De kritiske spørgsmål til rapporten

Har vi begået nogle fejl?

Hvad kunne vi have gjort bedre?

Er de konsekvenser vi drager for vidtrækkende?

De nationale test har problemer

Prøverne måler som vinden blæser.

Der kan stilles spørgsmålstegn ved, om prøverne måler noget, er relevant.

Prøverne er ikke blevet ordentligt validerede.

Ideen om adaptive test er en dårlig ide.

Den adaptive algoritme fungerer dårligt.

Nogle af opgaverne er for dårlige.

Mange elever har det dårligt med at blive testet på den måde, som det foregår som det foregår i de nationale test.

DNT skal derfor evalueres

Ministeriets opgavebeskrivelsen understreger at DNT har følgende grundlæggende karakteristika:

- **Hver test består af tre faglige profilområder.**
- **De er it-baserede.**
- **De er adaptive. Vælger opgaver, der svarer til hvor dygtig eleven er.**
- **De er selvscorende.**
- **Der gives tilbagemelding pr. profilområde samt en samlet vurdering.**
- **En test kan gennemføres på 45 minutter.**

Men glemmer at understrege at

- **Validiteten af DNT er afprøvet ved item-analyse vha. Rasch modellen**
- **Opgavernes sværhedsgrader er givet ved modellens item parametre**
- **Elevernes dygtighed måles vha. estimater af personparametre.**

Evalueringen skal vurdere to forskellige forhold

Tekniske forhold (STIL):

- **Regner de nationale test rigtigt**
- **Måler de tre profilområder samme eller kvalitativt forskellige aspekter af det, der måles**

Betydning af og holdningerne til DNT (VIVE):

- **Femten forskellige temaer**

Vores rapport viser hvordan de tekniske forhold skal håndteres ved at besvare de to spørgsmål for prøven i læsning i 8. klasse.

Den siger intet i forhold til VIVEs femten temaer.

Vores analyser reproducerer de tidligere DNT-analyser

	DNT 2009 & 2014	JB & SK 2017
Model	Rasch	Rasch
Program	RUMM2030	TAM & DIGRAM
Sværhedsgrader	Pairwise	Marginal
Dygtighed	ML	WML
Usikkerhed	SEM	SEM
Standardisering	2010 population	-
Item fits	+	(+)
Person fits	-	(+)

Konklusioner

Der er højsignifikante forskelle på DNTs sværhedsgrader og estimater af sværhedsgraderne i 2017.

Der vil derfor være elever der får ”uhensigtsmæssige” opgaver.

Dygtigheden vil derfor bliver forkert beregnet.

Der er tilfælde hvor DNT undervurderer hvor dygtig eleven er, fordi testforløbet mislykkes.

I øvrigt er målingerne for usikre selvom ovennævnte fejl ikke eksisterede, men at fejlene bidrager til usikkerheden.

Betyder det noget? Det kan naturligvis diskuteres.

Rasch modellen

Sandsynligheden for et korrekt svar på en opgave afhænger af forskellen på hvor dygtig eleven er, θ, og hvor vanskelig opgaven er, β_i.

$$\frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}$$

Opgavernes sværhedsgrader blev estimeret i 2009 og i 2014 på meget store datasæt og er efterfølgende blevet brugt som kendte parametre.

Om dygtighed og sværhedsgrader i flg. Rasch modellen

Sværhedsgrader og dygtighed måles på såkaldte logit-skalaer.

Sandsynlighed = p. $\quad$ **Odds =** $\frac{p}{1-p}$**.** $\quad$ **Logit = ln(Odds)**

$$p = \frac{e^{logit}}{1+e^{logit}}$$

Målinger i Rasch modellen er relative

Forskelle på dygtighed for flere elever eller sværhedsgrader for flere opgaver måles ved logit-forskelle.

Hvornår er logit forskelle udtryk for store forskelle?

To elever med dygtighed lig med θ og $\theta + \delta$

β = sværhedsgraden for en opgave

Sandsynligheder i flg. Rasch

	Forkert	Rigtigt
Elev nr. 1	$\frac{1}{1+\exp(\theta-\beta)}$	$\frac{\exp(\theta-\beta)}{1+\exp(\theta-\beta)}$
Elev nr. 2	$\frac{1}{1+\exp(\theta+\delta-\beta)}$	$\frac{\exp(\theta+\delta-\beta)}{1+\exp(\theta+\delta-\beta)}$

Odds-ratio = $e^{\delta} \approx$ forholdet mellem sandsynlighederne for et korrekt svar på en meget vanskelig opgave.

Man stiller ikke meget vanskelige opgaver i pædagogiske test. For at vurdere betydningen af forskelle i dygtighed er det bedre at se på opgaver der ligger midt mellem de to personer.

