Høgskolen i Innlandet

Jo Kleiven, Linda Sandholt og Per Normann Andersen

# Norsk versjon av «Yellow Red»

**- et testprogram på Android nettbrett for eksekutive funksjoner hos barn.**

Skriftserien 18 - 2022

Trykk: Flisa Trykkeri A/S

Utgivelsessted: Elverum



Forfatteren er selv ansvarlig for sine konklusjoner. Innholdet gir derfor ikke nødvendigvis uttrykk for høgskolens syn.

I Høgskolen i Innlandets skriftserie publiseres både internt og eksternt finansierte FoU-arbeider.

Skriftserien nr. 18-2022

ISBN trykt utgave: 978-82-8380-360-0

ISBN digital utgave: 978-82-8380-361-7

ISSN: 2535-5678

# Sammendrag

Yellow Red-testen tar sikte på å måle noen sider ved barns eksekutive funksjoner. Den er utformet som fire dataspill for bruk på Android-nettbrett. Selv om spillene opprinnelig var laget for bruk på ett enkelt barn, kan den administreres i mindre grupper. De fleste barn lar seg engasjere i spillene, og gjennomfører som planlagt.

Resultatene fra en pilotundersøkelse virker i all hovedsak tillitvekkende. Svarfordelingene samsvarer f.eks. både med rimelige forventninger om normalfordeling, og med vanlige endringer og modning hos barn.

Testen er rettighetsbelagt. Den er tilgjengelig som en nedlastbar app, men krever passord for å la seg bruke. På kort sikt er det derfor neppe hensiktsmessig å søke etter forbedringer av testprosedyren. Kanskje bør man heller se nærmere på de nåværende reglene for skåring av responsene. Forsøk med alternative kodingsalgoritmer kan muligens trekke ut annen relevant og interessant informasjon.

**Emneord:** Eksekutive funksjoner, testing, skåring

**Oppdragsgiver:** Oppland fylkeskommune

# Abstract

The Yellow Red-test is intended to assess aspects of children's executive functions. It is designed as four data games and is available for Android data tablets. Originally planned for use on individual children, the games may nonetheless be adminstered to small groups. Most children find the games challenging, and will complete them as intended.

The results from a pilot study are by and large encouraging. The response distributions appear trustworthy, matching, e.g., reasonable expectations of normality as well as known and common changes and maturation in children.

The games are designed a copyrighted app. It is freely downloadable, but requires a password to function. In the short run, therefore, changing the procedures may not be an efficient approach to improving the test. Focusing on the rules for scoring of the responses may be a more promising approach. Experimenting with alternate coding algorithms may yield relevant and interesting information.

**Keywords:** Executive functions, testing, scoring

**Financed by:** Oppland county

# Forord

Pilotprosjektet «Kunsten å lære» (Hundevadt & Klausen, 2019) ble gjennomført i perioden 2017-2019, finansiert av bl.a. Oppland fylkeskommune. Fra dette arbeidet fikk psykologmiljøet ved høgskolen erfaring med bruk av *Yellow Red*-testen, og har bygget opp et datamateriale fra et utvalg med barn som er testet tre ganger.

Testen tas nå i fortsatt bruk i det såkalte KÅL-prosjektet, som bl.a. skal undersøke hvilke virkninger et kunstbasert skoleprogram har på barns eksekutive funksjoner. Men videre bruk av testen er ikke bare planlagt i Norge, men også i andre land.

Det kan derfor være behov for informasjon om denne testen, og ikke minst om den norske utgaven. Denne rapporten gir derfor en kort gjennomgang av hva testen inneholder, og deretter noen første analyser av de tilgjengelige data.

Testens chilenske opphavsmann, prof. Ricardo Rosas, legger også vekt på at *Yellow Red* er et instrument under fortsatt utvikling, og derfor ikke nødvendigvis har fått noen endelig utforming. Kanskje kan vår gjennomgang av testmetoden også komme til nytte i denne sammenhengen.

Lillehammer, april 2022

Jo Kleiven  Linda Sandholt  Per Normann Andersen

# Innholdsfortegnelse

# 1. KORT HISTORIKK

Den såkalte Yellow Red-applikasjonen tar sikte på å måle eksekutive funksjoner hos barn. Ifølge Diamond (2013) dekker dette begrepet noen overordnede mentale prosesser, som er avgjørende når man konsentrerer seg og er særlig oppmerksom. Det er særlig tre slike prosesser som nevnes i litteraturen (Lehto et al., 2003): Inhibisjon, arbeidsminne og kognitiv fleksibilitet.

*Centro UC Tecnologias de Inclusión* (Sentret for inkluderende teknologi) ved det psykologiske instituttet på *Pontificia Universidad Católica de Chile* (Det Katolske Universitetet i Chile) har lenge arbeidet med computerbasert testing av barn (Margolis et al., 2006; Rosas et al., 2003; Tenorio et al., 2014). Den opprinnelige utformingen av applikasjonen var derfor på spansk, og man ønsket å måle de eksekutive funksjonene på en måte som er relativt lite avhengig av språkferdigheter og som kan brukes på enkle Android-nettbrett.

Dette arbeidet ble utført i prosjektet *«Yellow Red: Tablet based executive functions test for children between 7 and 10 years*», hvor de sentrale medarbeiderne er Ricardo Rosas, Victoria Espinoza og Marion Garolera. Prosjektet har som siktemål å samle tverrkulturell evidens om utviklingen av eksekutive funksjoner hos barn. Data er derfor ikke bare innsamlet i Chile, men også fra Argentina, Peru, Mexico, Tyskland, Skottland, England, Wales, Norge, India og Australia. Et internt arbeids-dokument fra 2019 (Garolera, 2019) rapporterer noen lovende data fra dette internasjonale prosjektet, bl.a. klart tilfredsstillende tall for Cronbach's alfa. Her skisseres det også mulige normer for bruk av testen. Noen data fra prosjektet ble også presentert på en konferanse i 2019 (Rosas-Días et al., 2019), og en foreløpig rapport fra prosjektet ble gitt ut i 2020 (Rosas et al., 2020).

Deler av testen har vist seg nyttige i tidligere forskning på barns utvikling av eksekutive funksjoner (Se f.eks. Rosas et al., 2017).

Den norske versjonen av testen ble utviklet i samarbeid med Per Normann Andersen ved Høgskolen i Innlandet. Det finnes også ungarske, tyske og engelske versjoner. Den norske versjonen er blitt brukt i en pilotundersøkelse av læringsutvikling hos norske barn (Hundevadt & Klausen, 2019), som under-søkte om læringsutviklingen blant norske barn kunne påvirkes av kunst i skolen. Her ble det samlet inn data før, under og etter barnas erfaringer med kunst som læringsform. Det er data fra denne pilotundersøkelsen vi har hatt tilgjengelige for videre analyser i denne omgang.

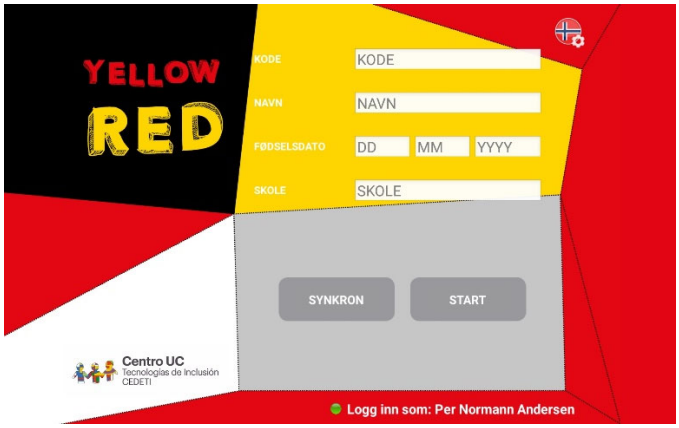
I en fagartikkel om denne undersøkelsen (Andersen et al., 2019) blir imidlertid data fra Yellow Red-testen ikke nevnt. Andersen et al. (2019) brukte imidlertid også BRIEF (Davidson et al., 2006), som er en helt annen test på eksekutive funksjoner. Den norske versjonen av BRIEF er mye brukt, og har tilfredsstillende måleegenskaper (Fallmyr & Egeland, 2011; Sørensen & Hysing, 2014). Det er derfor mulig å sammenholde data fra Yellow Red med data fra BRIEF i vårt materiale.

# 2. FEM TRINN I TESTEN

Testen har fem faser. Først kommer en innledende prosedyre og så fire test-deler. De fire test-delene er utformet som dataspill, og benevnes derfor som fire spill. Disse inneholder noe ulike oppgaver.

Barna som skal testes møtes i et klasserom, og får utdelt hvert sitt nettbrett. En instruktør styrer det hele, og gir løpende beskjeder om hva som skal gjøres underveis. For å sikre at alle barna får samme informasjon gis det instruksjoner i forkant med tydelige forklaringer på det som skal skje. Det gis også informasjon om hvor lang tid dette vil ta. Det kan variere mellom 40 og 50 minutter. Det gis beskjed om at det er fire spill som skal spilles, og at det vil bli gitt instruksjoner for hvert spill før det starter. Spillene har også øvingsoppgaver som gjør at elevene kan prøve seg først. Det er fokus på at vi gir hverandre arbeidsro mens vi spiller, og at vi venter med å begynne et nytt spill til instruksjoner er gitt.

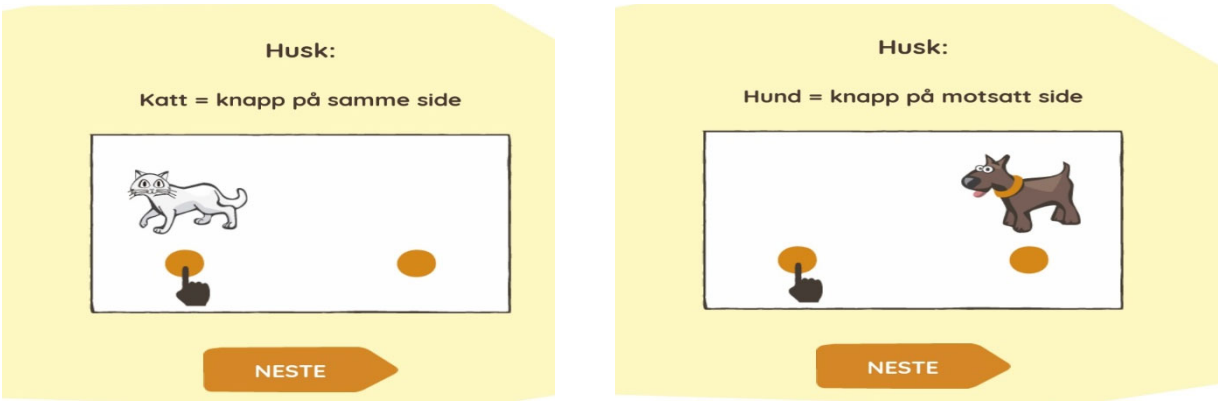
På skjermen vises først en påloggingsside hvor instruktør skal oppgi noen enkle opplysninger.




