrouwt. Je zult zien dat dit bijzonder weinig uitmaakt voor het resultaat.