Betydning af logit-forskelle på sandsynligheder for korrekte svar på en opgave mellem de to personer

δ	Elev 1: θ	Elev 2: $\theta+\delta$
,40	,45	,55
,80	,40	,60
1,00	,38	,62
2,00	,27	,73
3,00	,18	,82
4,00	,12	,88

Sandsynlighederne fortæller to forskellige historier om eleverne.

Hvor store skal δ være for at I synes, at historierne er *meget* forskellige?

Estimation af dygtigheden

Estimatet af dygtigheden, $\hat{\theta}$, er en funktion af antallet af korrekt besvarede opgaver, S, og sværhedsgraderne af de opgaver, som eleven har besvaret.

$$S = \sum_s s \frac{\exp\left(s\hat{\theta}\right)\gamma_s}{\sum_t \exp\left(t\hat{\theta}\right)\gamma_t}$$

γ_s er det elementære symmetriske polynomium af s'te grad af værdierne $e^{-\beta_i}$

Standard fejlen på $\hat{\theta}$ omtales som "standard error of measurement" (SEM).

Hvornår er SEM for stor?

Yderpunkterne i 95 % konfidensintervallet

$$\hat{\theta} - 1.96 \times SEM \leq \hat{\theta} \leq \hat{\theta} + 1.96 \times SEM$$

Angiver logit-værdier for sandsynligheden, p, for korrekte svar på opgaver, der svarer til eleven, $\beta = \hat{\theta}$

$$\frac{\exp(-1.96 \times SEM)}{1 + \exp(-1.96 \times SEM)} \leq p \leq \frac{\exp(+1.96 \times SEM)}{1 + \exp(+1.96 \times SEM)}$$

Hvis sandsynlighederne i yderpunkterne fortæller næsten samme historie om eleven er målingen sikker nok. Ellers er den for usikker.

Sandsynligheder for korrekte svar i yderpunkterne af konfidens-intervallet

SEM	Nedre	Øvre	Relative chancer	Odds-Ratio
,10	,45	,55	1,22	1,49
,15	,43	,57	1,35	1,82
,20	,40	,60	1,49	2,23
,25	,38	,62	1,65	2,72
,30	,35	,65	1,87	3,49
,40	,32	,68	2,12	4,48
,50	,27	,73	2,72	7,39
,75	,18	,82	4,48	20,09
1,00	,12	,88	7,39	54,60

I forbindelse med udviklingen af DNT var målet en SEM på 0,30.

I dag accepterer undervisnings ministeriet SEM værdier på 0,55

Hvor synes I der bliver fortalt den samme historie i yderpunkterne?

Adaptive test

SEM afhænger af antallet af opgaver og af sværhedsgraderne.

Jo tættere sværhedsgraderne er på θ, jo mindre vil SEM være.

De nationale test er adaptive, fordi adaptive test giver estimater af $\hat{\theta}$ med mindst mulig SEM.

De første tre opgaver vælges i midten af opgavebanken, hvorefter programmet løbende estimerer dygtigheden efterhånden som opgaverne besvares således at testsystemet kan vælge opgaver, der ligger så tæt på dygtigheden som muligt.

Et eksempel

Sværhedsgrader i 2017

Item	Score		Sværhedsgr.
0108020115072	1	1	1.38
01080200301234810-1	0	1	1.47
01080200301234966-1	1	2	1.37
01080200301234959-1	0	2	1.99
01080200301234953-1	1	3	1.42
01080200301234951-1	0	3	2.05
0108020111026	1	4	0.87
0108020110139-1	0	4	1.29
0108020111022	1	5	1.53
01080200301234848-1	0	5	2.09
0108020111019	1	6	1.37
01080200301238021-1	1	7	1.34
0108020111006	0	7	1.39
0108020110268	0	7	2.28
0108020110166-1	1	8	2.12
01080200301239338-1	1	9	2.98
01080200301239337-1	0	9	3.39

Estimation af dygtigheden baseret på sværhedsgraderne i 2017

Item		score	WML	SEM	Next
1	0108020115072	1			
2	01080200301234810-1	1			
3	01080200301234966-1	2	1.92	1.05	1.99
4	01080200301234959-1	2	1.55	0.99	1.42
5	01080200301234953-1	3	1.87	0.91	2.05
6	01080200301234951-1	3	1.61	0.83	0.87
7	0108020111026	4	1.77	0.78	1.29
8	0108020110139-1	4	1.48	0.73	1.53
9	0108020111022	5	1.69	0.69	2.09
10	01080200301234848-1	5	1.55	0.65	1.37
11	0108020111019	6	1.70	0.62	1.34
12	01080200301238021-1	7	1.83	0.60	1.39
13	0108020111006	7	1.65	0.57	2.28
14	0108020110268	7	1.56	0.55	2.12
15	0108020110166-1	8	1.73	0.53	2.98
16	01080200301239338-1	9	1.93	0.52	3.39
17	01080200301239337-1	9	1.88	0.51	

$$\hat{\theta} = 1.88 \quad \mathbf{SEM} = 0.51 \quad \mathbf{Corr(WML, Next)} = 0.59$$