**Figur 1:** Påloggingsside for testen

## 2.1 Hund og katt

Det første spillet kalles *Hund og katt* på norsk. Den må trolig forstås som en *generell* test på eksekutive funksjoner, og bygger på «Hearts and flowers»-testen av Adele Diamond (Davidson et al., 2006; Wright & Diamond, 2014).




**Figur 2:** To skjermbilder fra instruksjonene til *Hund og katt*-oppgaven

### 2.1.1 Prosedyre

Først kommer en øvelsesfase med 'kongruente' oppgaver. Den konkrete arbeidsoppgaven her er å trykke på sirkelen under bildet av *katten*, altså på *samme side* som katten (som i det venstre bildet i figur 2). Så følger øvelser med 'inkongruente' oppgaver, hvor det vises bilder av en *hund* og gis beskjed om at man da skal trykke på knappen på *motsatt side* av hunden (som i det høyre bildet).

Etter disse øvelsesrundene følger det så 33 bilder hvor 'kongruente' og 'inkongruente' oppgaver er blandet, og hvor det raskt skal trykkes på riktig knapp. Her må man altså først se om det er hund eller katt, og så velge respons som følge av dette. Her vises hvert bilde i ett sekund, og hvert mellomrom er på ½ sekund.

Barna får selv ikke vite underveis om de svarer riktig eller feil, og får heller ingen informasjon om hvordan den samlede poengsummen regnes ut.

### 2.1.2 Skåring

Om det ikke gis noe svar mens bildet vises, gis det 0 poeng for dette bildet (*omissiones*). Riktig svar gir 1 poeng på bildet (*respuesta correcta*), mens feil svar gir -2 poeng (*respuesta incorrecta*). De 33 responsene summeres, og til slutt trekker man fra det antallet responser som ble gitt før 0,2 sekunder var gått (*respuestas anticipatorias*). Den endelige sumskåren er altså sammensatt på en noe kompleks måte.

Det kan også nevnes at Garolera (2019) rapporterer en meget høy reliabilitet på denne skalaen i data fra det internasjonale prosjektet. Cronbach's alfa oppgis der til 0.833.

## 2.2 Trio

### 2.2.1 Prosedyre

I det neste spillet er hensikten å måle *kognitiv fleksibilitet*. Her blir fire figurer vist samtidig, hvor tre av figurene ('trioen') har en felles egenskap (farge, form eller størrelse). Det gis ingen beskjed om *hva* som kan være felles. Den kritiske egenskapen endres stadig, og uten at det gis noen beskjed om dette. For hvert bilde som vises, går oppgaven ut på å finne de tre figurene som har noe til felles.

Det er 21 trinn i dette spillet. Det vises først fem bilder med *farge* som felles egenskap, så fem med *form*, og deretter fem med *størrelse* som det kritiske. Til slutt vises seks bilder med en blanding av de tre kritiske egenskapene. På hvert trinn får barna tre forsøk på å velge de tre figurene som har noe til felles. De som ikke lykkes, blir tatt videre til neste trinn. I dette spillet bruker barna høretelefoner, og dette er det eneste spillet med lyd. Ved riktig svar så vises det et fyrverkeri, og man hører lyden av et fyrverkeri. Ved feil svar så vil figurene knuse i mindre biter, og man vil høre lyden av knust glass. Hensikten med høretelefoner er at det bare er barnet selv som hører lyden av riktig eller feil svar.

**Figur 3:** Skjermbilde fra *Trio*-spillet

### 2.2.2 Skåring

Som nevnt får barna selv vite om feil gjennom både lyd og bilde. De får imidlertid ingen informasjon om hvordan den samlede poengsummen regnes ut.

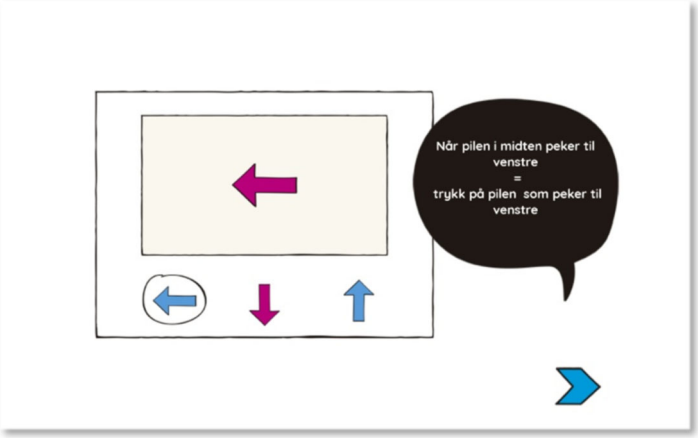
Men uløste oppgaver noteres som manglende data (SYSMIS). Ellers gis det 1,0 poeng når barnet treffer på første forsøk; 0,6 for andre; og 0,3 om man først lykkes på det tredje. Men om barnet faktisk gir tre feil responser så forstås dette som en '*error persevatido*', og straffes med -2 poeng. Poengene fra de 21 trinnene summeres til slutt til en samlet skåre for *Triader*. Også denne er altså en kompleks summert skåre, som er sammensatt av informasjon om flere ulike forhold.

Her oppgir Garolera (2019) en verdi for Cronbach's alfa på 0.782. Også dette må forstås som en klart akseptabel reliabilitet.

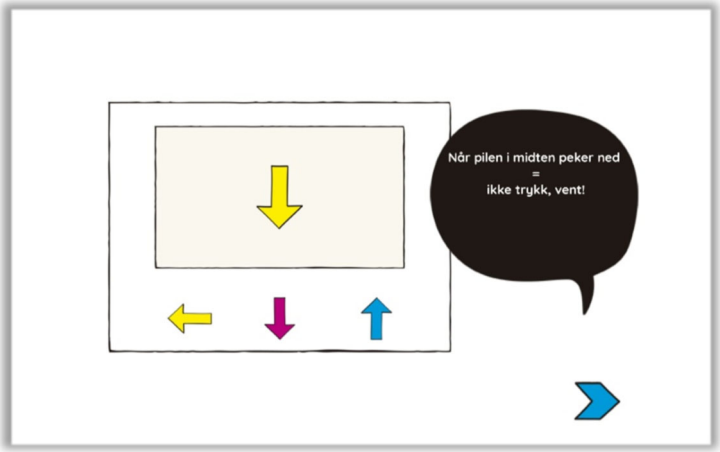
## 2.3 Piler

### 2.3.1 Prosedyre

Det tredje spillet dreier seg om *inhibisjon*, evnen til å holde tilbake en respons. Her vises ei stor, kritisk pil i ei ramme, og tre mindre piler utenfor ramma. Vanligvis skal barna trykke på den pila av de tre som peker samme vei (venstre, opp, eller høyre) som den store pila, som i figur 4 på neste side. Men dette gjelder ikke når pila peker *ned*, som i figur 5. Da er oppgaven å *ikke trykke* i det hele tatt, altså å holde tilbake responsen.




**Figur 4:** Skjermbilde fra *piler*-spillet, 'vanlig' instruksjon




**Figur 5:** Skjermbilde fra *piler*-spillet, 'avvikende' instruksjon

Før spillet starter, blir det gitt noen øvingsoppgaver. Her får elevene beskjed om de svarer rett eller galt.

Det er 36 oppgaver i selve spillet, og åtte av dem gjelder piler som peker ned. De første 15 oppgavene blir vist i vel 2 sekunder. De påfølgende 21 oppgavene blir vist i bare ett sekund, og mellomrommet mellom dem er ½ sekund. Responser som kommer mindre enn 0,2 sek etter oppgaven, blir ikke regnet med. Her får barna ingen beskjed om en respons er 'feil' eller 'riktig' – eller i hvilken grad de lykkes med dette spillet.

### 2.3.2 Skåring

Her gis det 1 poeng for hvert riktig svar, og -1 poeng for feil. Men når det gis respons til piler som peker ned, så straffes dette med -2 poeng. Den summerte skåren er altså også her sammensatt av informasjon om noe ulike forhold.

Også her synes reliabiliteten i de internasjonale dataene å være tilfredsstillende. Garolera (2019) oppgir en verdi for Cronbach's alfa på 0.796.

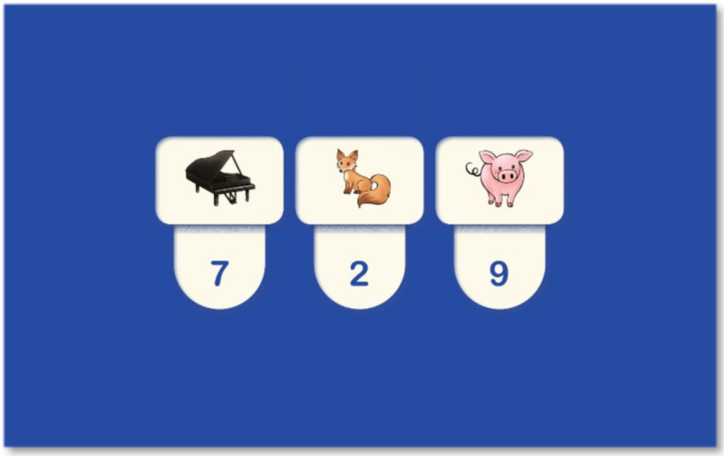
## 2.4 Tallforbindelser

Det fjerde spillet gjelder utviklingen av *arbeidsminnet*. Den gjennomgående utfordringen er å huske hvilke tall som er knyttet til bestemte ting eller gjenstander.

