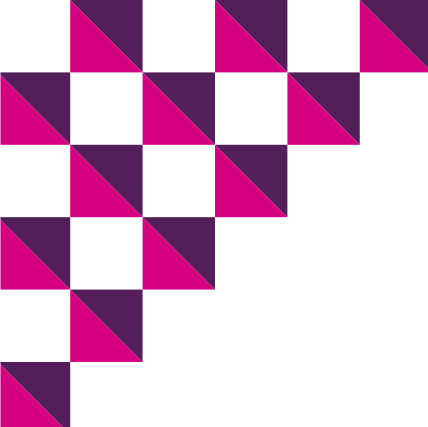


RAPPORT OM

KOMPE- TENCETEST I **HITTE PÅ-** PROJEKTET



INDHOLD

1	Baggrund
1	Relevante kompetencer til entreprenørskab
2	Udvikling af testredskab
3	Test med indsats og kontrol ved base- og endline
4	Udvikling af scoringskategorier
7	Scoring og dobbeltscoring
8	Rasch-analyse
13	Elevers kompetencer målt med instrumentet
14	Undersøgelse af udvikling
18	Konklusioner og perspektiver
19	Referencer

Rapport om kompetencetest i Hitte På-projektet
Jeppe Bundsgaard, DPU, Aarhus Universitet

Nationalt Videncenter for Læsning
December 2016
Alt indhold må printes og eftertrykkes.



Rapport om kompetencetest i Hitte På-projektet

Af professor Jeppe Bundsgaard

Baggrund

Hitte På-projektet er et interventionsprojekt oprindelig udtænkt af Janus Madsen og udført i regi af Nationalt Videncenter for Læsning med deltagelse af skoler, lærere og elever fra Kolding Kommune, WeLearn og en række konsulenter. Projektet er finansieret af Nordea Fonden og Fonden for Entreprenørskab – Young Enterprise Denmark (FFE-YE).

Målet med projektet er at udvikle nye rammer for innovativ undervisning i danskfaget, hvor eleverne gennem kreativitet og innovation udvikler literacy-kompetencer i første klasse i faget dansk. I den første pilotfase er der skabt tre læringsforløb til anvendelse i danskundervisning i 1. klasse. Læs mere om projektet på hittepå.dk og i (Lützen, 2016).

Denne rapport omhandler den kompetencetest som vi udviklede i tilknytning til projektet for at måle i hvilket omfang eleverne i første klasse udvikler nogle af de kompetencer som var intenderet i Hitte På-projektet.

Relevante kompetencer til entreprenørskab

Hitte På-projektet har som målsætning at ”forme foretagsomme, idérige og samarbejdsvillige elever” (hittepå.dk). Denne intention ligger i forlængelse af en generel bevægelse både i Danmark, i EU og internationalt. Der har således været en række projekter som har handlet om at formulere hvad det vil sige at være foretagsom og idérig – eller entreprenant og innovativ som det også kaldes.

FFE-YE har således udviklet en taksonomi til bestemmelse af udvikling i entreprenørielle kompetencer som består i fire hovedområder: handling, kreativitet, omverdensrelation og personlig indstilling (Rasmussen & Moberg, 2016).

De definerer de fire kompetencer således:

Handling: Eleven kan selvstændigt og i samarbejde med andre initiere, organisere, planlægge og udføre simple værdiskabende projekter i en enkel kontekst og forholde sig refleksivt hertil.

Kreativitet: Eleven kan arbejde vedholdende, eksperimenterende og undersøgende i enkle kreative processer på baggrund af grundlæggende faglig viden.

Omverdensrelation: Eleven kan ud fra sin forståelse af egen kulturel baggrund reflektere over muligheder i enkle sociale, kulturelle og økonomiske sammenhænge.

Personlig indstilling: Eleven har en grundlæggende tro på egne evner, kan selvstændigt og i samarbejde med andre deltage engageret i løsningen af usikre og åbne opgaver og håndtere egne emotionelle reaktioner i denne forbindelse (Rasmussen & Moberg, 2016, s. 15).

Som det fremgår, er der tale om meget omfattende kompetencemål som ikke umiddelbart kan omsættes til en konkret test – særligt ikke når eleverne går i første klasse og er i gang med at lære at læse og skrive. Vi besluttede derfor at udvikle en test som adresserede de to førstnævnte kompetenceområder, og som fokuserede på evnen til at planlægge, på kreativ idéskabelse og den omtalte grundlæggende faglige viden.

I første klasse handler den grundlæggende faglige viden om at lære at læse og skrive. Eleverne vil i første klasse for nogens vedkommende lære at læse og skrive deres første ord, mens andre allerede har gennemskuet skriftens principper og kan arbejde på at blive bedre læsere, stavere og skrivere. I løbet af første klasse vil de fleste have forstået skriftens principper og være i gang med at udvikle alle de mere avancerede aspekter af disse færdigheder. Det er naturligvis afgørende vigtigt at undervisningen støtter disse processer. Og derfor var det vores mål at udvikle et redskab der kunne give os indblik i denne udvikling, samtidig med at vi også kunne få indblik i elevernes kreative og handlingsrettede kompetencer.

At få ideer er en grundlæggende menneskelig egenskab – vi kan altid forestille os noget der var anderledes, eller måder at løse et problem på. Men kreativ idéskabelse handler om at få ideer som ikke er helt indlysende, eller som løser et mere kompliceret problem, og det handler i anden omgang om at kunne begrunde hvorfor der er tale om en god, realiserbar idé. Det fordrer at man kan forestille sig en verden der ikke er helt som den man kender. Inden for nyere didaktik kaldes en sådan kompetence for scenariekompetence (Bundsgaard & Foug, Forthcoming).

En central del af at kunne løse problemer er at kunne forestille sig hvad der skal til for at løsningen kommer i stand, altså hvilke handlinger der skal udføres, hvilke værktøjer der skal bruges, hvad der skal være på plads før man kan gå i gang, hvem der skal inddrages i løsningen osv. Det er dette aspekt der ligger i FFE-YE's handle-kompetenceområde. Elever i første klasse vil typisk kunne forestille sig nogle ret overordnede aspekter af en sådan proces, som fx hvilke faser en opgave skal løses i, eller hvilke delelementer den endelige løsning består i.

Udvikling af testredskab

Testredskabet blev udviklet af konsulent og læremiddeludvikler Janus Madsen, konsulent Peter Heller Lützen (Nationalt Videncenter for Læsning) og professor Jeppe Bundsgaard (DPU). På baggrund af ovenstående overvejelser besluttede vi at udvikle en simpel A3-papirbaseret test som bestod i fire simple spørgsmål til eleverne. I første version af testen var spørgsmålene: "1. Tegn det bedste sted på skolen, 2. Skriv om det bedste sted på skolen, 3. Tegn noget, som vil gøre skolen bedre, 4. Skriv om det, der vil gøre skolen bedre," Under spørgsmålene var der en ramme til at tegne i eller en række streger til at skrive på. Hvert ark havde desuden plads til at skrive skolens navn, klassenavn og uni-login på eleven der tog testen, således at før- og eftertest kunne knyttes sammen.

Vi udarbejdede en vejledning til læreren med manuskript til oplæsning ved starten af testen samt en række svar på spørgsmål der kunne opstå undervejs. Vi afprøvede disse spørgsmål i foråret 2015 på eleverne i en 1. klasse. Vores analyser af elevernes besvarelser viste at de ikke fik tilstrækkelig lejlighed til at vise procesbevidsthed og kreativitet i idéskabelsen. Der var mange ideer til svømmehaller og slikbutikker, og ikke mange anvisninger på hvordan disse ideer kunne realiseres.

Afprøvningen af testen gav anledning til at revidere spørgsmålene så de første to gik ud, og testen fokuserede mere på elevernes ideer og strategier for at realisere dem. Spørgsmålene lød nu: "1. Tegn en idé, der vil gøre din skoledag bedre, 2. Skriv om din idé, 3. Tegn, hvordan du vil lave din idé, 4. Skriv om, hvordan du vil lave din idé," Vi besluttede desuden at indføre fire felter i det ene tegnefelt (ved spørgsmål 3) for at opfordre eleverne til at tænke i processer.

Vi afprøvede denne udgave ved sommertid 2015 og foretog herefter nogle mindre ændringer. Det endelige testark så ud som i figur 1.

Navn: _____		Klasse: _____	Skole: _____	Unilogin: _____	
 1. Tegn en ide, der vil gøre din skoledag bedre!		 2. Skriv om din ide!			
 3. Tegn, hvordan du vil lave din ide!		 4. Skriv om, hvordan du vil lave din ide!			