Sværhedsgrader i flg. DNT

Item	Score		DNT	(2017)
0108020115072	1	1	1.95	(1.38)
01080200301234810-1	0	1	2.57	(1.47)
01080200301234966-1	1	2	1.97	(1.37)
01080200301234959-1	0	2	2.79	(1.99)
01080200301234953-1	1	3	2.23	(1.42)
01080200301234951-1	0	3	2.65	(2.05)
0108020111026	1	4	2.30	(0.87)
0108020110139-1	0	4	2.70	(1.29)
0108020111022	1	5	2.48	(1.53)
01080200301234848-1	0	5	2.65	(2.09)
0108020111019	1	6	2.62	(1.37)
01080200301238021-1	1	7	3.04	(1.34)
0108020111006	0	7	2.92	(1.39)
0108020110268	0	7	3.31	(2.28)
0108020110166-1	1	8	3.46	(2.12)
01080200301239338-1	1	9	3.72	(2.98)
01080200301239337-1	0	9	3.71	(3.39)

Estimation af dygtigheden baseret på sværhedsgraderne i DNT

Item		score	WML	SEM	Next
1	0108020115072	1			
2	01080200301234810-1	1			
3	01080200301234966-1	2	2.69	1.06	2.79
4	01080200301234959-1	2	2.32	1.00	2.23
5	01080200301234953-1	3	2.65	0.91	2.65
6	01080200301234951-1	3	2.36	0.84	2.30
7	0108020111026	4	2.61	0.78	2.70
8	0108020110139-1	4	2.40	0.72	2.48
9	0108020111022	5	2.61	0.68	2.65
10	01080200301234848-1	5	2.43	0.64	2.62
11	0108020111019	6	2.62	0.62	3.04
12	01080200301238021-1	7	2.82	0.60	2.92
13	0108020111006	7	2.68	0.57	3.31
14	0108020110268	7	2.59	0.55	3.46
15	0108020110166-1	8	2.77	0.53	3.72
16	01080200301239338-1	9	2.96	0.52	3.71
17	01080200301239337-1	9	2.88	0.50	

$$\hat{\theta} = 2.88 \quad \mathbf{SEM} = 0.50 \quad \mathbf{Corr(WML, Next)} = 0.82$$

DNT beregner forkerte tal for dygtigheden fordi DNT bruger forkerte sværhedsgrader.

2017: $\hat{\theta} = 1.88 \quad \mathbf{SEM} = 0.51 \quad \mathbf{Corr(WML, Next)} = 0.59$

DNT: $\hat{\theta} = 2.88 \quad \mathbf{SEM} = 0.50 \quad \mathbf{Corr(WML, Next)} = 0.82$

Sandsynlighed for korrekt svar på en opgave med sværhedsgrad = 2.38

2017	DNT
0,38	0,62

Er sikkerheden god nok, når SEM = 0,50?

Sandsynlighed for korrekt svar på hvis sværhedsgrad = dygtighed

Nedre grænse	Øvre grænse
0,27	0,73

Vores Konklusioner.

I dette eksempel er der for stor forskel på det DNT fortæller om eleven og det som analysen ved hjælp af de rigtige sværhedsgraderne fra 2017 ville have fortalt.

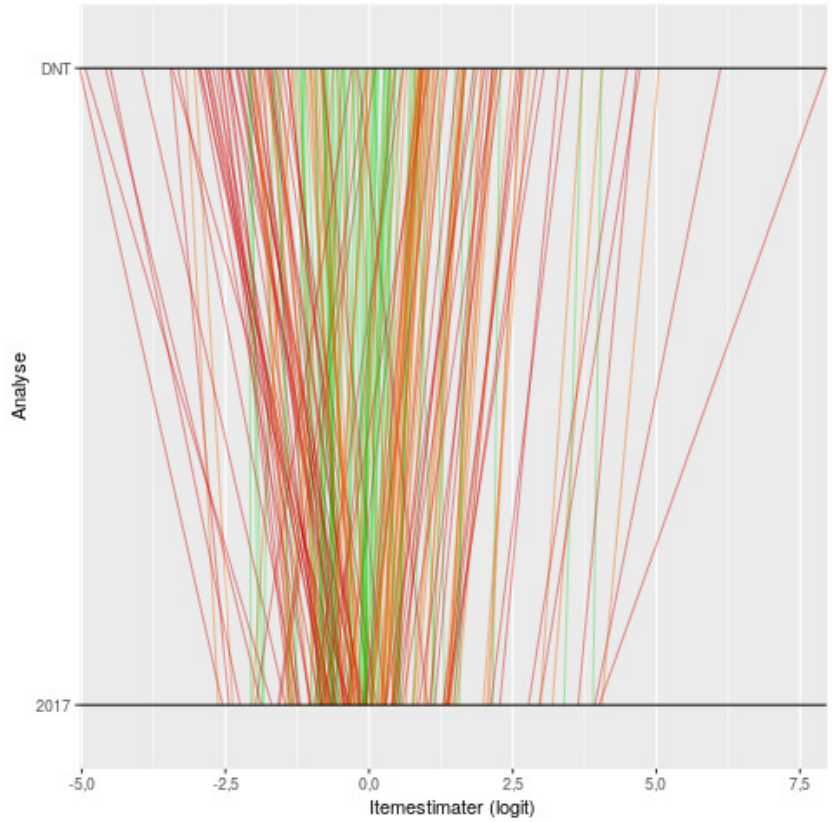
Usikkerheden med en SEM på 0,5 er alt for stor.