### 2.4.1 Prosedyre

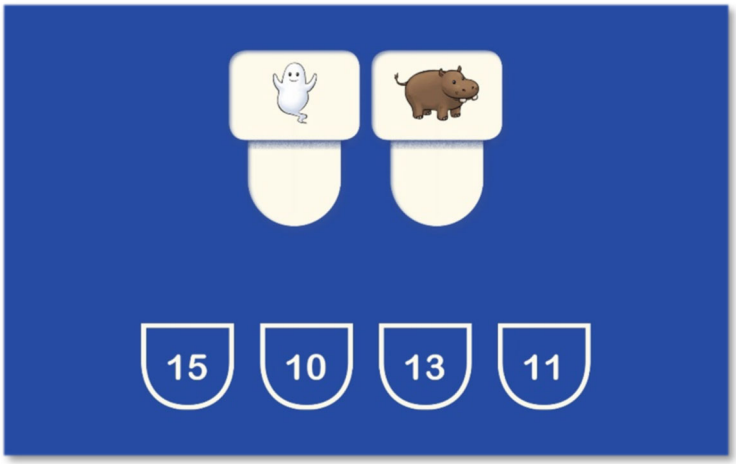
Også dette spillet starter med en øvingsoppgave, Barnet får her beskjed om det svarer rett eller galt, og det må svares korrekt for å komme videre. Etter øvingsoppgaven så gis det ikke beskjed om svaret er rett eller galt.

I selve spillet er det 27 oppgaver, med økende vanskegrad. For hver oppgave vises det derfor først et bilde med noen forskjellige ting. Til hver ting er det knyttet et bestemt tall, som i figur 6.




**Figur 6:** Skjermbilde fra *Forbindelser*-spillet

Deretter vises et bilde hvor (litt færre) ting og (litt flere) tall vises hver for seg, som i figur 7. Oppgaven gjelder så å dra/flytte det korrekte tallet bort til hver av tingene.




**Figur 7:** Skjermbilde fra *Forbindelser*-spillet

De første oppgavene er forholdsvis enkle. Barnet starter med å se to ulike bilder, og under bildene er det to ulike figurer eller tall. Her har barnet en fire sekunders svarfrist. Etter hvert blir spillet vanskeligere, og det vil dukke opp flere bilder med ulike kombinasjoner av figurer og etter hvert tall for de som kommer lengre i spillet. Spillet avsluttes når det er gitt tre feilresponser.

### 2.4.2 Skåring

Feilresponser gir alltid 0 poeng, men riktige svar belønnes i samsvar med den økende vanskegraden i oppgavene. For oppgavene 1 - 10 gis det derfor 1 poeng, for 11 - 22 gis det 2 poeng, og for oppgavene 23 – 27 gis det 3 poeng for riktige svar.

Denne skåringen forutsetter altså at oppgavene 11-22 er dobbelt så vanskelige som de første, mens oppgavene 23 – 27 er tre ganger så vanskelige. Også dette virker litt komplekst, og alternativer kan være mulige og kanskje interessante å vurdere.

Har kan det nevnes at reliabiliteten på denne skalaen for de internasjonale dataene også her er høy. Ifølge Garolera (2019) er Cronbach's alfa på hele 0.860.

# 3. NOEN RESULTATER OG ANALYSER

Designet for 'KÅL-prosjektet' (Andersen et al., 2019; Hundevadt & Klausen, 2019) er ikke helt enkelt. For det første har vi data fra tre ulike tidspunkter. Det er:

1. Før intervensjonen (med kunst som læringsform)
2. Rett etter intervensjonen
3. Seks måneder etter intervensjonen

For det andre er informantene delt i en intervensjonsgruppe (som fikk erfaring med kunst som læringsform) og en kontrollgruppe (som ikke fikk denne erfaringen). Om vi begrenser oss til informanter som har gitt data til alle de tre tidspunktene, får vi N = 139. Av disse er 83 i intervensjons-gruppa, og 56 i kontrollgruppa.

## 3.1 Skalaenes måletekniske egenskaper

Det er rimelig først å undersøke de målene som er brukt i det norske materialet, med 'klassiske' psykometriske problemstillinger. Hvor konsistente eller *reliable* er målene fra de fire spillene? Og i hvilken grad er de fire målene *valide*, dvs. at de måler det de skal måle? I første omgang ser vi da på de tre datainnsamlingene enkeltvis.

### 3.1.1 Reliabilitet

Vurderingen av reliabilitet er ikke helt enkel, da utregningen av *Cronbach's alfa* ga noen over-raskelser. Med 'listwise deletion' av manglende data forsvinner en stor del av våre data, da relativt mange respondenter ikke har gitt noe svar på en eller flere oppgaver. Ikke minst gjelder dette datasettet fra før intervensjonen, hvor frafallet er relativt stort for både Hund/katt- og Trio-spillet. Disse tallene for *alfa* er derfor ikke representative for utvalget som helhet.

Tabell 1 på neste side viser alfa-verdiene for data fra de fire spillene. Som Pedhazur and Schmelkin (1991) påpeker, finnes det ikke noen allment akseptert standard for hva som er akseptable verdier. Men for våre formål vurderer vi alfa-verdiene som akseptable for Hund/katt og Piler, mens verdiene for Trio er lavere enn ønskelig.

**Tabell 1**: Cronbach's alfa (og N) for data fra fire spill i tre datainnsamlinger

| | Før intervensjonen | Rett etter | Seks mnd. etter |
|---|---|---|---|
| **Hund/katt** (33 oppg.) | 0.87 (N = 18) | 0.80 (N=103) | 0.73 (N = 117) |
| **Trio** (21 oppg.) | 0.49 (N = 50) | 0.49 (N = 83) | 0.36 (N = 94) |
| **Piler** (36 oppg.) | 0.69 (N = 122) | 0.76 (N = 128) | 0.68 (N = 125) |
| **Tallforbindelser** (27 oppg.) | * (N = 0) | * (N = 0) | * (N = 0) |

Men det er enda mer kritisk for *Tallforbindelser*, hvor *ingen* har svart på alle oppgavene ved noen av de tre datainnsamlingene. Da blir det ikke mulig å regne ut *alfa* i det hele tatt. Dette problemet er i stor grad knyttet til de fem siste oppgavene, hvor ingen har gitt riktig svar og mange får notert 'manglende data'.

Det hjelper imidlertid ikke å ta disse fem variablene ut av analysen. Om vi forsøker det, viser det seg at den gjennomsnittlige kovariansen mellom oppgavene er *negativ* ('negative average covariance among items'). Dette bryter for det første med forutsetningene for denne analysen. Dessuten betyr det vel uansett at vårt mål på denne skalaen hverken er konsistent eller pålitelig.

Disse tallene synes vanskelige å forene med de resultatene som er oppgitt av Garolera (2019). Trolig fortjener disse forskjellene derfor noe oppmerksomhet.

Det er imidlertid litt enklere å regne *alfa* for en samlet sumskåre for Yellow Red som *helhet*. Selv om vi mangler data fra enkeltoppgaver i et spill, har jo alle informantene fått en sammenlagt-skåre for spillet. Derfor kan vi bruke denne sammenlagt-skåren som deler av en samlede sumskåre.

Som tabell 2 viser, får vi da en relativt svak *alfa* (mellom 0.62 og 0.63) for alle de tre tidspunktene. Det betyr at den *interne konsistensen* i denne samleskalaen ikke er spesielt god.

Analysen antyder kanskje også at Tallforbindelse-skalaen bidrar mest til dette problemet. På alle de tre tidspunkter blir *alfa* litt lavere dersom Hund/katt-, Trio- eller Pil-skalaen fjernes. Det skjer imidlertid ikke om det er som Tallforbindelser ikke regnes med.

**Tabell 2**: Cronbach's alfa (og N) for en *samlet* sumskåre av fire sumskårer fra tre datainnsamlinger

| | Før intervensjonen | Rett etter | Seks mnd. etter |
|---|---|---|---|
| Alle fire skalaer summert | **.617** | **.633** | **.629** |
| Uten Hund/katt | .524 | .553 | .507 |
| Uten Trio | .529 | .569 | .571 |
| Uten Piler | .457 | .495 | .473 |
| Uten Tallforbindelser | .625 | .617 | .639 |

### 3.1.2 Validitet

*Validiteten* av Yellow Red-data er enda vanskeligere å vurdere med konvensjonell psykometrisk tenkning. I materialet finnes det ikke andre data på eksekutive funksjoner, bortsett fra BRIEF-testen. Denne synes imidlertid ikke å være særlig godt egnet for validering av Yellow Red målene.

Men først må vi merke oss at de elleve indeksene fra BRIEF måler graden av bestemte problemer, slik at høyere tall betyr mere problemer. Yellow Red skalaene måler imidlertid graden av suksess på oppgavene, dvs. at høyere tall betyr flere løste problemer. Dersom *mindre problemer* i BRIEF henger sammen med *flere løste oppgaver* på Yellow Red, skulle dette derfor vises ved *signifikante* og *negative* korrelasjoner.

Som vi ser i tabell 3 på neste side, finner vi dette i noe begrenset grad i data fra før intervensjonen. For det første har skårene fra Hund/katt-skalaen *bare* en sammenheng med Arbeidshukommelse. Den har ingen statistisk signifikant sammenheng med noen av de øvrige BRIEF-indeksene. Det samme gjelder Trio-skalaen. Også denne henger sammen med Arbeidshukommelse, og ikke med noen av de ti øvrige BRIEF-indeksene.

For Pil-skalaen, derimot er det flere signifikante korrelasjoner. Den har sammenheng med både Igangsettelse, Arbeidshukommelse, Orden, Metakognisjon og Selvstyring. Med litt velvilje kan man derfor kanskje tolke dette som en viss validering av Pil-skalaen – at den i det minste måler noe av det samme som noen av BRIEF-indeksene gjør. Men Tallforbindelse-skalaen viser ingen sammenheng i det hele tatt med noen av BRIEF-indeksene. Også her kan det altså se ut til at det er denne skalaen som har minst tilfredsstillende egenskaper.