Figur 1. Ark anvendt ved test.

Test med indsats og kontrol ved base- og endline

Testen af eleverne blev foretaget ved baseline i november 2015 og endline i marts 2016, 383 elever fra 13 indsats- og 9 kontrolklasser deltog ved både baseline- og endlinetesten. Det er disse elevers besvarelser der omtales i det følgende.

Der var afsat 20 minutter til testen, og læreren skulle læse følgende tekst op før eleverne gik i gang:

I skal på arket tegne og skrive om en god idé, som kan gøre skolen bedre. En idé kan fx være en ny legeplads, ting til klassen, bedre legeregler, at vi taler pænt til hinanden, et sted med ro, osv.

På den øverste del af arket skal I tegne og skrive om en god idé, som I synes vil kunne gøre jeres skoledag bedre.

På den venstre side uden linjer skal I tegne.

På den højre side med linjer skal I børneskrive, om det I har tegnet. Når I børneskriver, skal I skrive alle de bogstaver. I kan høre, og de ord I kender. Husk, at voksne ikke har lært at børneskrive og derfor ikke kan hjælpe jer.

På den nederste del af arket skal I tegne og skrive om, hvordan I vil lave jeres idé.

I skal også her tegne i feltet uden linjer og børneskrive i feltet med linjer.

Udvikling af scoringskategorier

Databehandlingen foregik i april-juni 2016. Den bestod i tre faser:

- 1) Udvikling af scoringskategorier
- 2) Scoring af elevbesvarelser
- 3) Statistisk behandling af scoringerne

Scoringskategorierne blev udviklet af Peter Heller Lützen og Jeppe Bundsgaard over en række møder. Elevernes besvarelser blev gennemgået og forskellige forslag til kategorier blev bragt op og diskuteret og dernæst afprøvet på et udvalg af elevbesvarelser.

Tre hovedområder blev identificeret: Skrivning, idé & proces samt tegning.

Eleverne befandt sig på ganske forskellige niveauer i deres skriveudvikling. Vi afsøgte en række teoretiske værker om begynderes skriveudvikling. Deri fandt vi blandt andet et bud på hvilke faser elever typisk går igennem når de tilegner sig skriftsproget. Her er disse gengivet efter (Hannibal, Korsgaard, & Vitger, 2010):

Skriftsprogsudviklingens fem faser

1. *Førkommunikativ skrivning*
 - *Skribling*
 - *bogstavskribling*
 - *logografisk skrivning*
2. *Semifonetisk skrivning*
3. *Fonetisk skrivning*
4. *Overgangsskrivning*
5. *Konventionel skrivning.*

Sådanne stadiebeskrivelser har en tendens til at være fokuseret på vejen mod korrekt stavning og i mindre grad lægge vægten på alle de andre træk ved skriveudviklingen, fx på kompleksiteten i sætninger (ordstillinger, brug af ledsætninger, variation, brug af mere sjældne ord, brug af tillægsord, biord osv.).

På baggrund af disse overvejelser og teoretiske input og på baggrund af analyser af elevteksterne udviklede vi kategorier der vedrørte længde, læsbarhed, variation, brug af verber, vokaler og ledsætninger samt variation i aktører. Vi valgte desuden et kriterium som vi kaldte staveudvikling som omhandlede hvor langt eleven var på vej mod konventionel stavning, og som fandt inspiration i de ovenfor citerede skriftsprogsudviklingsfaser.

Vi var interesserede i at identificere elevernes kompetencer til at få ideer og til at finde måder at realisere dem på. Vi udviklede derfor kriterier der vedrørte ideer og omfattede om eleven havde identificeret et problem, om ideen var innovativ, samt om der var begrundelse for ideen. I vores vurdering af elevernes kompetencer til at forestille sig en proces der kunne føre til løsningen, udviklede vi kriterier for om der var begrundelse for design eller løsningsvalg, om eleven beskrev et forløb eller en progression, om eleven inddrog en eller flere aktører i sin beskrivelse af løsningsprocessen, og endelig om eleven analyserede fænomenet hun foreslog at udvikle, i enkeltelementer.

Ud over de to hovedområder skrivning og idé og proces, der begge knytter sig tæt til entreprenørskabskompetencer, så valgte vi også at undersøge elevernes tegnekompetencer, idet vi vurderede at også tegning er en faglig forudsætning der kan kvalificere formidlingen af idé og procesforestillinger, ligesom den i mange tilfælde vil bidrage til udviklingen af ideer og processer. Vores fokus lå derfor ikke på æstetik eller naturtrohed, på repræsentation, men på variation og brug af et mere avanceret visuelt sprog. Et sådant sprog kunne man kalde en *diagrammatisk modalitet*, og det omhandler klarhed, præcision, overblik. Kan både eleven selv og andre overskue og forstå de delprocesser eller delelementer, som optræder?

Vi fokuserede kategorierne på antal elementer, læsbarhed, brug af ikoniske virkemidler (fx pile, bokse og andre tegn), brug af perspektiv, omfang af interaktion imellem aktører, samt brug af tekst og talebobler.

Da et første udkast til scoringskategorier var udviklet, blev det afprøvet af begge udviklere på elevbesvarelser fra to klasser. Scoringerne blev samlet i et regneark og sammenlignet. Der hvor der var uoverensstemmelser, blev kategorierne diskuteret igennem og tilrettet. De endelige kategorier blev dernæst afprøvet igen og scoringerne sammenlignet.

Efter afprøvning og yderligere tilretning ved et scorertræningsmøde blev de endelige scoringskategorier formuleret som det fremgår af tabel 1. Den yderste højre kolonne angiver hvilke omkodninger der blev foretaget i databehandlingen på baggrund af de Rasch-analyser som omtales nedenfor.

Tabel 1. Scoringskategorier. Højre kolonne angiver omkodning foretaget i databehandlingen.

Skrivning		
<i>Kategori</i>	<i>Beskrivelse og underkategorier</i>	<i>Omkodning</i>
<i>Tekst indskrevet</i>	<i>Teksten skrives med samme stavemåde, men der tilføjes mellemrum der hvor ord identificeres. Punktummer, streger og lignende som eleven har brugt til at markere mellemrum, erstattes af mellemrum. Slutpunktummer bevares.</i>	
<i>Mellemrum</i>	0: Ingen eller relativt meget få mellemrum. 1: Inkonsekvente mellemrum. 2: Tæt på konsekvent brug af mellemrum.	2 → 1
<i>Antal læsbare ord</i>	Tæl antal ord som kan forstås (også selv om der ikke er mellemrum). Tal tæller ikke som ord.	Udgår
<i>Antal (næsten) korrekte ord</i>	Tæl antallet af ord der er korrekt stavet (evt. med manglende endelse). Tal tæller ikke som ord.	Udgår
<i>Andel korrekte ord</i>	Beregnet som $\frac{\text{Antal (næsten) korrekte ord}}{\text{Antal læsbare ord}}$	1: Flere end 60 procent 2: Flere end 80 procent
<i>Antal unikke ord</i>	Tæl antallet af unikke ord. Forskellige stavemåder tælles som ét ord. Tal tæller ikke som ord.	1: 5 eller flere 2: 12 eller flere

<i>Læsbarhed</i>	0: Ulæsbar. 1: Delvis læsbar. 2: Fuldt læsbar.	
<i>Vokalbrug</i>	0: Ingen eller meget få vokaler. 1: Flere, op til næsten alle korrekte vokaler (forholdsvis få (konsekvente) fejl). 2: Næsten alle korrekte vokaler.	
<i>Staveudvikling</i>	0: Ingen fonetik. 1: Nogen sammenhæng mellem bogstaver og lyde (en del lyde glemmes eller fejlidentificeres). 2: Fonetisk. 3: Konventionel (inklusiv stumme bogstaver).	2 → 1 3 → 2
<i>Verbbrug</i>	0: Ingen eller kun infinitte verber. 1: Være, have, blive, måtte, ville, skulle, kunne, lave, sige og synes (herunder som hjælpeverber). 2: Et eller flere finit bøjede verber udover 1.	1 → 0 2 → 1
<i>Aktør</i>	0: Ingen aktør. 1: Jeg, mig, min, man, vi, os. 2: Andre end første person (herunder maskiner og institutioner der handler).	
<i>Brug af ledsætninger</i>	0: Ingen ledsætninger. 1: Substantiviske (som, at) og adjektiviske ledsætninger (indskudte). 2: Adverbielle ledsætninger (hvis, fordi, så).	