Er dette et usædvanligt tilfælde?

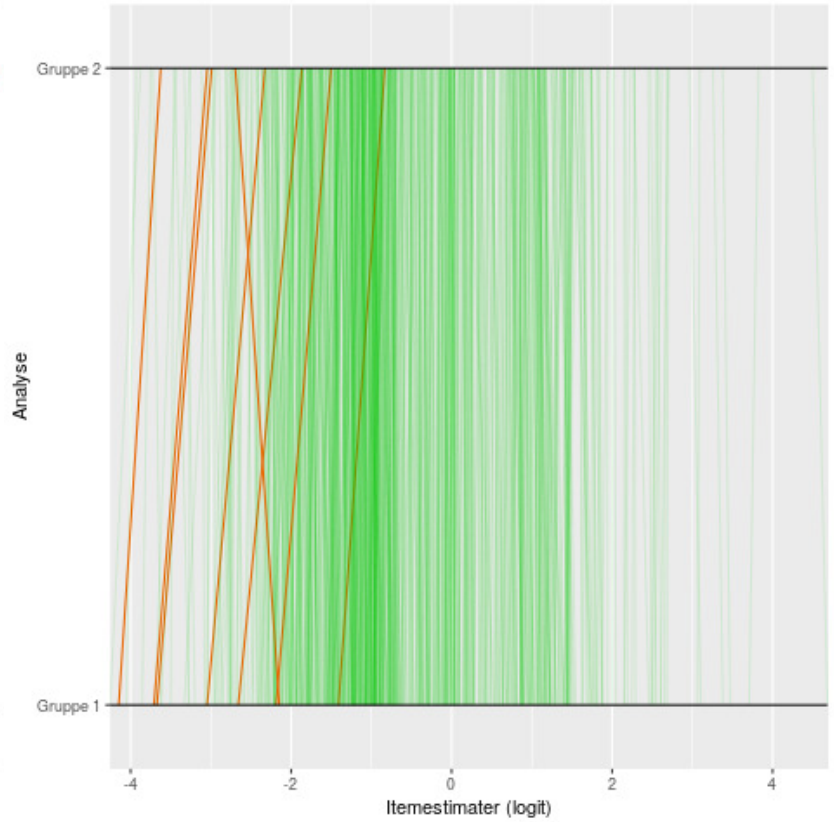
For at besvare dette spørgsmål, må vi se på resultaterne for alle 48.481 elever i 8. klasse

Højsignifikante forskelle på DNTs sværhedsgrader og sværhedsgraderne i 2017.