**Tabell 3:** Korrelasjon mellom 11 BRIEF-indekser og 4 Yellow Red skalaer, 1. runde med data

| | | Hund/katt sumskåre | Trio sumskåre | Piler sumskåre | Tallforbindelser sumskåre |
|---|---|---|---|---|---|
| Impulskontroll | Pearson r | -0,095 | -0,125 | -0,100 | -0,012 |
| | Sig. (2-tailed) | 0,273 | 0,149 | 0,248 | 0,893 |
| | N | 135 | 135 | 135 | 135 |
| Fleksibilitet | Pearson r | -0,037 | -0,112 | -0,118 | -0,089 |
| | Sig. (2-tailed) | 0,667 | 0,195 | 0,172 | 0,304 |
| | N | 135 | 135 | 135 | 135 |
| Emosjonell kontroll | Pearson r | 0,055 | -0,100 | -0,072 | 0,002 |
| | Sig. (2-tailed) | 0,525 | 0,246 | 0,403 | 0,986 |
| | N | 136 | 136 | 136 | 136 |
| Igangsetting | Pearson r | -0,094 | -0,136 | -,265** | -0,128 |
| | Sig. (2-tailed) | 0,275 | 0,114 | 0,002 | 0,136 |
| | N | 136 | 136 | 136 | 136 |
| Arbeids-hukommelse | Pearson r | -,189* | -,225** | -,309** | -0,120 |
| | Sig. (2-tailed) | 0,028 | 0,009 | 0,000 | 0,168 |
| | N | 134 | 134 | 134 | 134 |
| Planlegging /organisering | Pearson r | -0,072 | -0,087 | -0,115 | -0,106 |
| | Sig. (2-tailed) | 0,406 | 0,312 | 0,181 | 0,220 |
| | N | 136 | 136 | 136 | 136 |
| Orden | Pearson r | -0,111 | -0,144 | -,185* | -0,038 |
| | Sig. (2-tailed) | 0,216 | 0,107 | 0,038 | 0,673 |
| | N | 126 | 126 | 126 | 126 |
| Monitorering | Pearson r | -0,059 | -0,131 | -0,119 | -0,056 |
| | Sig. (2-tailed) | 0,489 | 0,125 | 0,166 | 0,517 |
| | N | 138 | 138 | 138 | 138 |
| Atferds-regulerings-indeks | Pearson r | -0,038 | -0,129 | -0,114 | -0,039 |
| | Sig. (2-tailed) | 0,665 | 0,144 | 0,195 | 0,656 |
| | N | 130 | 130 | 130 | 130 |
| Meta-kognisjons-indeks | Pearson r | -0,117 | -0,176 | -,256** | -0,117 |
| | Sig. (2-tailed) | 0,202 | 0,055 | 0,005 | 0,201 |
| | N | 120 | 120 | 120 | 120 |
| Global selv-styringsindeks | Pearson r | -0,070 | -0,160 | -,199* | -0,096 |
| | Sig. (2-tailed) | 0,464 | 0,091 | 0,036 | 0,312 |
| | N | 112 | 112 | 112 | 112 |

Spørsmålet blir da om disse sammenhengene er stabile over tid, eller om de to neste datainnsamlingene gir et annet bilde. Tabell 4 på neste side viser sammenhengene i 2. datarunde.

**Tabell 4:** Korrelasjon mellom 11 BRIEF-indekser og 4 Yellow Red skalaer, 2. runde med data

| | | Hund/katt sumskala | Trio sumskala | Piler sumskala | Tallforbind. sumskala |
|---|---|---|---|---|---|
| Impulskontroll | Pearson r | 0,015 | -0,052 | -0,157 | -0,016 |
| | Sig. (2-tailed) | 0,873 | 0,570 | 0,087 | 0,860 |
| | N | 120 | 120 | 120 | 120 |
| Fleksibilitet | Pearson r | -0,026 | -0,040 | -0,087 | -0,055 |
| | Sig. (2-tailed) | 0,777 | 0,662 | 0,344 | 0,552 |
| | N | 121 | 121 | 121 | 121 |
| Emosjonell Kontroll | Pearson r | 0,075 | 0,047 | -0,043 | -0,020 |
| | Sig. (2-tailed) | 0,408 | 0,609 | 0,638 | 0,826 |
| | N | 123 | 123 | 123 | 123 |
| Igangsetting | Pearson r | -0,122 | -0,159 | -,264** | -0,087 |
| | Sig. (2-tailed) | 0,177 | 0,078 | 0,003 | 0,337 |
| | N | 124 | 124 | 124 | 124 |
| Arbeids-hukommelse | Pearson r | -0,110 | -,179* | -,318** | -0,079 |
| | Sig. (2-tailed) | 0,224 | 0,046 | 0,000 | 0,378 |
| | N | 125 | 125 | 125 | 125 |
| Planlegging/ organisering | Pearson r | -0,124 | -0,045 | -0,161 | -0,006 |
| | Sig. (2-tailed) | 0,177 | 0,627 | 0,078 | 0,946 |
| | N | 121 | 121 | 121 | 121 |
| Orden | Pearson r | -0,001 | 0,067 | -0,116 | -0,033 |
| | Sig. (2-tailed) | 0,988 | 0,464 | 0,208 | 0,719 |
| | N | 120 | 120 | 120 | 120 |
| Monitorering | Pearson r | -0,016 | -0,040 | -,178* | -0,044 |
| | Sig. (2-tailed) | 0,862 | 0,664 | 0,050 | 0,634 |
| | N | 122 | 122 | 122 | 122 |
| Atferds-regulerings-indeks | Pearson r | 0,032 | 0,006 | -0,118 | -0,026 |
| | Sig. (2-tailed) | 0,728 | 0,946 | 0,206 | 0,780 |
| | N | 117 | 117 | 117 | 117 |
| Meta-kognisjons-indeks | Pearson r | -0,090 | -0,082 | -,247** | -0,044 |
| | Sig. (2-tailed) | 0,342 | 0,387 | 0,008 | 0,640 |
| | N | 113 | 113 | 113 | 113 |
| Global selv-styringsindeks | Pearson r | -0,029 | -0,036 | -,196* | -0,039 |
| | Sig. (2-tailed) | 0,765 | 0,709 | 0,041 | 0,690 |
| | N | 109 | 109 | 109 | 109 |

Her synes det meste å være likt tabell 3, og det er nok det viktigste. Et lite unntak ser vi for Hund/katt, som ikke lenger korrelerer med noen av BRIEF-indeksene. Og Piler korrelerer nå med Monitorering, og ikke med Orden. Endringene må likevel vurderes som begrensede.

Men når vi til slutt ser på data fra 3. runde i tabell 5 på neste side, så har det vært mer langt betydelige endringer. Både Hund/katt og Piler er nå signifikant korrelert med de fleste BRIEF-indeksene. Her får altså både Hund/katt- og Piler-skalaene en viss validering fra BRIEF-målene på eksekutive funksjoner.

**Tabell 5:** Korrelasjon mellom 11 BRIEF-indekser og 4 Yellow Red skalaer, 3. runde med data

| | | Hund/katt sumskåre | Trio sumskåre | Piler sumskåre | Tallforbindelser sumskåre |
|---|---|---|---|---|---|
| Impulskontroll | Pearson r | -,208* | -0,065 | -,320** | -0,105 |
| | Sig. (2-tailed) | 0,041 | 0,526 | 0,001 | 0,306 |
| | N | 97 | 97 | 97 | 97 |
| Fleksibilitet | Pearson r | -,208* | -0,112 | -,295** | -0,110 |
| | Sig. (2-tailed) | 0,039 | 0,268 | 0,003 | 0,280 |
| | N | 99 | 99 | 99 | 99 |
| Emosjonell kontroll | Pearson r | -0,143 | -0,059 | -,221* | -0,179 |
| | Sig. (2-tailed) | 0,165 | 0,566 | 0,031 | 0,082 |
| | N | 96 | 96 | 96 | 96 |
| Igangsetting | Pearson r | -,298** | 0,010 | -,271** | -0,099 |
| | Sig. (2-tailed) | 0,003 | 0,925 | 0,007 | 0,333 |
| | N | 97 | 97 | 97 | 97 |
| Arbeids-hukommelse | Pearson r | -,382** | -0,111 | -,334** | -0,140 |
| | Sig. (2-tailed) | 0,000 | 0,272 | 0,001 | 0,166 |
| | N | 99 | 99 | 99 | 99 |
| Planlegging /organisering | Pearson r | -,319** | 0,056 | -,224* | -0,088 |
| | Sig. (2-tailed) | 0,001 | 0,589 | 0,028 | 0,394 |
| | N | 97 | 97 | 97 | 97 |
| Orden | Pearson r | -0,198 | 0,092 | -0,101 | -0,099 |
| | Sig. (2-tailed) | 0,051 | 0,367 | 0,323 | 0,333 |
| | N | 98 | 98 | 98 | 98 |
| Monitorering | Pearson r | -,289** | -0,071 | -,363** | -0,159 |
| | Sig. (2-tailed) | 0,004 | 0,487 | 0,000 | 0,117 |
| | N | 99 | 99 | 99 | 99 |
| Atferds-regulerings-indeks | Pearson r | -,208* | -0,077 | -,323** | -0,155 |
| | Sig. (2-tailed) | 0,046 | 0,463 | 0,002 | 0,138 |
| | N | 93 | 93 | 93 | 93 |
| Meta-kognisjons-indeks | Pearson r | -,318** | 0,020 | -,309** | -0,116 |
| | Sig. (2-tailed) | 0,002 | 0,852 | 0,003 | 0,270 |
| | N | 92 | 92 | 92 | 92 |
| Global selv-styringsindeks | Pearson r | -,282** | -0,016 | -,342** | -0,139 |
| | Sig. (2-tailed) | 0,008 | 0,883 | 0,001 | 0,195 |
| | N | 88 | 88 | 88 | 88 |

### 3.1.3 Foreløpig vurdering

Disse første konvensjonelle analysene viser ikke uten videre at skalaenes måleegenskaper er gode og forsvarlige. Verdiene på Cronbachs *alfa* er noe lavere enn ønskelig, og antyder nok at konsistensen eller sammenhengene mellom de forskjellige oppgave som inngår i hver skala gjerne skulle ha vært sterkere. Sett fra tradisjonell psykometri er det derfor rimelig å konkludere med at *reliabiliteten* kunne ha vært bedre, da disse 'testleddene' ikke måler nøyaktig det samme i tilstrekkelig grad.

Et sannsynlig bidrag til dette problemet synes å være at skåringsprosedyrene gir en stor mengde 'manglende data'. Særlig gjelder dette Tallforbindelser, som vist i resultatkapitlet.