Tegning

<i>Kategori</i>	<i>Beskrivelse og underkategorier</i>	
Antal elementer	Tæl antallet af forskelligartede elementer. Tekst og ikoner (der tælles som sådanne) er ikke elementer. Elementer der gentages, tælles som én. Et element er en sammenhængende helhed. En bogreol med bøger er således én ting. En aktør som bruger en ting, tæller som aktør og ting (fx en person der sjipper er to ting). En ting som analyseres i delementer i 4-feltet, tælles som sine enkeltelementer og som tingen (et gyngestativ i 1, der vises i 3 som gyngestativ, stolper og snore er i alt fire ting).	1: 2 eller flere elementer 2: 6 eller flere elementer
Læsbarhed	0: Ingen elementer kan identificeres. 1: Nogle elementer kan identificeres. 2: Alle eller næsten alle elementer kan identificeres.	

Ikoniske virkemidler	0: Ingen ikoner vises. 1: Nogle eller flere ikoner vises (fx procespile, relationsstreger, krydser osv.).	
Perspektiv	0: Ingen perspektiv. 1: Ting/aktører placeret på hinanden (2 dimensioner). 2: Ting/aktører placeret rumligt i forhold til hinanden (3 dimensioner).	
Brug af tekst	0: Ingen elementer identificeres. 1: Nogle eller alle elementer identificeres.	
Talebobler	0: Ingen talebobler. 1: Én taleboble eller flere som ikke er i dialog. 2: Flere talebobler som bruges til dialog.	Udgår

Ide og proces

Identificeret problem	0: Intet problem. 1: Problem påpeges (der er meget larm, derfor stillerum).	
Innovativ idé	0: Ingen idé eller forslag til mere af noget der er i forvejen. 1: Forslag til noget der ikke er der i forvejen.	
Begrundelse for idé	0: Ingen begrundelse. 1: Begrundelse uden argument, begrundelse der udtrykker egen glæde/at ting ville være godt (jeg synes skolen ville blive bedre hvis der var en svævebane). 2: Argumenteret begrundelse (jeg vil have en skraldespand, for så bliver skolen renere).	2 → 1
Begrundelse for design/løsningsvalg	0: Ingen begrundelse. 1: Begrundelse for valg af design eller løsning (der skal være ryglæn, så man ikke falder ned).	Udgår
Forløbsbevidsthed – progression	0: Ingen forløb. 1: Beskrivelse af to faser/delopgaver. 2: Beskrivelse af tre eller flere faser/delopgaver.	2 → 1
Antal aktanter	0: Ingen aktakter (ingen der udfører arbejdet der skaber løsningen). 1: 1 aktant der udfører processen/skaber løsningen (fx mig, tømreren, læreren, butikken, maskinen). 2: 2 eller flere aktanter.	
Analyse i enkeltelementer	0: Ingen analyse. 1: Analyse i to dele. 2: Analyse i flere end to dele.	1 → 0 2 → 1

Scoring og dobbeltscore

Efter den indledende afprøvning af scoringskategorierne som blev foretaget af forskerne, blev scoringen foretaget af en studerende og en adjunkt tilknyttet Nationalt Videntcenter for Læsning. Disse blev uddannet ved at scorertræningsseminar med deltagelse af forskerne og scorerne.

Scorerne foretog en del af scoringen i fællesskab således at de havde mulighed for at diskutere tvivlstilfælde, og de mødtes undervejs med den ene forsker og tog spørgsmål op med ham.

For at sikre at der var konsistens i scoringerne, blev besvarelsene fra tre klasser i begge runder scoret af begge scorere (samlet 125 besvarelser). Scorerne kunne give samlet 68 point. Af tabel 2 fremgår fordelingen af scorerens uenighed om scoren. I gennemsnit var scorerne uenige om 8,87 % af scoren. Dette tal ligger fint under de grænser for scoreruoverensstemmelse der typisk sættes i internationale storskalatester. Ved den følgende Rasch-skalering blev en række af kategorierne slået sammen (se næste afsnit). Vi har ikke på nuværende tidspunkt testet scoreruoverensstemmelse på disse data, men vi vurderer at dette vil føre til lavere procentuel scoreruoverensstemmelse.

Tabel 2. Fordeling af scorerens uenighed målt i point.

<i>Antal uenige point</i>	<i>Antal tilfælde</i>	<i>Procent af samlet antal tilfælde</i>
0	15	10,9
1	3	2,5
2	11	9,2
3	12	10,0
4	14	11,7
5	17	14,2
6	14	11,7
7	12	10,0
8	8	6,7
9	10	8,3
10	3	2,5
Flere end 10	6	5,0

Rasch-analyse

I en klassisk test opgør man antallet af point og rangordner elevernes besvarelser på denne baggrund. Men en sådan rangordning tager ikke højde for at opgaver kan have forskellig sværhedsgrad, og for at opgaver kan være forkert konstruerede og derved slet ikke være et mål for det man ønsker at måle.

Den danske matematiker Georg Rasch udviklede derfor en metode til at afgøre om opgaver måler det samme, og til at fastsætte opgavers sværhedsgrad og på baggrund heraf elevers dygtighed. Dette kaldes Rasch-analyse eller Item Response Theory, og denne metode anvendes ved alle moderne internationale storskalatester.

Ved Rasch-analysen i denne undersøgelse anvendtes programmet ConQuest som også er blevet anvendt ved en række storskalatests, fx PIRLS, TIMSS, ICILS og PISA.

Rasch-analysen tager udgangspunkt i at opgaver har sværhedsgrader, og elever har dygtigheder, og at disse er på samme skala. Analysen bygger således på den antagelse at sandsynligheden for at en person med en given dygtighed besvarer spørgsmålet korrekt, hænger sammen med opgavens sværhedsgrad. Hvis personen har samme dygtighed som en opgaves sværhedsgrad, har hun 50 procents sandsynlighed for at svare korrekt. Er hun mindre dygtig end opgavens sværhedsgrad, falder sandsynligheden for at hun kan besvare den. Er hun dygtigere, stiger sandsynligheden.

Ved at regne frem og tilbage (estimere) på opgaver og elevbesvarelser, kan man afgøre hvor svære opgaver er, og derudfra hvor dygtige elever er til det såkaldt bagvedliggende træk (*latent trait*). Men det fordrer at opgaverne faktisk måler det samme bagvedliggende træk. Om det er tilfældet, kan afgøres af en række statistiske mål der kan beregnes på baggrund af estimeringen. Når man gennemfører Rasch-analyser må man således undersøge resultaterne af estimeringen og vurdere om der er opgaver der ikke måler det, man forventede. På denne baggrund kan man enten udelukke nogle af opgaverne fra estimeringen, eller man kan omkode dem således at opgaver med flere niveauer får færre niveauer. I Hitte På-testen er der ikke tale om opgaver i traditionel forstand, men om en performance-test med besvarelser der er scoret på en række kategorier. For nemhedens skyld anvendes alligevel ordet opgaver i det følgende.

Rasch-analyser skal foregå på variable der er på formen rigtig-forkert eller evt. på nogle få niveauer (fx 0, 1, 2 og 3 point). Vores mål for længde, antal ord, korrekt stavede ord, unikke ord og for antal elementer i tegningen var alle mål med mange værdier. Vi omkodede derfor disse værdier til værdier fra 0-3 (således at værdierne fordelte sig nogenlunde over de tre værdier). Vi gjorde flere forsøg med omkodning for at undersøge om disse kategorier kunne bringes til at være et mål for skrive- henholdsvis tegnekompetence, men det lykkedes, som det fremgår i det følgende, kun for tre af de fem kategorier.

Dimensioner

Vores umiddelbare forventning var at data ville fordele sig i tre dimensioner: Skrivning, tegning og idé- og proceskompetence. Det fungerede, men det krævede at en del af de scorede kategorier måtte udelades. Vi genovervejede derfor de scorede kategorier og besluttede at undersøge en alternativ opdeling i fire dimensioner, hvor den første dimension alene bestod af de korrekthedsorienterede dele af skrivekompetencen (brug af mellemrum, læsbarhed, vokalbrug og staveudviklingsstadiet), og hvor vi desuden omregnede antal korrekte ord til *andel* korrekte ord af de skrevne ord. Den anden dimension bestod af de kategorier fra skrivekompetencen som omfattede mere avancerede indholdsrelaterede kategorier (brug af verber, variation i aktører og brug af ledsætninger) kombineret med de dele af idé- og procesdimensionen der fremgik af den skriftlige del af elevernes besvarelse og som omhandlede udvikling af og begrundelse for en idé (identificeret problem, innovativ idé, begrundelse for idé (kategorien for begrundelse for design/løsningsvalg viste sig ikke at kunne indgå, hvilket skyldes at meget få elever fik point på denne kategori)). Vi kalder denne dimension for innovativ fremstillingskompetence. Vi bevarede dimensionen tegnekompetence, mens de procesorienterede kategorier fra idé- og procesbevidsthedsdimensionen (forløbsbevidsthed, antal aktanter og analyse i enkeltdele) udgjorde sin egen dimension, som vi kalder proceskompetence.

