UNDERVISNINGS
MINISTERIET
STYRELSEN
FOR IT OG LÆRING

Beregning af sværhedsgrad for opgaver i opgavebanken i de nationale test

Introduktion

De Nationale Test tester elevernes dygtighed i udvalgte områder og fag. I hvert fag testes eleverne inden for tre hovedområder, der kaldes profilområder. Elevernes dygtighed beregnes ud fra de besvarelser, der er fremkommet ved, at eleven har besvaret en række opgaver (også kaldet items). Elevens dygtighedsparameter bestemmes i hvert profilområde for sig. Hvert profilområde består af en række opgaver, som til sammen danner opgavebanken.

I notatet beskrives, hvordan opgaverne afprøves og deres sværhedsgrad beregnes.

Dygtighedsmåling vha. Rasch-modellen

De Nationale Test er it-baserede, selvscorende og adaptive. Testene er baseret på Rasch-modellen (www.rasch.org). Rasch-modellen er en sandsynlighedsmodel, der i den simpleste udgave, kaldet det dikotome tilfælde, giver sandsynligheden for, at en elev nummer *n* med dygtighedsparameteren β_n svarer rigtigt (svarende til scoringen $X_{ni}=1$) på item[1] nummer *i* med item sværhedsparameteren δ_i:

$$P\{X_{ni} = x\} = \frac{e^{x(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$$

Sandsynligheden for, at en elev svarer rigtig på et item, afhænger således kun af elevens dygtighed og opgavens sværhed.

I Rasch-modellen optræder itemsværheder og elevdygtigheder på samme skala.

Rasch-modellen kan udvides til at inkludere items med flere subitems, således at scoringen x kan antage højere heltalsværdier end 1 svarende til, at flere subitems er besvaret korrekt, det såkaldte polytome tilfælde.

Det polytome tilfælde resulterer i en lidt mere kompliceret matematisk model med flere parametre, de såkaldte tærskelværdier τ, og en størrelse m_i, der angiver den maksimale scoring for det polytome item:

$$P\{X_{ni} = x\} = \frac{e^{-\tau_{1i} - \tau_{2i} \ldots - \tau_{xi} + x(\beta_n - \delta_i)}}{\sum_{z=0}^{m_i} e^{-\tau_{1i} - \tau_{2i} \ldots - \tau_{zi} + z(\beta_n - \delta_i)}}$$

[1] I Rasch-modellen anvendes begrebet item i stedet for opgave.

Det ses, at det dikotome tilfælde er indeholdt i modellen med m_i=1 og τ_{1i}=0.

I de nationale test anvendes en blanding af dikotome og polytome items. Rasch-modellen sikrer, at elevernes dygtigheder er sammenlignelige, selvom eleverne ikke har svaret på de samme items. Endvidere muliggør brugen af Rasch-godkendte items, at man i den adaptive testning af eleverne kan finde de items, der giver den optimale statistiske sikkerhed i bestemmelsen af elevens dygtighed.

Tærskelværdierne for de enkelte items i opgavebanken er beregnet på baggrund af opgaveafprøvningerne.

Opgaveafprøvning af nye opgaver

Opgaverne udarbejdes af faglige opgavekommissioner. Alle opgaver i opgavebanken er afprøvet af elever på det klassetrin testen er målrettet til. I dag bliver alle nye opgaver afprøvet på ca. 700 elever. Skoler, der skal deltage i opgaveafprøvningen, udvælges tilfældigt blandt landets folkeskoler. Udvælgelsen sker stratificeret efter skolens beliggenhed og størrelse.

Opgaveafprøvningen foregår som en lineær test, hvor eleverne får 2-3 sæt på ca. 30 opgaver i hvert sæt. Et sæt af opgaver kan besvares på 45 minutter. Opgaverne i hvert sæt randomiseres, så opgaverne kommer i forskellig rækkefølge til de enkelte elever.

Udvælgelse af items der er anvendelige til dygtighedsbestemmelse efter Rasch-modellen

På baggrund af elevernes besvarelser fra opgaveafprøvningen foretages en Rasch-analyse.

Det er Rasch-analysens formål:

1) at sikre at besvarelsesmønsteret for items passer med Rasch-modellen og
2) at beregne itemsværhederne

De nye items, der ikke passer til Rasch-modellen fjernes. De resterende items betegnes 'Rasch-godkendte'. Det bemærkes, at Rasch-godkendte items kun er godkendte i forhold til de andre items i det pågældende profilområde. Rasch-godkendte items fra et profilområde kan ikke nødvendigvis bruges sammen med items fra et andet Rasch-godkendt profilområde.

Items undersøges også for bias mht. 1) 'Class interval' (svarer til om modellen passer til elevscoringer), 2) Køn, 3) Skolestørrelse samt 4) Geografi i en Differential Item Functioning (DIF) analyse.

Items, der udviser statistisk signifikant DIF mht. class interval, køn, skolestørrelse eller geografi inklusiv interaktion mellem class interval og henholdsvis køn, skolestørrelse samt geografi, fjernes.

Alle analyser af besvarelser fra opgaveafprøvningerne foretages i software programmet RUMM (www.rummlab.com.au).

Til sidst undersøges, om der er statistisk signifikant negativ rangordenskorrelation mellem scoringerne for hvert item og personparameteren. Items, der udviser signifikant negativ korrelation, fjernes.

Efter Rasch-analysen af de nye items, sammenføjes disse med items i den eksisterende opgavebank. Efter sammenføjningen foretages en fastlåsning af tærskelværdierne for de eksisterende items. Denne fastlåsning kaldes 'anchoring' i RUMM. Dette sikrer, at tærskelværdierne på eksisterende items ikke ændres, således at elevdygtigheder over tid stadig umiddelbart kan sammenlignes.

For at itemsværhederne i de nytilkomne Rasch-godkendte items er relateret til de tidligere items, er det nødvendigt, at der i forbindelse med opgaveafprøvningen sker et overlap af items mellem blokkene af de tidligere og nye items. De items, der fungerer som overlapningsitems, kaldes 'link items'. Link items udvælges blandt eksisterende opgaver i opgavebanken og afprøves af elever sammen med de nye opgaver. Typisk udvælges ca. 10 link items i hvert profilområde i forbindelse med afprøvning af nye opgaver.

For at sikre, at link items ikke introducerer en forskydning i sværheden af items over tid, undersøges det, om disse link items udviser DIF mht. periode (tidspunkterne for opgaveafprøvningerne). Over tid kan der ske et skifte i elevernes interaktion med items: Som et tænkt eksempel kunne man forstille sig, at elever, der generelt var dygtige indenfor dansk sprog, også var dygtige til items relateret til salmevers i 1950'er, hvilket måske ikke var tilfældet i 1970'erne eller i dag. Altså den konsistens mellem forskelige items mht. elevdygtighed, der var gældende i 1950'erne, ikke nødvendigvis var gældende på et senere tidpunkt.

Link items, der udviser signifikant DIF mht. periode, gennemgår af ovennævnte årsag en procedure, der kaldes item-splitning. Item-splitningen kompenserer for den ændring i itemsværhed, der kan være opstået pga. et skifte i relationen mellem item'ets indhold og elevernes evne til at besvare item'et korrekt. For disse link items estimeres et nyt sæt af tærskelværdier.

Efter Rasch-analysen er de godkendte items parate til at indgå i itembanken med et sæt af tærskelværdier.