Et annet punkt som kan fortjene oppmerksomhet, er spørsmålet «hva er det som skal måles?». Dette henger som kjent enda tettere sammen med *validiteten*. Men denne er det enda vanskeligere å vurdere, da forundersøkelsen dessverre ikke inneholder andre data enn BRIEF om eksekutive funksjoner hos respondentene.

Som vi så i tabellene 3 og 4, så er det gjennomgående ikke noe tett og godt samsvar mellom data fra Yellow Red og BRIEF. Noen korrelasjoner finnes nok, f.eks. viser *Arbeidshukommelse*-faktoren i BRIEF en viss sammenheng med to/tre av de fire skalaene i Yellow Red; og Pil-skalaen korrelerer signifikant med flere av BRIEF-indeksene. Men i tabellene som helhet er det likevel *manglende* sammenhenger som er det mest påfallende.

Tallene i tabell 5 viser imidlertid noe annet. Her er det tydelige samsvar mellom de fleste BRIEF-indeksene og både Hund/katt- og Piler-skalaene, som antyder at de to metodene delvis måler det samme. Disse to Yellow Red skalaene får derfor en viss validering fra de fleste BRIEF-indeksene. Det gjelder imidlertid *ikke* for Trio- og Tallforbindelser-skalaene, som viser en nesten påfallende uavhengighet i forhold til BRIEF-indeksene og deres informasjon om eksekutive funksjoner.

Vi ser imidlertid også at antallet respondenter (N) klart er lavere ved denne 3. runden med data-innhenting. Det er derfor et spørsmål om dette frafallet kan ha hatt noen betydning for disse sammenhengene. Kan f.eks. mange av de 'frafalne' respondentene ha gitt mindre klare data og mere 'støy' enn det flertallet som gjenstår ved slutten av undersøkelsen?

Det kan ellers finnes både gode og akseptable grunner til at Yellow Red og BRIEF gir ulik informasjon. F.eks. forteller kanskje BRIEF om 'vanlige' eksekutive funksjoner under komplekse dagliglivsforhold, mens det mer testpregede opplegget i Yellow Red i større grad gir informasjon om 'peak performance' under enklere, mer standardiserte betingelser.

## 3.2 Endring over tid

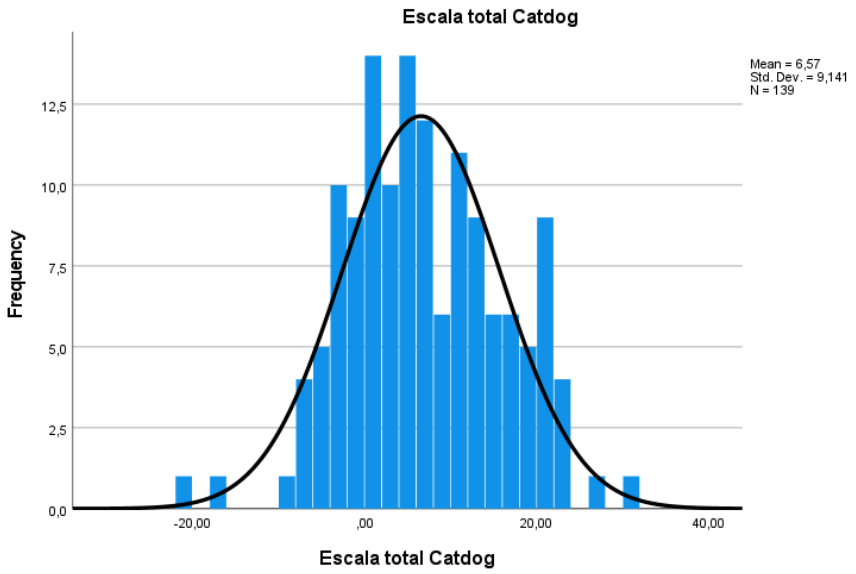
Med data fra tre ulike tidspunkter, blir det selvsagt viktig å sammenligne disse. Er f.eks. disse målene i hovedsak stabile over tid, eller skjer det noen endringer?

For hvert tidspunkt ligger det data for tre versjoner av 'sumskårene' for fire delskalaer (Hund/katt, Trio, Piler og Tallforbindelser). Først ligger de tilsynelatende 'enkle' totalskårene for delskalaene. Så følger z-skårer og T-skårer for disse verdiene. I tillegg til dette er z-skårene for de fire skalaene for hver person summert til samleindeksen «*Suma de Z de pruebas*» (Summen av prøvene). Denne indeksen er dessuten omregnet til T-skåren «Global Index».

Både z-skårer og T-skårer er direkte avhengige av de opprinnelige totalskårene for Cat/Dog, Trios, Arrows og Binding. Det blir da naturlig å begynne arbeidet med disse fire variablene, og deretter se på samleindeksen «*Suma de Z de pruebas*».

### 3.2.1 Hund og katt-skalaen

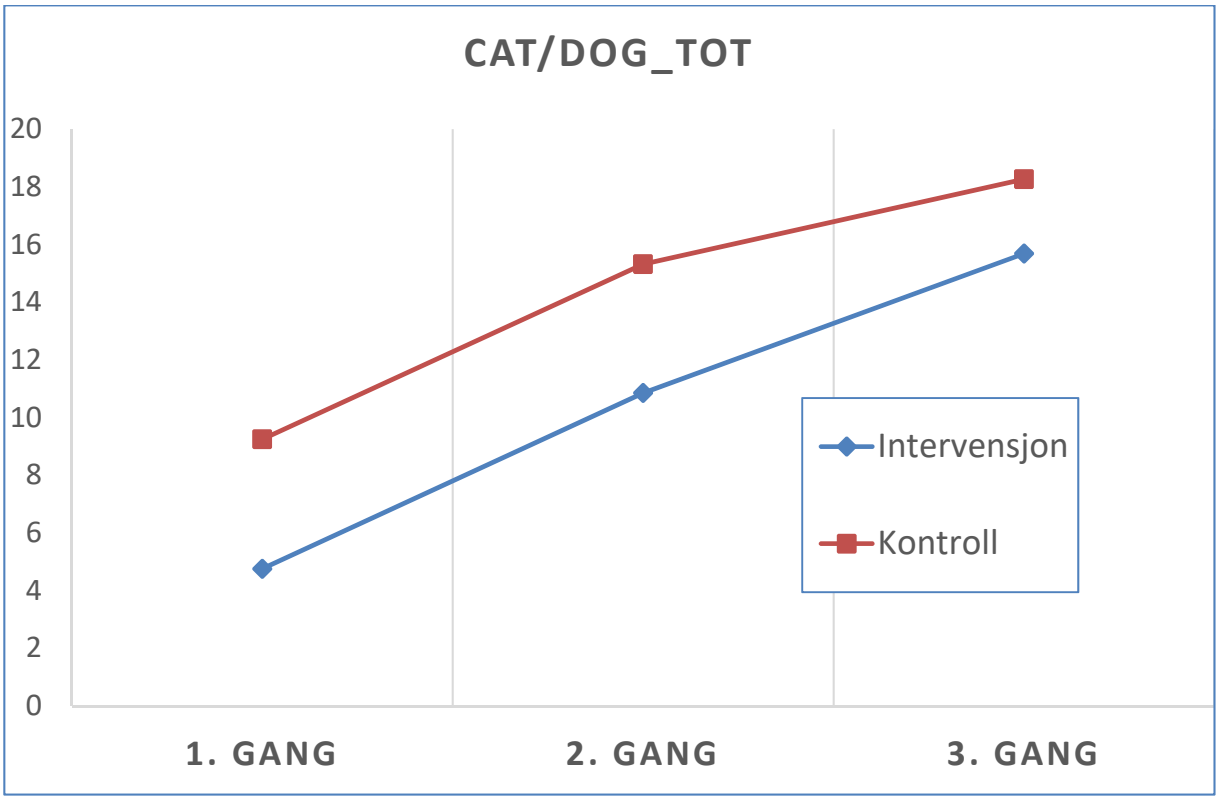
Dette spillet har 33 oppgaver, og sumskårer mellom 0 og 33 kunne derfor ha vært en rimelig forventning. Men som nevnt er dette *ikke* enkle sumskårer, og -2 for feilsvar kan ha stor betydning. Skårene for 1. tidspunkt ligger derfor mellom -21 og +30, og gjennomsnittet er på 6,6. I figur 8 på neste side ser vi at skårene har en forholdsvis klokkeformet fordeling.




**Figur 8:** Fordeling på variabelen *catdog_TOT* på 1. tidspunkt (før intervensjonen).

Slike klokkeformede fordelinger får vi også for 2. og 3. tidspunkt. Der er gjennomsnittene på henholdsvis 12,7 og 16,7. Max- og min-observasjoner er på henholdsvis -16/+32 og -15/+32.

Figur 9 viser gjennomsnittet på de tre tidspunktene for kontrollgruppe og intervensjonsgruppe. Tallene for kontrollgruppa ser gjennomgående høyere ut enn tallene fra intervensjonsgruppa. Dette bekreftes også av ANOVA (MS = 1483,875; F = 8,241; p = 0,005). Kan dette bety at man har vært uheldige med utvalget av grupper? Eller er det andre mulige forklaringer på denne forskjellen?




**Figur 9:** Gjennomsnitt for *Cat/dog_TOT* i to grupper på tre tidspunkter

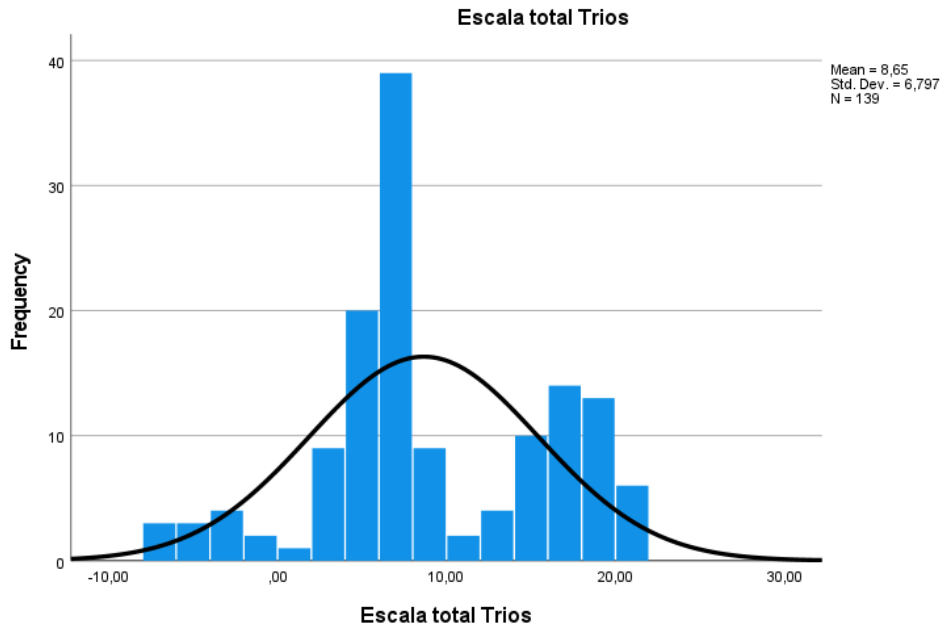
Figuren viser også at tallene gjennomgående øker fra 1. gang via 2. gang til 3. gang. ANOVA bekrefter også dette (MS = 3379,413; F = 75,506; p < 0,001). En nærliggende tolkning av dette er at denne delskalaen har en betydelig læringseffekt. En annen forklaring kan være modning.