Rasch-analyser af de to forskellige opdelinger viste at den firedimensionelle havde bedre statistiske egenskaber end den tredimensionelle, og at den desuden målt med Akaike Information Criterion (AIC), som er et mål for hvor meget information der *ikke* anvendes i en analyse, anvendte mere af den indgående information (AIC var på 16922) end den tredimensionelle (hvor AIC var på 17342). Vi testede også en model med kun en dimension, og her var AIC på 17517. De fire dimensioner er beskrevet i tabel 3.

Tabel 3. Beskrivelse af de fire dimensioner

Dimension	Beskrivelse
<i>Skrivefærdighed</i>	Et mål for hvor korrekt eleven skriver. Målt gennem brug af vokaler, mellemrum, andel korrekte ord og staveudviklingsfase.
<i>Innovativ fremstillingskompetence</i>	Et mål for hvor langt, varieret, idérigt og begrundet eleven skriver. Målt gennem antal unikke ord, brug af verber, aktører, ledsætninger, og graden af innovative ideer, begrundelse for ideer og udfoldelse af forløbet i en løsningsbeskrivelse.
<i>Tegnekompetence</i>	Et mål for elevens brug af diagrammatisk modalitet til klart, præcist og kommunikativt at skitsere et problem og en løsning. Målt gennem antal elementer, læsbarhed, ikoniske virkemidler og brug af perspektiv og tekster.
<i>Proceskompetence</i>	Et mål for elevens kompetence til at identificere hvilket forløb der skal til for at nå til løsning på et skitseret problem. Målt gennem forløbsbevidsthed, antal aktører i sin beskrivelse af løsningsprocessen og af analysere i enkeltelementer.

Følgende ændringer viste sig nødvendige for at opnå et instrument som måler de fire underliggende dimensioner. For det første viste det sig at de kategorier som omhandlede længden af tekst, antal ord og antal korrekt stavede ord, ikke måler det samme underliggende fænomen som de øvrige kategorier. Men som omtalt omkodede vi antal korrekt stavede ord og antal ord til *andel* korrekt stavede ord ved at dividere den første med den anden, og dette mål korrelerede med de øvrige korrekthedsorienterede mål.

Det viste sig desuden nødvendigt at omkode alle kategorier hvor vi havde scoret i fire scoreniveauer (0-3), idet der ikke kunne opnås tilstrækkeligt lave t-værdier med fire scoreniveauer. For de fleste kategoriers vedkommende blev 3 omkodet til 2, og i nogle tilfælde blev 2 omkodet til 1. Se de konkrete omkodninger i tabel 1.

Resultaterne af de endelige analyser fremgår af tabel 4 og 5. I første række i tabel 4 ses opgave-id'et (item number), dernæst en forkortelse af kategoriens navn (som det fremgår er de første 11 opgaver knyttet til skrivning, fra 12-17 er tegneopgaverne og fra 18-24 findes idé- og procesopgaverne). Dernæst kan opgavens sværhedsgrad ses (på en logaritmisk skala fra omkring -7 til +7). I tabel 5 vises items med flere scoringsniveauer. De to kolonner med *fit*-data angiver hvor godt opgaven måler på det bagvedliggende træk. MNSQ er et mål for hvor præcist den teoretiske Rasch-score beskriver de empiriske data, eller sagt på en anden måde: MNSQ er et mål for tilfældighed i data. Hvis MNSQ-værdien for et item er 1, følger de empiriske data den teoretisk forventede fordeling. Jo højere værdien er, des mere uforudsigelig er elevernes svar, og det betyder at itemmet diskriminerer mindre end den teoretiske kurve (den empiriske kurve er fladere end den skulle være

ifølge teorien). Hvis MNSQ ligger under 1, er den for forudsigelig (itemmet bidrager ikke med viden om eleverne).

Unweighted MNSQ (også kaldet outfit MNSQ) tager højde for såkaldte outliers i data, fx elever som har svaret rigtigt på svære spørgsmål og forkert på lette. Weighted MNSQ (også kaldet infit MNSQ) tager højde for inliers, dvs. for informationsværdien i data.

Empiriske data følger aldrig teorien helt, og derfor må man opstille nogle retningslinjer for hvornår en værdi er acceptabel, og hvornår den ikke er (Boone, Staver, & Yale, 2014, s. 167; Osteen, 2010, s. 68). Wright og Linacre foreslår at MNSQ for surveys som skal ligge mellem 0,6 og 1,4, mens multiple choice-opgavesæt skal ligge mellem 0,7 og 1,3 og for high stakes test mellem 0,8 og 1,2 (Wright & Linacre, 1994, s. 370). For den vægtede MNSQ (infit) falder værdierne for alle på nær éns vedkommende inden for Wright og Linacres forslag til grænseværdier for high stakes test. Dette betragter vi som meget tilfredsstillende. For den uvægtede MNSQ-værdi (outfit) er der lidt flere (8 af 53) værdier som ikke falder inden for grænserne. Det vil sige at en del elever har fået flere eller færre point på de pågældende items, end man ville forvente ud fra deres øvrige svar. Dette vil skulle undersøges nærmere i en senere udgave af instrumentet. Ifølge John M. Linacre (Linacre, u.å.) er dårligere outfit, altså at der er mange outliers, en mindre trussel mod et instruments målekvalitet end infit, så vi vurderer at de pågældende værdier ikke i for høj grad påvirker instrumentets kvalitet.

Værdien i T-søjlen er resultatet af en t-test af nul-hypotesen: Stemmer data fra dette item perfekt med modellen? Hvis værdien er større end 2, afvises nul-hypotesen (og itemmet måler ikke sådan som man ville forvente på baggrund af modellen). Også i dette tilfælde er der tale om tommelfingerregler i forhold til hvordan værdier over 2 behandles. Boone med flere anbefaler at man ser bort fra høje t-værdier så længe MNSQ ligger inden for det acceptable interval (Boone m.fl., 2014, s. 166). For vægtede MNSQ-værdier er der alene tale om høje t-værdier ved items som ligger inden for de ønskede intervaller. For de uvægtede er det også her lidt mere problematisk. Vi vurderer dog af samme grunde som ovenfor at instrumentet har god kvalitet.

Tabel 4. Resultaterne af Rasch-analyse. Item-parametre estimeret med ConQuest.

TERM 1: item				UNWEIGHTED FIT			WEIGHTED FIT		
VARIABLES									
item	ESTIMATE	ERROR [^]		MNSQ	CI	T	MNSQ	CI	T
1 Andel korre,,	2,260	0,104		1,51 (0,90, 1,10)	8,6		1,12 (0,90, 1,10)	2,5	
4 Antal unikk,,	-3,219	0,112		0,95 (0,90, 1,10)	-1,0		0,96 (0,90, 1,10)	-0,8	
5 Mellemrum	-2,701	0,166		1,25 (0,90, 1,10)	4,4		1,06 (0,73, 1,27)	0,5	
6 Læsbarhed	-1,066	0,106		1,04 (0,90, 1,10)	0,8		0,97 (0,87, 1,13)	-0,4	
7 Vokalbrug	0,022	0,095		0,94 (0,90, 1,10)	-1,2		0,94 (0,90, 1,10)	-1,2	
8 Staveudvikl,,	1,485*	0,139		0,55 (0,90, 1,10)	-10,4		0,86 (0,77, 1,23)	-1,2	
9 Verbumbrug	-0,162	0,107		0,92 (0,90, 1,10)	-1,6		0,99 (0,89, 1,11)	-0,1	
10 Aktører	-2,326	0,094		0,95 (0,90, 1,10)	-1,0		0,94 (0,91, 1,09)	-1,3	
11 Brug af led,,	-0,523	0,085		1,05 (0,90, 1,10)	1,0		1,13 (0,88, 1,12)	2,1	
12 Antal eleme,,	-1,139	0,078		1,04 (0,90, 1,10)	0,7		1,03 (0,91, 1,09)	0,8	
13 Læsbarhed	-2,578	0,120		0,97 (0,90, 1,10)	-0,6		1,00 (0,89, 1,11)	0,0	
14 Ikoniske vi,,	1,010	0,096		0,99 (0,90, 1,10)	-0,2		1,00 (0,88, 1,12)	0,0	
15 Perspektiv	1,887	0,174		1,01 (0,90, 1,10)	0,3		1,02 (0,90, 1,10)	0,4	
16 Brug af tek,,	0,819*	0,092		1,05 (0,90, 1,10)	1,0		1,05 (0,90, 1,10)	0,9	
18 Identificer,,	2,175	0,182		0,75 (0,90, 1,10)	-5,3		1,05 (0,69, 1,31)	0,4	
19 Innovativ i,,	2,897	0,235		3,50 (0,90, 1,10)	30,3		1,08 (0,56, 1,44)	0,4	
20 Begrundelse,,	1,158*	0,135		0,69 (0,90, 1,10)	-6,8		1,03 (0,81, 1,19)	0,3	
22 Forløbsbev,,	0,736	0,087		0,72 (0,90, 1,10)	-6,1		0,88 (0,88, 1,12)	-2,1	
23 Antal aktan,,	0,823	0,089		0,85 (0,90, 1,10)	-3,1		0,90 (0,88, 1,12)	-1,7	
24 Analyse i e,,	-1,559*	0,100		1,40 (0,90, 1,10)	7,0		1,17 (0,92, 1,08)	3,8	
An asterisk next to a parameter estimate indicates that it is constrained									
Separation Reliability = 0,996									
Chi-square test of parameter equality = 3682,14, df = 16, Sig Level = 0,000									
[^] Empirical standard errors have been used									