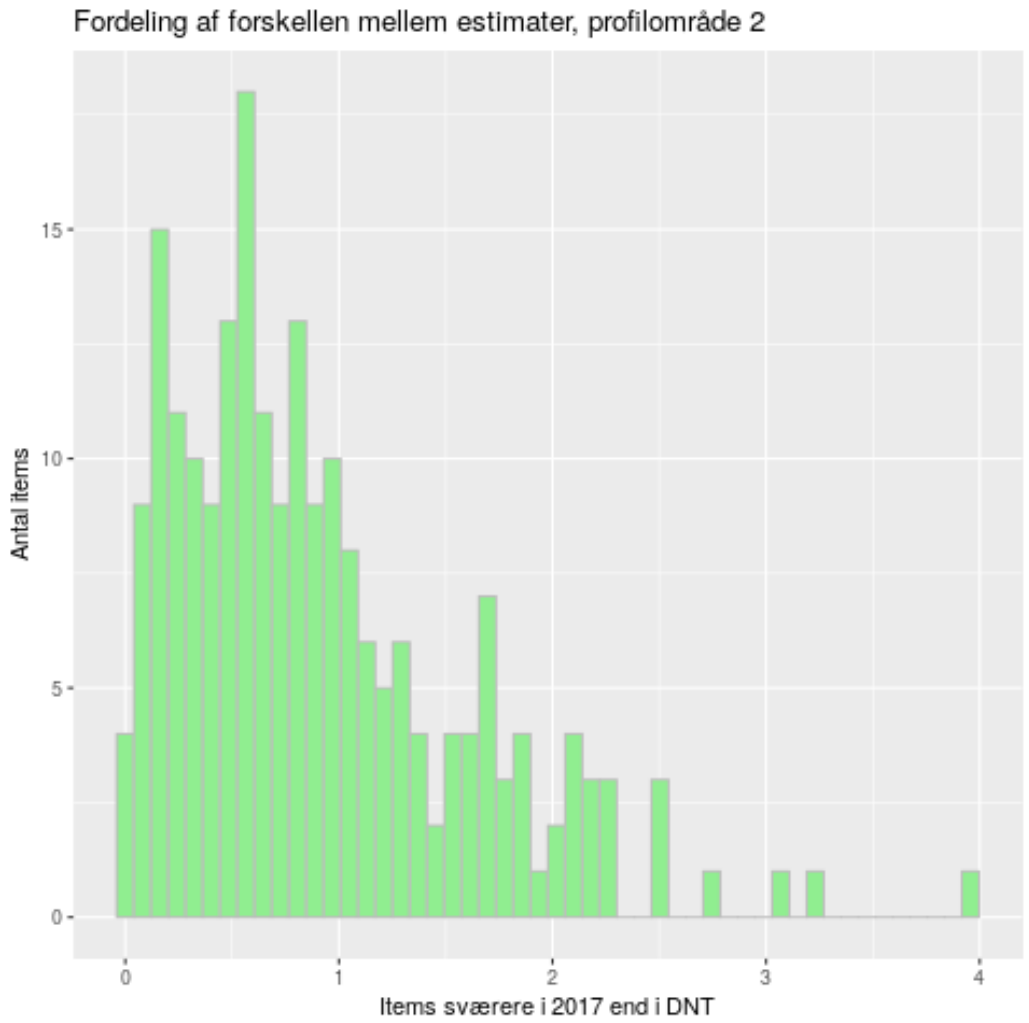
Forskelle på DNT og 2017



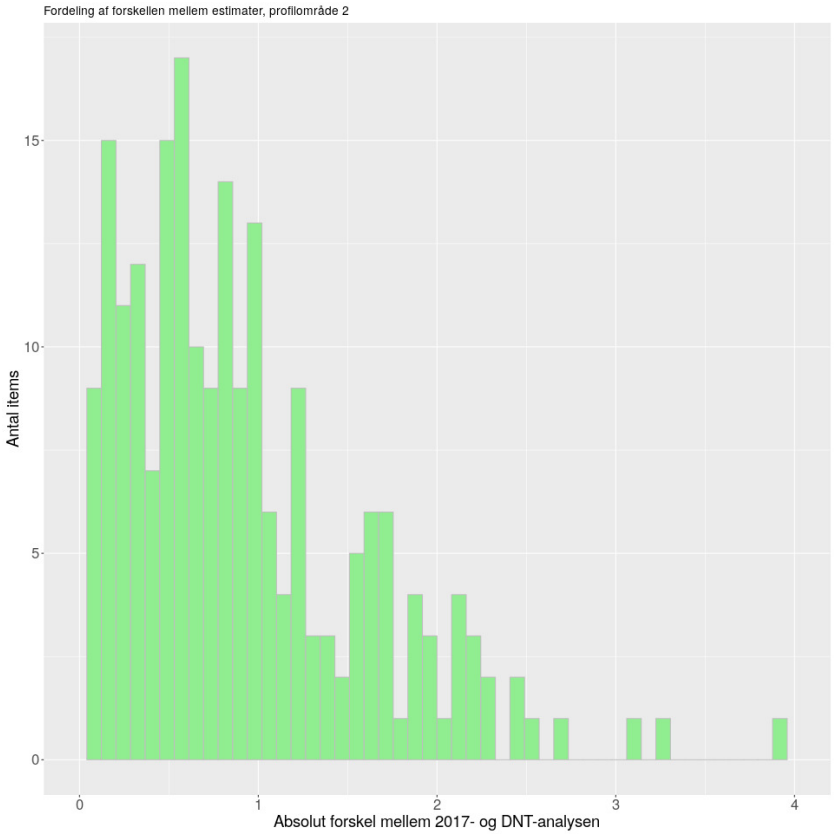
Opdeling af 2017 sample i to forskellige dele