For «Art of Learning»-prosjektet ville det trolig ha vært mer tilfredsstillende dersom økningen i intervensjonsgruppa hadde vært betydelig større enn i kontrollgruppa. I så fall ville det ha gitt en interaksjonseffekt mellom faktorene «gjentakelse» og «grupper». Men det ser vi ikke i figuren, og ANOVA viser heller ikke noen slik effekt (MS = 40,132; F = 0,897; p = 0,409).

### 3.2.2 Trio-skalaen

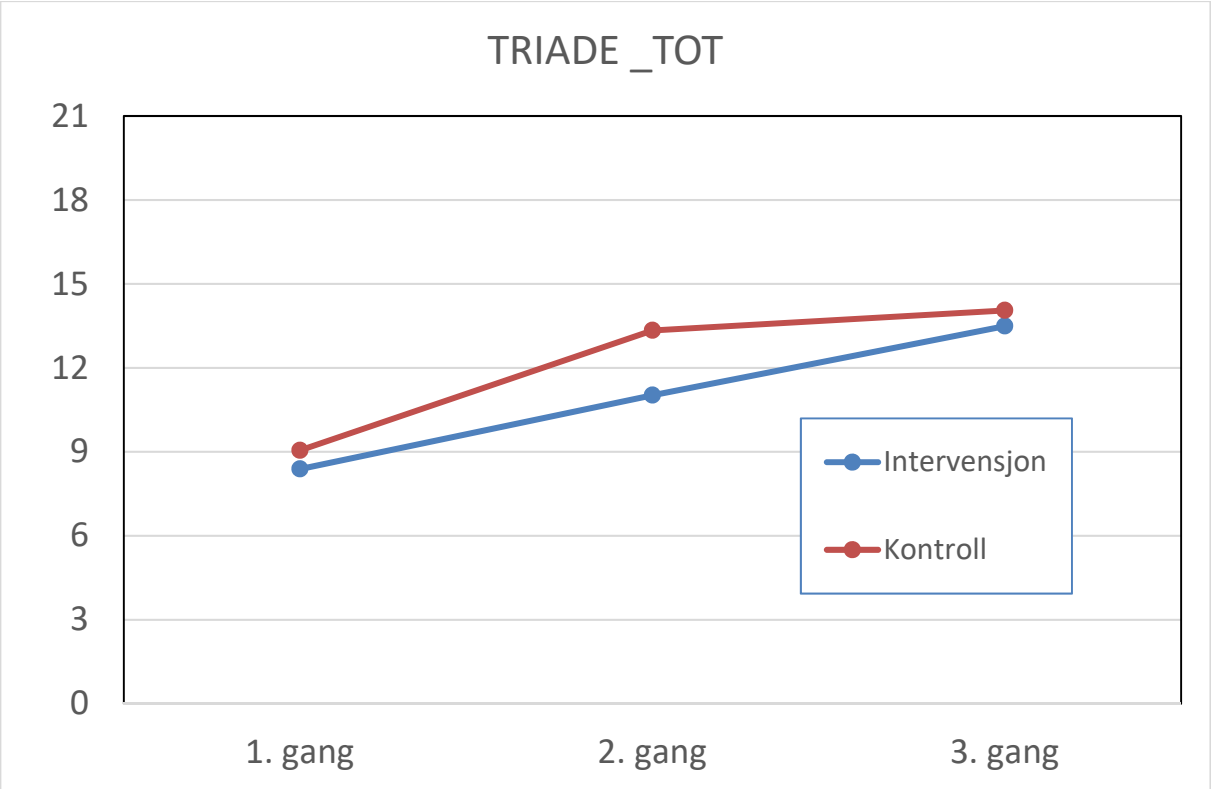
Denne skalaen har 21 oppgaver, og også her er «totalskåren» en sammensatt sumskåre. Skårene for 1. tidspunkt varierer fra -7 til +21, med et gjennomsnitt på 8,7. At max. skåre er 21, kan virke rimelig, og vi ser at de negative skårene for feilsvar spiller en viss rolle også her. Men som figur 10 viser, har vi også her en forholdsvis 'vanlig' fordeling av skårene.

Omtrent tilsvarende fordelinger får vi også her for 2. og 3. tidspunkt. Der er gjennomsnittene på henholdsvis 12,0 og 13,8. Max- og min-observasjoner er henholdsvis -9/+21 og -6/+21. *Heller ikke her er det klart hvordan disse totalskårene blir regnet ut.*




**Figur 10:** Fordeling på variabelen *Trios_TOT* på 1. tidspunkt (før intervensjon).

Og som vi ser i figur 11 på neste side, ligger kontrollgruppa også her litt høyere enn intervensjonsgruppa. Men her viser ANOVA at denne gjennomgående forskjellen er liten og ikke statistisk signifikant (MS = 139,903; F = 1,533; p = 0,218).

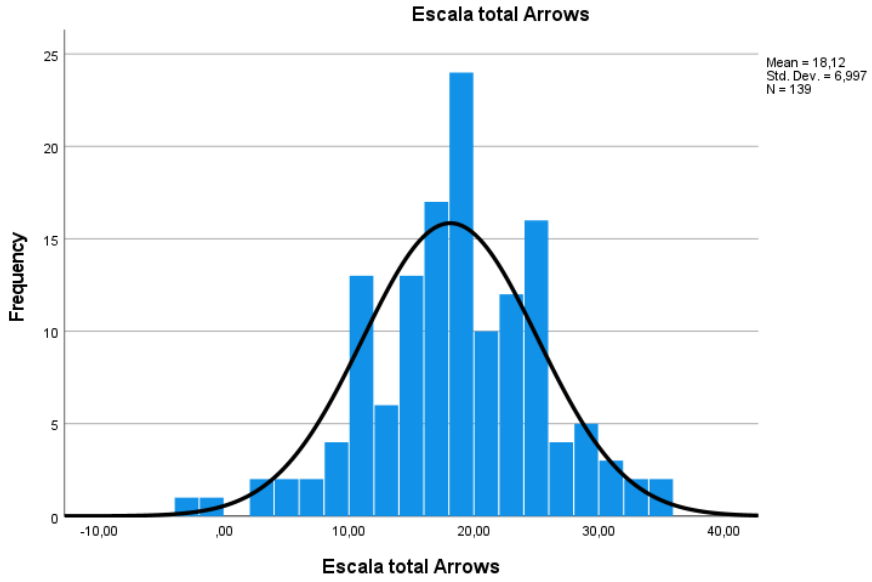



**Figur 11:** Gjennomsnitt for *Triade_TOT* i to grupper på tre tidspunkter

Det skjer derimot klare endringer over de tre tidspunktene også her. Tallene øker både fra 1. til 2. gang, og fra 2. til 3. gangs observasjon. ANOVA bekrefter at denne endringen er signifikant (MS = 893,179; F = 38,993; p < 0,001). Heller ikke her er det noen interaksjonseffekt (MS = 32,363; F = 1,413; p = 0,245).

### 3.2.3 Piler-skalaen

Denne skalaen har 36 oppgaver. Også her innebærer Total-skalaen 'straffepoeng' for feilsvar. Skårene for 1. tidspunkt går derfor fra -3 til +35, og har et gjennomsnitt på 18,1.

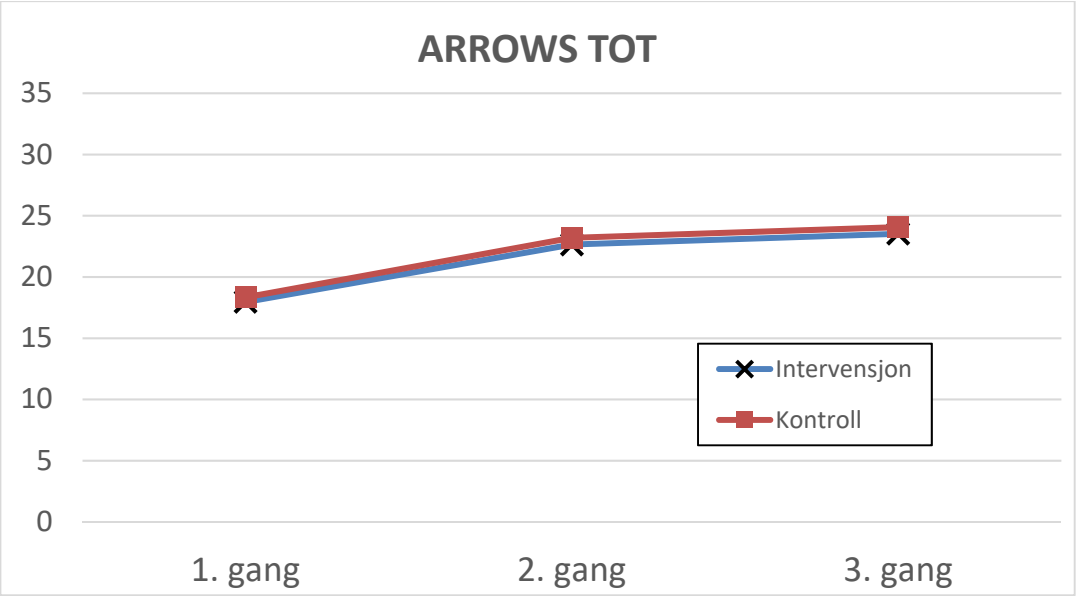



**Figur 12**: Fordeling på variabelen *Arrows_TOT* på 1. tidspunkt (før intervensjonen).

For 2. tidspunkt er gjennomsnittet 22,9. Her er min. verdien 0, og max. verdien +35. Det 3. tidspunktet gir et gjennomsnitt på 23,7, og min. og max. verdier på henholdsvis +5 og +35.