Tabel 5. Resultaterne af Rasch-analyse, item*step-parametre estimeret med ConQuest.

TERM 2: item*step				UNWEIGHTED FIT			WEIGHTED FIT		
VARIABLES									
item	category	ESTIMATE	ERROR^	MNSQ	CI	T	MNSQ	CI	T
1	Andel korre,, 0			1,24 (0,90, 1,10)	4,3		1,06 (0,93, 1,07)	1,6	
1	Andel korre,, 1	-0,923	0,097	1,09 (0,90, 1,10)	1,8		0,99 (0,95, 1,05)	-0,3	
1	Andel korre,, 2	0,923*		10,37 (0,90, 1,10)	69,2		1,09 (0,86, 1,14)	1,2	
4	Antal unikk,, 0			0,74 (0,90, 1,10)	-5,6		0,95 (0,86, 1,14)	-0,7	
4	Antal unikk,, 1	-1,022	0,101	0,97 (0,90, 1,10)	-0,5		0,98 (0,94, 1,06)	-0,7	
4	Antal unikk,, 2	1,022*		0,95 (0,90, 1,10)	-1,0		0,97 (0,91, 1,09)	-0,6	
6	Læsbarhed 0			1,00 (0,90, 1,10)	0,0		0,88 (0,76, 1,24)	-1,0	
6	Læsbarhed 1	-0,819	0,131	1,07 (0,90, 1,10)	1,3		1,04 (0,90, 1,10)	0,8	
6	Læsbarhed 2	0,819*		1,04 (0,90, 1,10)	0,8		1,04 (0,89, 1,11)	0,6	
7	Vokalbrug 0			0,38 (0,90, 1,10)	-15,8		0,83 (0,76, 1,24)	-1,5	
7	Vokalbrug 1	-2,359	0,121	0,97 (0,90, 1,10)	-0,5		0,98 (0,93, 1,07)	-0,5	
7	Vokalbrug 2	2,359*		0,97 (0,90, 1,10)	-0,5		1,00 (0,91, 1,09)	0,0	
8	Staveudvikl,, 0			0,23 (0,90, 1,10)	-22,1		0,78 (0,71, 1,29)	-1,6	
8	Staveudvikl,, 1	-4,361	0,187	0,56 (0,90, 1,10)	-10,0		0,87 (0,77, 1,23)	-1,1	
8	Staveudvikl,, 2	4,361*		0,41 (0,90, 1,10)	-14,7		1,00 (0,60, 1,40)	0,0	
10	Aktører 0			0,73 (0,90, 1,10)	-5,7		0,94 (0,89, 1,11)	-1,1	
10	Aktører 1	-1,330	0,093	1,00 (0,90, 1,10)	0,1		0,99 (0,95, 1,05)	-0,2	
10	Aktører 2	1,330*		0,83 (0,90, 1,10)	-3,5		0,99 (0,91, 1,09)	-0,3	
11	Brug af led,, 0			1,08 (0,90, 1,10)	1,5		1,02 (0,90, 1,10)	0,4	
11	Brug af led,, 1	-0,217	0,108	0,86 (0,90, 1,10)	-2,7		0,97 (0,90, 1,10)	-0,7	
11	Brug af led,, 2	0,217*		1,52 (0,90, 1,10)	8,6		1,18 (0,85, 1,15)	2,3	
12	Antal eleme,, 0			1,02 (0,90, 1,10)	0,3		1,00 (0,85, 1,15)	-0,0	
12	Antal eleme,, 1	-1,283	0,085	1,02 (0,90, 1,10)	0,4		1,02 (0,97, 1,03)	1,3	
12	Antal eleme,, 2	1,283*		1,07 (0,90, 1,10)	1,3		1,05 (0,93, 1,07)	1,3	
13	Læsbarhed 0			0,74 (0,90, 1,10)	-5,5		0,97 (0,60, 1,40)	-0,1	
13	Læsbarhed 1	-0,959	0,138	1,00 (0,90, 1,10)	-0,1		1,01 (0,93, 1,07)	0,4	
13	Læsbarhed 2	0,959*		0,99 (0,90, 1,10)	-0,1		1,01 (0,93, 1,07)	0,3	
15	Perspektiv 0			1,02 (0,90, 1,10)	0,4		1,02 (0,93, 1,07)	0,6	
15	Perspektiv 1	-1,543	0,218	1,02 (0,90, 1,10)	0,4		1,02 (0,93, 1,07)	0,6	
15	Perspektiv 2	1,543*		0,90 (0,90, 1,10)	-1,9		0,98 (0,23, 1,77)	0,1	
23	Antal aktan,, 0			0,81 (0,90, 1,10)	-3,9		0,88 (0,90, 1,10)	-2,4	
23	Antal aktan,, 1	-0,920	0,129	0,88 (0,90, 1,10)	-2,4		0,94 (0,92, 1,08)	-1,6	
23	Antal aktan,, 2	0,920*		0,65 (0,90, 1,10)	-7,9		1,01 (0,81, 1,19)	0,1	

An asterisk next to a parameter estimate indicates that it is constrained

Et andet mål for kvaliteten af måledimensionerne er de såkaldte *MLE Person separation reliability* og *EAP/PV reliability*. Disse ligger på følgende for dimension 1: 0,642 og 0,571, for dimension 2: 0,710 og 0,644, for dimension 3: 0,392 og 0,445 og for dimension 4: 0,206 og 0,587. Tallene er tilfredsstillende for dimension 1 og 2, mens de er lidt små for dimension 3 og 4. Dette kan blandt andet skyldes at få elever har været i stand til at score højt på disse skalaer, og at der er forholdsvis få items i dimensionerne. Disse to dimensioner bør derfor undersøges nærmere i senere projekter, evt. med ældre elever.

I tabel 6 ses et såkaldt Wright- eller item-map som på venstre side viser elevernes dygtigheder på en logaritmisk skala fra -7 til +6 som krydser. Jo flere elever, des flere krydser (hvert kryds repræsenterer 3,4 elever). På højre side ses opgavernes sværhedsgrader på samme skala.

Det fremgår at items der måler korrekthed og innovativ fremstillingskompetence, fordeler sig bredt over skalaen, således at der er items som er lette for de fleste elever, og items som er svære for de fleste. Det ville have været at foretrække med lidt flere items i midterfeltet (så eleverne i mellemgruppen kunne differentieres med større sikkerhed). Tegne- og procesitems fordeler sig nogenlunde, men dog i den svære ende for de testede elever i første klasse. Instrumentet vil således

sandsynligvis egne sig bedre til at måle lidt ældre elevers tegne- og proceskompetencer. Men som det skal underbygges nedenfor, så måler instrumentet en udvikling over de fem måneder der går, og det er derfor tilstrækkeligt nuanceret til at kunne bruges også af elever i første klasse.

Tabel 6. Kort over fordeling af elevdygtigheder i de fire dimensioner sammenholdt med opgavernes sværhedsgrader.