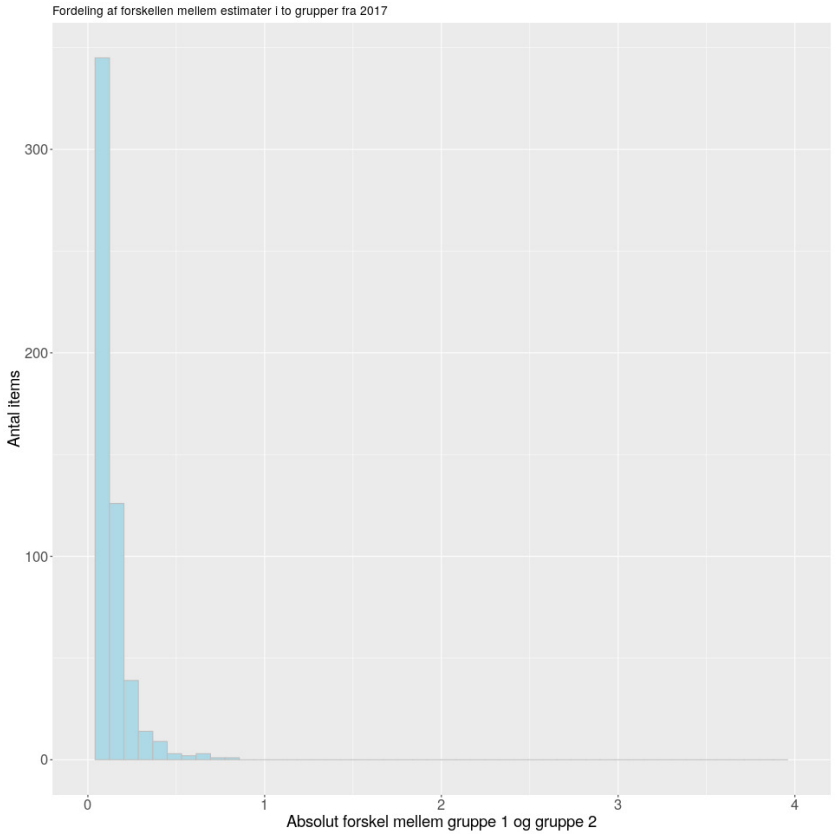


Forskelle på sværhedsgrader

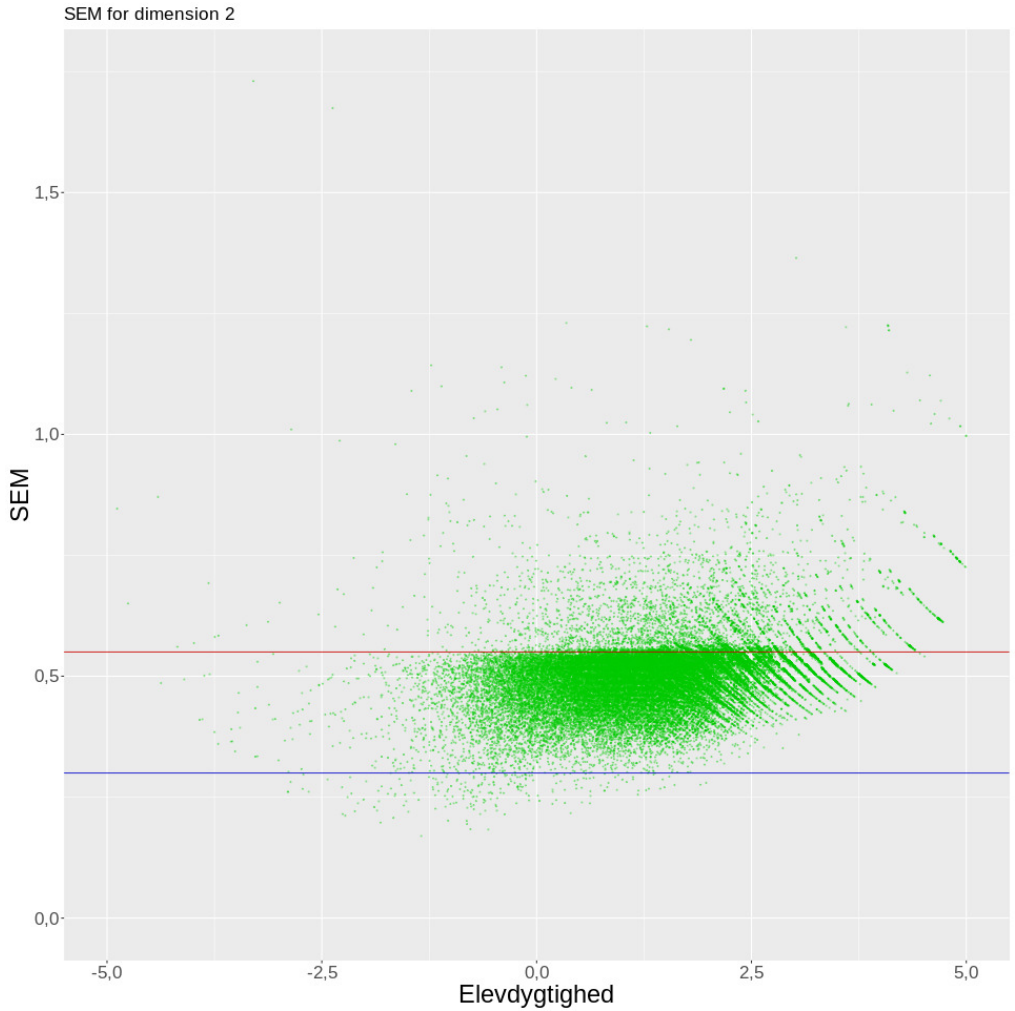


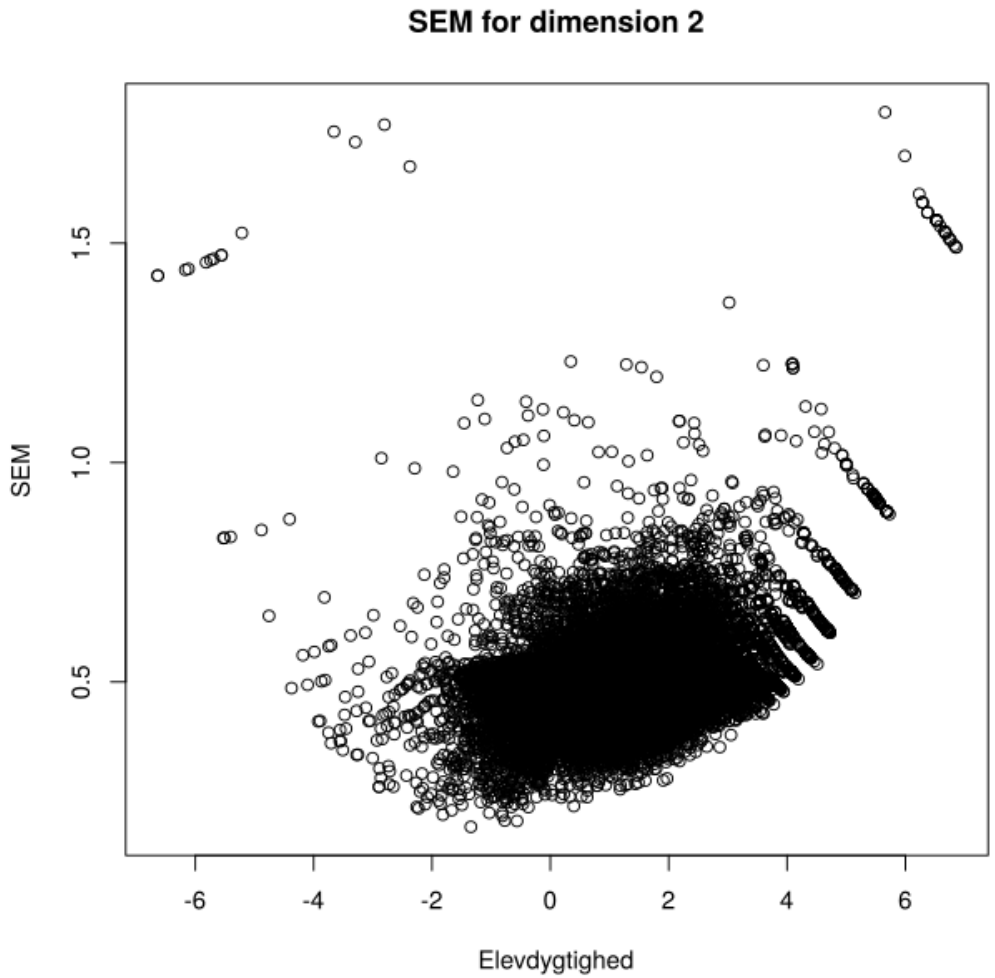
Forskelle på DNT estimater af dygtigheden og 2017-estimaterne





Hvor usikkert måler DNT?




SEM for dimension 2
SEM
0.5
1.0
1.5
-6
-4
-2
0
2
4
6
Elevdygtighed

Konklusioner

Mange eksempler på meget store forskelle på de to sæt sværhedsgrader.

Mange eksempler på meget store forskelle på dygtigheden beregnet vha. de rigtige sværhedsgrader og den dygtighed som DNT beregner.

Mange tilfælde med uacceptabelt stor usikkerhed.

Test af fit til Rasch modellen

Intet samlet test af fit til data

Intet test af differentiel item functioning (DIF)

Intet test af lokal afhængighed

Item fit statistics

Person fit statistics

Analyse af tilpasning mellem enkelte elevers svarmønstre og Rasch modellen

Item responses: $X = \{X_1,\ldots,X_k\}$ med samlet Score : $R = \sum_{i=1}^{k} X_i$

Delmængde af items $Z \subset \{X_1,\ldots,X_k\}$ med subscore : $S = \sum_{i:X_i \in Z} X_i$

Vi beregner sandsynligheden for S givet R

$$\text{Prob}(S=s|R=r) = \frac{\gamma_s(Z)\gamma_{r-s}(X\backslash Z)}{\gamma_r(X)}$$

og forkaster tilpasningen, hvis denne sandsynlighed er meget lille.