Som vi ser i figur 13 på neste side, er det ingen større forskjeller mellom intervensjons- og kontroll-gruppene på *Arrows_TOT*. ANOVA bekrefter dette (MS = 23,843; F = 0,243; p = 0,623).

Men også her øker gjennomsnittene med gjentakelsene; tallene vokser både fra 1. til 2. gangs og fra 2. til 3. gangs observasjon. Dette bekreftes av ANOVA (MS = 1236,224; F = 53,627; p < 0,001). Vi kan også merke oss at det definitivt ikke finnes noen interaksjonseffekt (MS = 0,315; F = 0,014; p = 0,986).

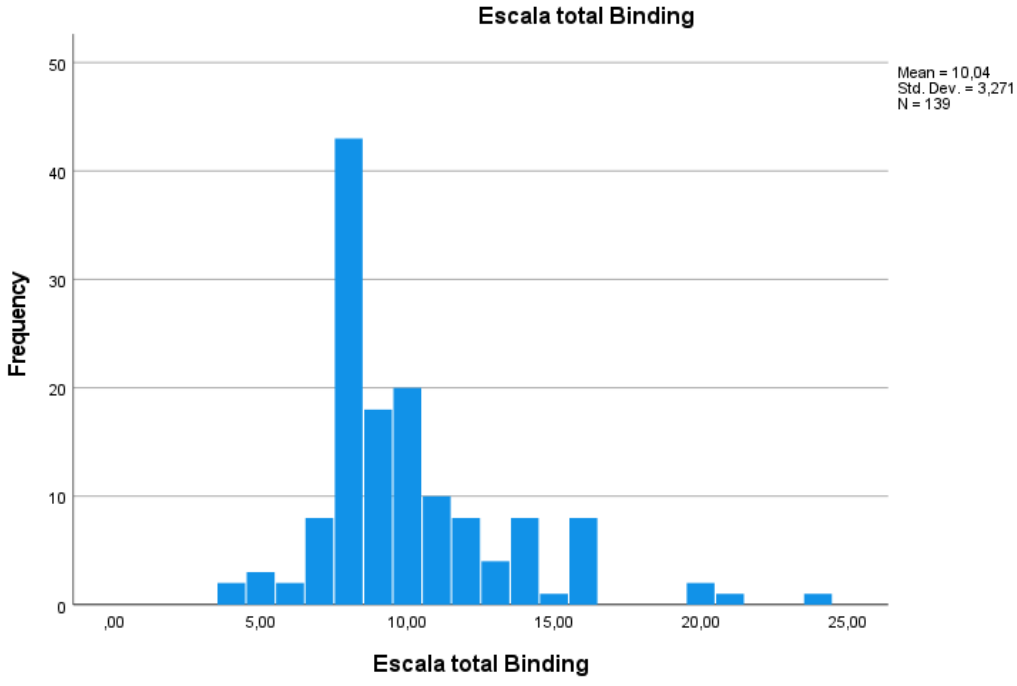



**Figur 13:** Gjennomsnitt for *Arrows_TOT* i to grupper på tre tidspunkter

## 3.2.4 Tallforbindelser-skalaen

Denne skalaen har 27 ledd eller oppgaver. Gjennomsnittsskåren for første tidspunkt er på 10,0. Minimumsskåren er på 4 og maksimum på 24. Siden feilsvar her gir 0 poeng (og ikke -2), så forekommer det ingen negative verdier her. Her skilles det dessuten mellom riktig svar på *enkle* og *vanskelige* oppgaver, og disse gir skårer på henholdsvis +1 og +2. Høye sumskårer er derfor mulige.

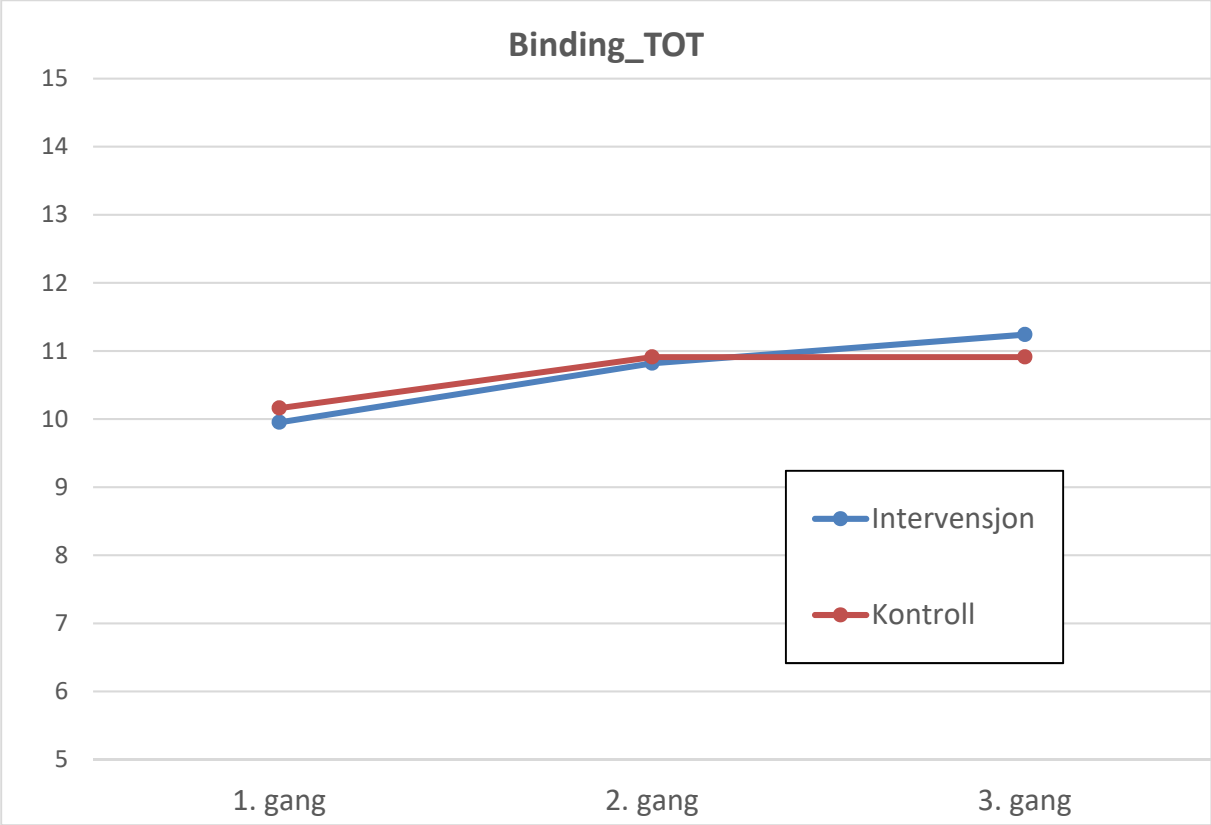
Spillet inneholder også noen få ekstra vanskelige oppgaver, hvor riktig svar ville ha gitt +3 poeng. Slike svar finnes imidlertid ikke i vårt materiale.




**Figur 14**: Fordeling på variabelen *Binding_TOT* på 1. tidspunkt (før intervensjonen).

Her viser variansanalysen for det første at forskjellen mellom de tre tidspunktene er signifikant (MS = 38,746; F = 3,751; p = 0,025). Men det er absolutt ingen forskjell mellom intervensjons- og kontroll-gruppe (MS = 0,010; F = 0,000; p = 0,984), og interaksjonen er også langt fra signifikant (MS = 2,688; F = 0,260; p = 0,771).

Figur 15 viser det samme. Den illustrerer dessuten at forskjellen mellom tidspunktene ikke er særlig stor på denne delskalaen, selv om den altså er statistisk signifikant.




**Figur 15:** Gjennomsnitt for *Binding_TOT* i to grupper på tre tidspunkter

## 3.2.5 Kommentarer til de fire skalaene

Med utgangspunkt i designet for undersøkelsen får vi altså litt ulike resultater fra de fire delskalaene. Tabell 6 på neste side gir en samlet oversikt. Den viser vel et rimelig klart resultat av undersøkelsen; skårene øker *både* fra 1. til 2., og fra 2. til 3. tidspunkt. Den interaksjonseffekten som ville ha vist en virkning av intervensjonen uteblir helt. Når kontrollgruppa skårer tydelig høyere enn intervensjons-gruppa på Cat/Dog-skalaen -- og bare der – er ikke umiddelbart forståelig. Den mest nærliggende forklaringen kan imidlertid være et uforstått metodeproblem.