=====				
Danish 1st grade writing, drawing and process competences tWed Nov 30 19:23 2016				
MAP OF LATENT DISTRIBUTIONS AND THRESHOLDS				
=====Build: Feb 25 2016=====				
Dimension			Generalised-Item Thresholds	
1	2	3	4	
6				
X			8,2	
X				
X				
5	X			
XX				
XX				
4	XXXX			
XXXX				
XXXXXX			15,2	
XXXXXX			X 1,2	
3	XXXXXX		X 19	
XXXXXX			X	
XXXXXXXX			X 7,2	
XXXXXXXXXX			XX 18	
2	XXXXXXXXXX		XXX 23,2	
XXXXXXXXXXXXXXXX	X		XXX	
XXXXXXXXXXXXXXXX	XX		XXXXXXXX 1,1	
1	XXXXXXXXXXXXXXXX	XXX	XXXXXXXX 14 20	
XXXXXXXXXXXXXXXX	XXXXX	XXXXX	XXXXXXXX 16 22	
XXXXXXXXXXXX	XXXXXX	XXXXXXXXXXXX	XXXXXXXX	
XXXXXXXXXXXX	XXXXXX	XXXXXXXXXXXX	XXXXXXXX 12,2 15,1	
0	XXXXXXXXXXXX	XXXXXXXXXXXX	XXXXXXXXXXXX 6,2 9 11,2	
XXXXXXXXXXXX	XXXXXXXXXXXX	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	XXXXXXXXXXXX 23,1	
XXXXXXXXXXXX	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	XXXXXXXXXXXXXXXXXXXX	XXXXXXXXXXXX	
-1	XXXXX	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	XXXXXXXXXXXX 10,2	
XXXXXX	XXXXXXXXXXXX	XXXXXXXXXXXXXXXXXXXX	XXXXXXXXXXXX 11,1	
XXXXX	XXXXXXXXXXXX	XXXXXXXXXXXXXXXXXXXX	XXXXXXXXXXXX 13,2 24	
-2	XXX	XXXXXXXXXXXX	XXXXXXXXXXXX	
XX	XXXXXXXXXXXX	XXXXX	XXXXXXXXXXXX 4,2 6,1	
X	XXXXXXXXXXXX	X	XXXXXXXXXXXX 7,1	
X	XXXXXXXXXXXX	X	XXXXXXXXXXXX 12,1	
-3	X	XXXXX	XXXXXXXXXXXX 5 8,1	
	XXXXXXXXXXXX		XXXXXXXXXXXX	
	XXXXXXXXXXXX		XXXXXXXXXXXX 10,1 13,1	
-4		XXXXX	XXXXX	
	XXXXX		XXXXX 4,1	
-5		XXXXX	XX	
	XXXXX		XX	
-6		XX	X	
	X			
-7				
=====				
Each 'X' represents 3,4 cases				
The labels for thresholds show the levels of item, and category, respectively				

Elevers kompetencer målt med instrumentet

Af tabel 7 fremgår elevernes gennemsnitlige scorer fordelt på indsats- og kontrolgruppe ved baseline og endline.

Tabel 7. Oversigt over elevernes score på de fire dimensioner fordelt på indsats- og kontrolgruppe ved baseline og endline.

Dimension	Gruppe	Baseline		Endline		Udvikling	
		Gennemsnit	Standard-afvigelse	Gennemsnit	Standard-afvigelse	Gennemsnit	Standard-afvigelse
Skrivefærdighed	Indsats	0,71	1,57	1,38	1,41	0,80	1,87
	Kontrol	0,66	1,66	1,30	1,45	0,74	1,86
Innovativ fremstillingskompetence	Indsats	-2,56	1,78	-1,62	1,58	1,09	2,02
	Kontrol	-2,61	1,81	-2,07	1,66	0,73	2,02
Tegnekompetence	Indsats	-0,76	0,67	-0,48	0,65	0,32	0,83
	Kontrol	-0,83	0,69	-0,63	0,63	0,23	0,81
Proceskompetence	Indsats	-1,55	2,10	-0,77	2,00	0,91	2,57
	Kontrol	-1,83	2,16	-1,17	2,15	0,86	2,50

Note. N: Baseline: Indsats: 276. Kontrol: 181. Endline: Indsats: 268. Kontrol: 172. Udvikling: Indsats: 238. Kontrol: 144. I beregning af udvikling indgår kun elever som har deltaget i både baseline og endline.

Undersøgelse af udvikling

Samlet set er der udviklet et redskab som ser ud til at kunne måle træk ved fire væsentlige kompetencer.

Målet med denne del af Hitte På-projektet var at undersøge om der kunne udvikles et instrument der kunne måle om de deltagende elever udviklede kompetencerne i højere grad end kontrolklassernes elever. Dette er emnet for dette kapitel.

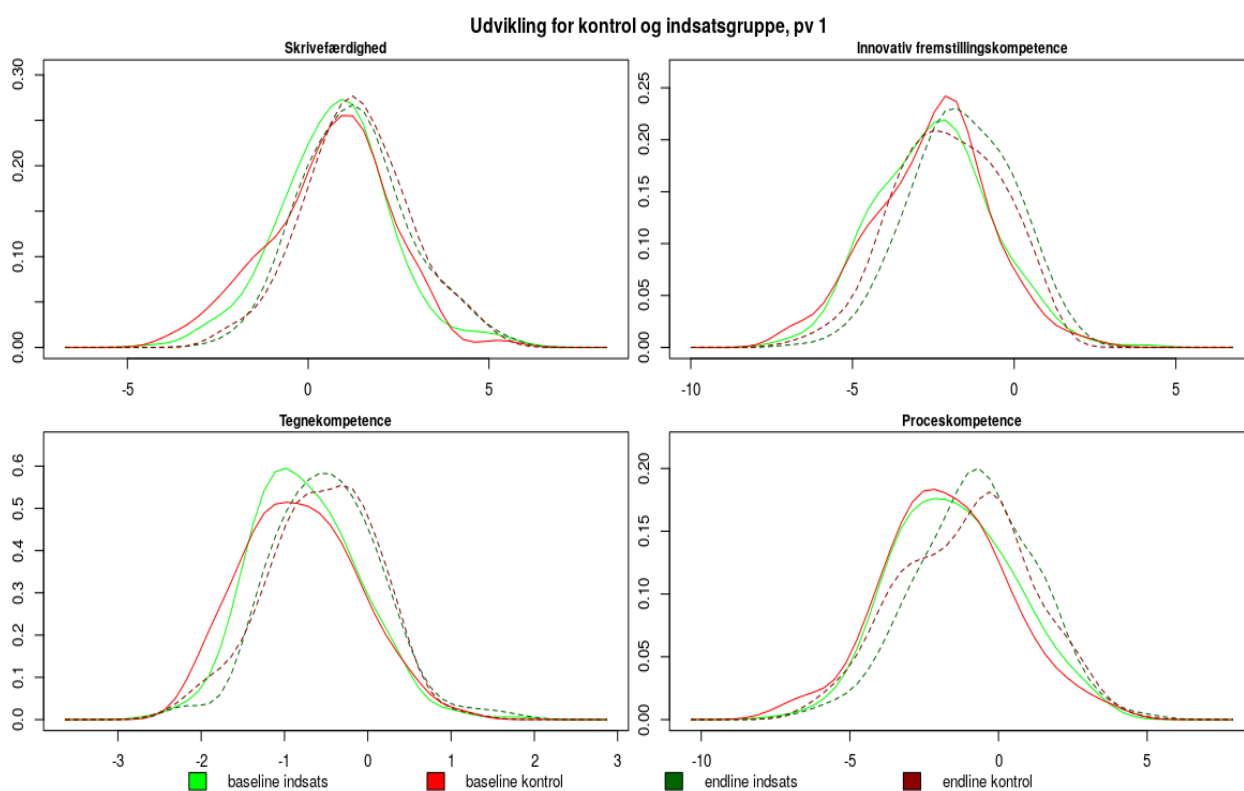
Undersøgelsen er designet som en prætest-posttest-undersøgelse (målinger ved baseline og endline) med indsatsgruppe og kontrolgruppe. De fem deltagende skoler (med 13 deltagende klasser) er alle fra Kolding. Skolerne meldte sig til projektet og blev udvalgt til deltagelse efter lodtrækning. De fire skoler som ikke deltog, fungerede som kontrolskoler (med 8 klasser). Dette er et centralt træk ved et såkaldt randomiseret kontrolleret forsøg. Men da vi ikke strikt håndhævede indsatsklassernes brug af Hitte På-materialet, og da vi heller ikke havde kontrol med eller nærmere viden om kontrolklassernes praksis i projektets løbetid, vil vi ikke karakterisere dette projekt som et ægte randomiseret kontrolleret forsøg.

Da der er tale om selvudvalgte skoler fra en enkelt dansk kommune, kan der heller ikke på baggrund af undersøgelsen generaliseres til populationen af danske elever.