Da vi ikke har adgang til opgaverne kan vi ikke definere subscores ud fra indholdet af opgaverne. I stedet ser vi på subscores defineret ved de første opgave eleven har besvaret

Eksempel

Item	Score		PCM Thresholds			
0108030606134 0005-2	0	0	-0.47			
0108030311018	0	0	0.19			
0108030606133 0044-3	0	0	0.20			
0108030606132 52-4	0	0	-0.26			
0108030320029	0	0	-2.37			
0108030606136 01-3	0	0	-2.33			
0108030606133 0109	0	0	-1.59			
0108030320022-3	1	1	-1.44			
0108030320040	1	2	-0.91			
0108030691 2-1_2	1	3	-0.78			
0108030612006-1	1	4	-1.46			
0108030320033	1	5	-1.54			
0108030606133 0104	1	6	-1.22			
0108030612019-1	0	6	-0.35			
0108030606133 0034-2	1	7	-1.07			
0108030310373	4	11	-0.89	-1.08	-1.30	-0.93
0108030320069-2	1	12	-0.93			
0108030306094 0011-1	1	13	-1.03			

Item		Conditional probabilites cumulated score	
1	0108030606134 0005-2	0	0.54490
2	0108030311018	0	0.37428
3	0108030606133 0044-3	0	0.25294
4	0108030606132 52-4	0	0.13697
5	0108030320029	0	0.01359
6	0108030606136 01-3	0	0.00103
7	0108030606133 0109	0	0.00010
8	0108030320022-3	1	0.00079
9	0108030320040	2	0.00457
10	01080306912-1_2	3	0.01799
11	0108030612006-1	4	0.03655
12	0108030320033	5	0.06353
13	0108030606133 0104	6	0.11350
14	0108030612019-1	6	0.04892
15	0108030606133 0034-2	7	0.10353
16	0108030310373	11	0.33251
17	0108030320069-2	12	0.59974
18	0108030306094 0011-1	13	1.00000

Responses to items 1 to 7 are significant. p = 0.0001

Estimat af personparameteren

```
Estimates of person parameter

Optimal SEM with 21 dichotomous items ~  0.471

   WML = -0.631   SEM =  0.441   Bias =  0.004
    ML = -0.590   SEM =  0.478   Bias =  0.004


Subscore from 8 to 18 = 13  Max = 14

   WML =  0.932    SEM =  0.788
```

Konklusioner fra person fit analyser

Tilpasningen blev afvist i 11 ud af 17 tilfælde, hvor estimater baseret på Rasch modellen giver alt for pessimistiske bud på elevens læsefærdighed.

De 17 tilfælde var ikke udvalgt tilfældigt, så analysen viser kun, at sådanne tilfælde findes og at det er nemt at finde dem med det blotte øje.

Efterfølgende analyser må afsløre, om det er et stort elle lille problem.

Under alle omstændigheder bør DNT nægte at beregne dygtigheden, når analyser afslører at testforløbet er mislykkedes.

De samlede konklusioner

DNT bruger forkerte sværhedsgrader og kan derfor risikere at vælge opgaver, der er alt for lette eller alt for vanskelige for eleverne.

Af samme årsag beregnes forkerte estimater af dygtigheden.

Usikkerheden på målingerne er alt for stor. Da gentagne test vil vælge opgaver, med forskellige fejl vil det få målingerne til at se endnu mere usikre ud.

Der er eksempler på mislykkede testforløb, hvor dygtigheden undervurderes.

Konsekvenser

Brugen af testene bør sættes i bero indtil tilsvarende analyser er lavet på andre klassetrin og for andre fag, og indtil fejlene er rettet.

Derefter må man overveje følgende tre muligheder

1) **At fortsætte med DNT som adaptive test**
2) **At ændre DNT til ikke-adaptive test**
3) **At ophøre med at bruge DNT som obligatoriske test og i stedet at finde en enklere og billigere måde at følge den faglige udvikling på populationsniveau.**

I forhold til spørgsmålet om DNT skal overleve, kan vores rapport naturligvis ikke stå alene. Svaret på dette spørgsmål må bero på det samlede resultat af den evaluering, som forhåbentlig snart sættes i gang.

Vores rapport er et bidrag til denne evaluering, både som resultat i sig selv for læsning i 8. klasse og som forslag til, hvorledes de test-tekniske problemer for de øvrige dele af DNT skal håndteres.