Da målet ikke var at gennemføre en generaliserbar undersøgelse, men at udvikle et instrument der kunne bruges til at vurdere indsatsens muligheder, og som desuden ville kunne bruges ved fremtidige undersøgelser, og for ikke at belaste skoler og elever unødigt, indsamlede vi ikke oplysninger om køn, socioøkonomisk baggrund, etnicitet eller lignende baggrundsvARIABLE som kunne bruges til at undersøge om de to grupper var ens i udgangspunktet. Men en t-test og et QQ-plot af data fra baseline viser at der ikke er signifikante forskelle i scorene for eleverne fra indsats- og kontrolgruppe i nogen af de fire dimensioner.

Elevers kompetencer kan ikke måles præcist – der vil altid være en eller anden grad af usikkerhed på målet. Ved behandlingen af resultater af tests statistisk skal man finde en måde at håndtere denne usikkerhed på. Når man skal sammenligne grupper, har det vist sig (von Davier, Gonzalez, & Mislevy, 2009) at give mere præcise udtryk for standardfejl og varians hvis man i stedet for én værdi for hver elev, udtrækker et antal (typisk 5) plausible værdier (*plausible values*) for den enkelte elevs kompetence. Sådanne værdier er tilfældige udtræk fra det normalfordelte sandsynlige resultat for den enkelte elev. Ved de statistiske analyser udregnes resultaterne for hver af disse plausible værdier, hvorefter man tager gennemsnittet af resultaterne. Derved får man mere sande værdier for varians og standardfejl, end hvis man havde anvendt én estimeret værdi for elevernes dygtigheder.

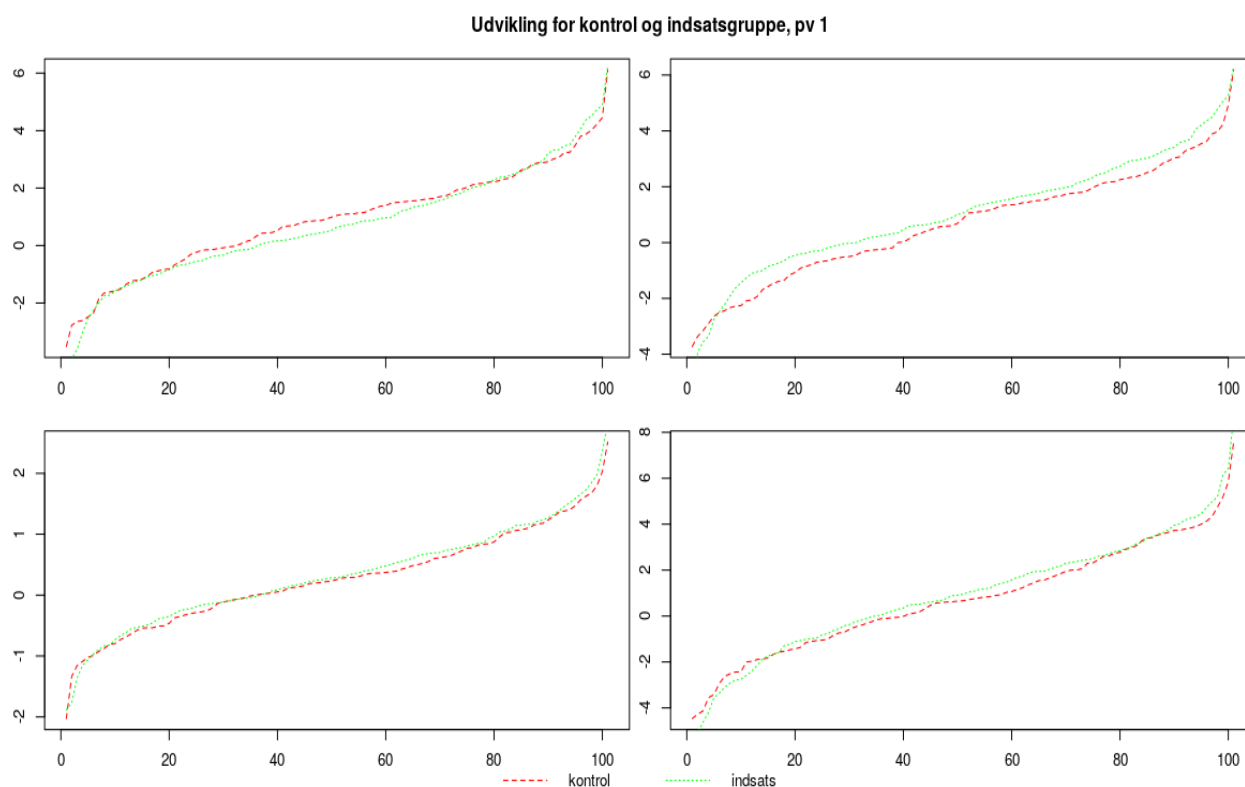
I figur 2 ses et overblik over hvordan udviklingen har været for eleverne i kontrol- og indsatsgrupperne fra base- til endline. Denne figur viser resultaterne for den første plausible værdi (PV). På grund af den særlige måde PV-erne er udtrukne, ser figurerne lidt anderledes ud for de andre værdier, men tendensen er den samme.



Figur 2, Udvikling for kontrol- og indsatsgruppe for de fire dimensioner. Vist for den første plausible værdi.

Note: N: Indsatsgruppe: baseline: 276, endline: 268, Kontrolgruppe: baseline: 181. endline: 172.

Som det fremgår bliver eleverne samlet set bedre i alle fire dimensioner. Det kan også anes at eleverne fra indsatsgruppen udvikler sig lidt mere end eleverne fra kontrolgruppen i den innovative fremstillingskompetence. Det bliver lidt mere synligt i figur 3 hvor de enkelte elevers udvikling er plottet ind sorteret fra mindste til største udvikling. Som det fremgår har nogle elever udviklet sig negativt – hvilket blandt andet skyldes at kompetencemålingsinstrumenter som ovenfor omtalt altid vil have en vis grad af usikkerhed forbundet med målingen.



Figur 3, Udvikling målt som elevernes resultat i endline minus resultat i baseline. Vist for PV 1.

Note: N=144 i kontrolgruppen og 238 i indsatsgruppen, Da der er flere elever i indsatsgruppen end i kontrolgruppen, er der først udtrukket en værdi fra hver kvartil i begge grupper.

For at afgøre om der er tale om en statistik signifikant forskel på udviklingen for indsatsgruppen og kontrolgruppen, foretages en multilevel-regressionsanalyse, som undersøger om forskellen på score ved baseline og endline varierer signifikant med deltagelse i indsatsen. En multilevel-regressionsanalyse tager højde for at eleverne modtager interventionen klassevis og derfor sandsynligvis vil dele en række fælles forudsætninger (tidligere undervisning, træk hos deres lærere, klassekultur mv.) som vil gøre at de vil "følges ad" med hinanden. Da der var flere klasser på nogle skoler, undersøgte vi om en model med klasser i skoler ville give bedre statistiske egenskaber, men AIC-kriteriet var på samme niveau både med og uden et skoleniveau, og analysen er derfor kørt alene med to niveauer (eleverne og klasserne).

I regressionsformlen indgår ud over indsats-variablen også elevernes score ved baseline. Det betyder at regressionen kontrollerer for elevernes score ved baseline, således at variation der skyldes forskelle i scores ved baseline, ikke får indflydelse på vores vurdering af indsatsens betydning.

Antagelserne for regressionsanalysen (Osborne & Waters, 2002) er testet med succes i forhold til normalitetsantagelsen med Q-Q-plots, i forhold til antagelsen om variabelnes lineære relation og homoscedasticitet med plot af standardiserede residualer.

Resultaterne af multilevel-regressionsanalyserne fremgår af tabel 7.

Tabel 7. Resultater af multilevel-regressionsanalyser for de fire dimensioner.

Skrivefærdighed

	<i>Estimat</i>	<i>Standardfejl</i>	<i>t-værdi</i>
(Intercept)	1,21	0,17	6,94
baseline_dim1	-0,81	0,04	-18,38
indsats	0,19	0,23	0,87

<i>Gruppe</i>	<i>Varians</i>	<i>Standardafvigelse</i>
klasseid	0,14	0,37

Innovativ skrivekompetence

	<i>Estimat</i>	<i>Standardfejl</i>	<i>t-værdi</i>
(Intercept)	-1,36	0,21	-6,61
baseline_dim2	-0,76	0,04	-17,00
indsats	0,48	0,21	2,27

<i>Gruppe</i>	<i>Varians</i>	<i>Standardafvigelse</i>
klasseid	0,09	0,30

Tegnekompetence

	<i>Estimat</i>	<i>Standardfejl</i>	<i>t-værdi</i>
(Intercept)	-0,48	0,09	-5,51
baseline_dim3	-0,82	0,05	-16,92
indsats	0,19	0,10	1,94

<i>Gruppe</i>	<i>Varians</i>	<i>Standardafvigelse</i>
klasseid	0,03	0,16

Proceskompetence

	<i>Estimat</i>	<i>Standardfejl</i>	<i>t-værdi</i>
(Intercept)	-0,69	0,23	-3,08
baseline_dim4	-0,77	0,05	-15,74
indsats	0,38	0,26	1,47

<i>Gruppe</i>	<i>Varsians</i>	<i>Standardafvigelse</i>
<i>klasseid</i>	0,12	0,34

T-værdien for estimatet for indsatsen er større end 2 for den innovative skrivekompetences vedkommende, hvilket betyder at der er 95% sandsynlighed for at resultaterne for indsats- og kontrolgruppen er forskellige med indsatsgruppens udvikling som den største. For de øvrige kompetencers vedkommende er der også tale om en numerisk større udvikling for indsatsgruppens vedkommende, men den er ikke statistisk signifikant, og man må derfor konkludere at der ikke kan iagttages en signifikant forskel på kontrolgruppen og indsatsgruppens elever for disse tre kompetenceområder. For tegneområdet vedkommende er der dog tale om en værdi meget tæt på 2, og man kan derfor kalde udviklingen for indsatsgruppen for borderline signifikant bedre end for kontrolgruppen.

For alle fire dimensioners vedkommende viser der sig en negativ korrelation mellem baselinescore og udvikling. Det betyder konkret at jo højere eleverne har scoret, des mindre (forholdsmæssigt) er deres udvikling. Eller sagt på en anden måde: Elever med lavere scorer udvikler sig mest. Dette fænomen ses også i fx nationale test, og skyldes den logistiske skala i Rasch-scoren.

Størrelsen på forskellen kan udregnes som en såkaldt effektstørrelse. En effektstørrelse er et mål for hvor meget eleverne har flyttet sig i forhold til en standardafvigelse. Hvis man flytter sig én standardafvigelse fra midten af elevgruppen, bliver man bedre end 34,1 procent elever (og omvendt hvis man bliver en standardafvigelse dårligere). Effektstørrelser udregnes som gennemsnittet af udvikling divideret med en standardafvigelse. Når der sammenlignes mellem en indsatsgruppe og en kontrolgruppe, findes der forskellige tilgange til beregning af effektstørrelse, idet standardafvigelsen kan være forskellig for indsats- og kontrolgruppen samt ved præ- og posttest. I et Monte Carlo-studie viser Morris (2008) at den mest præcise formel til beregning af effektstørrelse indebærer at den pooled standardafvigelse fra baseline anvendes (Morris, 2008, s. 384).

De tre ikke statistisk signifikante effektstørrelser er på 0,07 for skrivefærdighed og 0,08 for proceskompetence, mens den er 0,18 for tegning. For innovativ fremstillingskompetence er effektstørrelsen på 0,28. I betragtning af at interventionerne var forholdsvis små og få i en forholdsvis kort periode, vurderer vi at denne effektstørrelse er markant og substantiel.

Konklusioner og perspektiver

Der er med Hitte På-projektet udviklet et instrument der på kort tid kan måle centrale træk af fire ganske væsentlige kompetenceområder i form af elevers skrivefærdigheder, innovative fremstillingskompetencer, tegne- og proceskompetencer. Ingen af de fire kompetencer er så vidt vides målt på tilsvarende vis tidligere. Der viser sig klare og signifikante udviklinger for alle eleverne over det lille halve år der er mellem målingerne.

Der viser sig desuden signifikant bedre udvikling hos indsatsgruppen på innovativ fremstillingskompetenceområdet. Vi vurderer dette som en klar indikator på at Hitte På-pædagogikken har fremmet nogle af de intendede entreprenørskab- og innovationskompetencer hos eleverne.

Lige så vigtigt er det måske at eleverne ikke har udviklet sig mindre end kontrolgruppen på skrivefærdighedsområdet. Så det er lykkedes at tilrettelægge en innovativ og motiverende undervisning der samtidig understøtter udvikling af basale færdigheder. Det skal dog understreges at vi ikke har viden om hvilken undervisning der er foregået i kontrolklasserne, og at vi derfor ikke kan vide om der her er foregået undervisning sådan som den almindeligvis foregår.

Det kan på denne baggrund anbefales at gennemføre et lignende projekt, men næste gang involvere flere elever, lade interventionerne være mere intensive og vare længere.

Til slut vil jeg kort berøre to faglige pointer som udviklingen af redskabet giver anledning til.

For det første har instrumentet vist at skriveudviklingsfaser udgøres af andet end staveudvikling, og at der ikke direkte er tale om det samme fænomen. Testen har desuden givet indblik i hvilke opgaver der er svære (og som eleverne derfor først sent bliver kompetente til at løse), og hvilke der er lette. En vigtig måde at arbejde videre med dette redskab på er at analysere kategorierne yderligere og derigennem opstille et empirisk funderet skriveudviklingsforløb der omfatter mere end staveudvikling. Ud fra tabellerne 3, 4 og 5 kan man fx se at det at sætte mellemrum og bruge nogenlunde korrekte vokaler kommer tidligt, brug af ledsætninger typisk følger senere, og mere avancerede verber først kommer ganske sent i skriveudviklingsprocessen.

Dette vil være et værdifuldt redskab til lærere når de skal vurdere hvad deres elever arbejder på, og hvordan de derfor bedst kan gives respons. Det vil desuden kunne bruges til at kvalificere læremidler der har fokus på skriveundervisning i indskolingen.

Den anden faglige pointe er at skrivefærdighed, herunder staveudvikling som vi målte det, viste sig ikke at være en del af det samme latente træk som det vi har kaldt for innovativ fremstillingskompetence. Faktisk er korrelationen mellem skrivefærdighed og innovativ fremstillingskompetence kun på 0,39, mens korrelationerne mellem innovativ fremstillingskompetence og tegning hhv. proceskompetence er på 0,54 og 0,71. Dette underbygger den opfattelse at stavning ikke i sig selv er et udtryk for hvor langt eleverne er i skrivekompetenceudviklingen.

Vi bemærkede flere gange at vores umiddelbare vurdering af en teksts kvaliteter var forbundet med vores oplevelse af umiddelbar læsbarhed (dvs. hvor tæt teksten var på konventionel stavning) (og tildels også af håndskriftens kvalitet), men ofte blev vi overraskede over de mange kvaliteter en tekst kunne have, uanset hvordan stavekvaliteten var. Disse to pointer – en statistisk og en erfaringsbaseret – giver anledning til at anbefale at lærere er opmærksomme på ikke i for høj grad at bedømme elevers skriveudviklingsfase på baggrund af deres staveudviklingsfase.

Referencer

- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht: Springer Netherlands. Hentet fra <http://link.springer.com/10.1007/978-94-007-6857-4>
- Bundsgaard, J., & Fougst, S. S. (Forthcoming). Faglighed og scenariedidaktik. I *Scenariedidaktik*. Aarhus: Aarhus Universitetsforlag.
- Hannibal, S., Korsgaard, K., & Vitger, M. (2010). *Opdagende skrivning - en vej ind i læsningen*. København: Dan sklærerforeningen.

- Linacre, J. M. (u.å.). What do Infit and Outfit, Mean-square and Standardized mean? Hentet 1. december 2016, fra <http://rasch.org/rmt/rmt162f.htm>
- Lützen, P. H. (2016). *Kvalitativ evaluering af Hitte På-projektet* (s. 9). København: Videncenter for Læsning.
- Morris, S. B. (2008). Estimating Effect Sizes From Pretest-Posttest-Control Group Designs. *Organizational Research Methods*, 11(2), 364–386.
<https://doi.org/10.1177/1094428106291059>
- Osborne, J., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical assessment, research & evaluation*, 8(2), 1–9.
- Osteen, P. (2010). An Introduction to Using Multidimensional Item Response Theory to Assess Latent Factor Structures. *Journal of the Society for Social Work and Research*, 1(2), 66–82.
<https://doi.org/10.5243/jsswr.2010.6>
- Rasmussen, A., & Moberg, K. (2016). *Taksonomi i Entreprenørskabsuddannelse - Perspektiver på mål, undervisning og evaluering* (2. udgave). Odense: Fonden for Entreprenørskab. Hentet fra <http://www.fey.dk/media/555474/taksonomi20i20entrepren3b8rskabsuddannelse20220udgave20onlineversion.pdf>
- von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? I *IERI Monograph Series Volume 2* (s. 9–36).
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8 (3), 370. *Rasch Measurement Transactions*, 8(3).