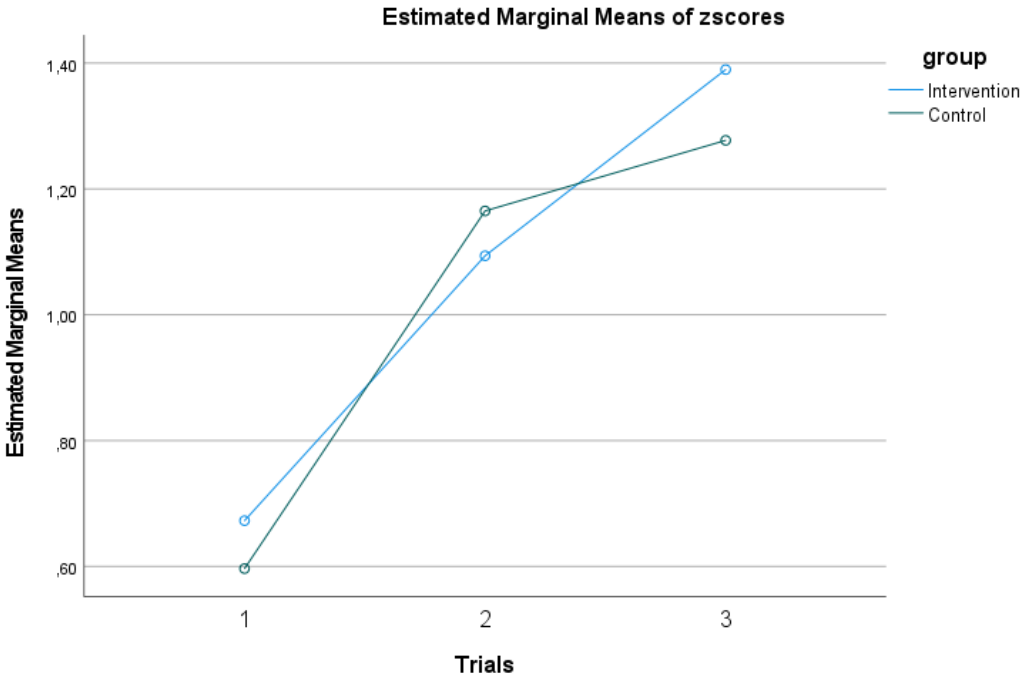
**Tabell 6:** Effekter i ANOVA (trials x grupper), p-verdier

| Delskala | Trials | Grupper | Interaksjon |
|---|---|---|---|
| **Cat/dog** | 0,001 | 0,005 | 0,409 |
| **Trios** | 0,001 | 0,218 | 0,245 |
| **Arrows** | 0,001 | 0,623 | 0,986 |
| **Binding** | 0,025 | 0,984 | 0,771 |

### 3.2.6 Tre-veis ANOVA: Trials x groups x measures

Det kan også være greit å se alle de fire delskalaene i sammenheng. Siden skalaversjonene med z-skårer på sett og vis er sammenlignbare, gir de en mulighet for dette. En tre-veis ANOVA hvor to av faktorene (3 trials og 4 delskalaer) blir forstått som gjentatte målinger, ble derfor forsøkt. Den tredje faktoren er intervensjonsgruppe vs. kontrollgruppe.

Som man vel kan vente, gir dette litt kompliserte resultater. Men for det første er det *ingen* gjennomgående forskjell mellom de to gruppene (MS = 0,409; F = 0,091; p = 0,763). For det andre er det en klar effekt av de tre gjentakelsene (MS = 45,923; F = 64,138; p < 0,001). Og heller ikke her finner vi noen interaksjonseffekt mellom grupper og gjentakelser (MS = 0,843; F = 1,177; p = 0,311). Figur 16 illustrerer disse forholdene.




**Figur 16:** Gjennomsnittlige z-skårer i to grupper på tre tidspunkter, fire delskalaer samlet

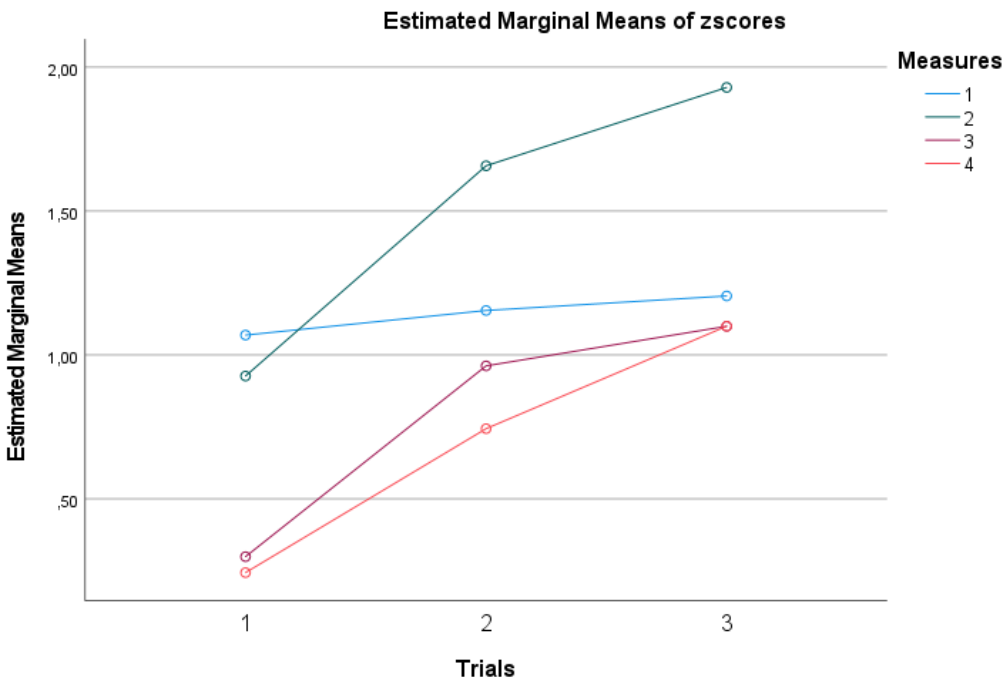
Så langt får vi altså i hovedsak bekreftet det vi fant når delskalaene tidligere ble analysert enkeltvis. Det viktige her er gjentakelsene, og det er fortsatt ingen interaksjonseffekt. Det er heller ingen *generell* forskjell mellom de to gruppene. Selv om vi tidligere fant en slik forskjell på Cat/Dogs-skalaen, gjelder den altså ikke generelt.

Bildet kompliseres likevel litt når vi trekker inn de fire delskalaene som et eget nivå i analysen. Da får vi vite at de utgjør en signifikant hovedeffekt (MS = 36,326; F = 32,935; p < 0,001), og at det dessuten er en interaksjonseffekt mellom delskalaene og tidspunktene (MS = 3,711; F = 6,568; p < 0,001).

Figur 17 viser hvordan dette henger sammen. Her må vi imidlertid merke oss at rekkefølgen på delskalene er omvendt fra tidligere. Det er altså *Tallforbindelser* (blått) som er nr. 1 i figuren, *Piler* nr. 2 (grønt), og *Trio* nr. 3 (fiolett); mens *Hund/katt* er nr. 4 (rødt).

Vi ser da at z-skårene for *Piler* ligger klart høyere enn de andre, mens *Hund/katt* ligger lavest. Det er altså forskjell på de fire delskalaene her. Vi bør derfor kanskje se litt nærmere på hvordan de er blitt regnet ut.

Det er også tydelig at de fire delskalaene har litt ulikt forløp over tid. *Piler* og *Hund/katt* øker forholdsvis jevnt over de tre tidspunktene. Men *Tallforbindelser* ser generelt ut til å øke mindre enn de øvrige, og *Trio*-skårene øker mindre fra 2. til 3. tidspunkt enn fra 1. til 2. Det er disse litt ulike forløpene som utgjør interaksjonseffekten. Hva disse forskjellene mellom skalaene betyr, er imidlertid ikke umiddelbart klart.




**Figur 17:** Gjennomsnittlige z-skårer for fire delskalaer på tre tidspunkter

## 3.3 Endringer i svartyper over tid

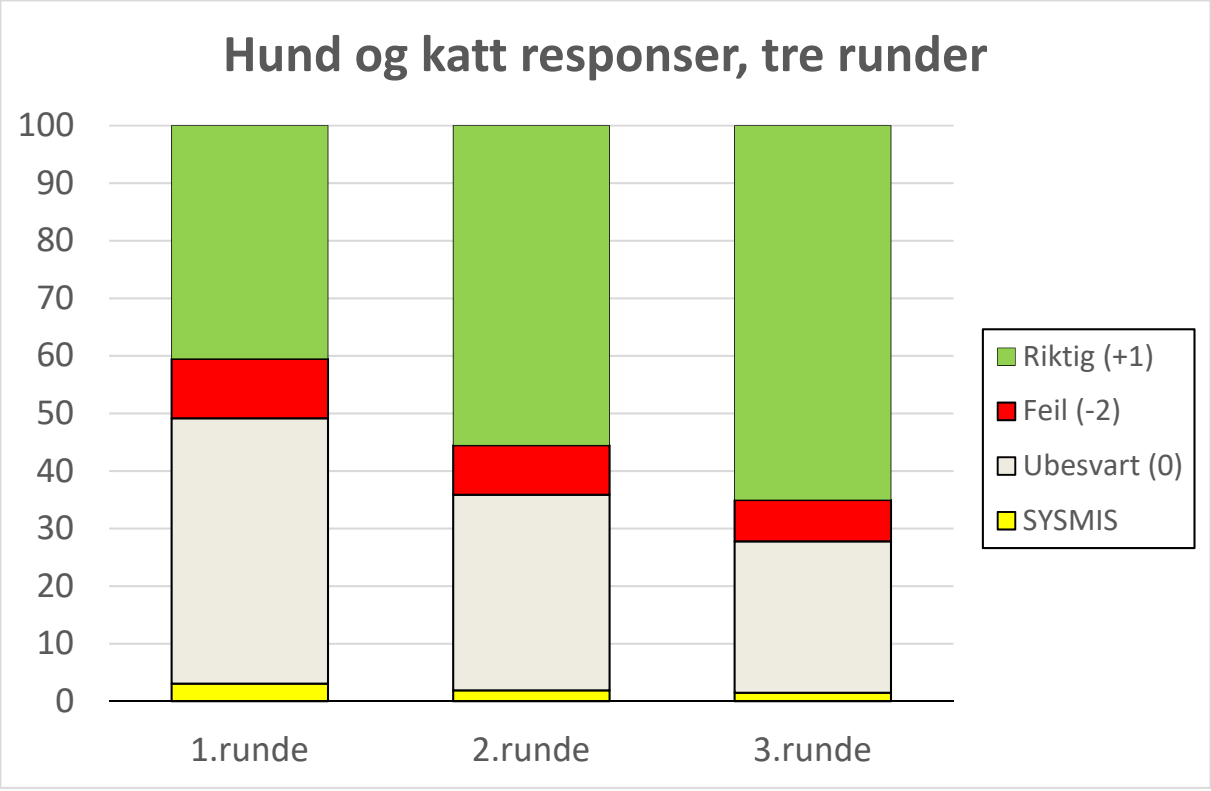
På de tre første spillene gis det som nevnt ulik mengde poeng til forskjellige typer av svar (riktige svar, feil svar, ubesvart og 'manglende data') for hver oppgave (eller testledd). Mens riktige svar får 1 poeng, får feil svar -2. Ubesvart gir 0 poeng, og 'manglende data' betyr at datapunktet ikke regnes med videre.

Når man da summerer alle svarene fra hvert spill for hver respondent, betyr dette at summen inneholder mange og ulike slags informasjon. Kanskje er det særlig viktig å være oppmerksom på at det gis 'straffepoeng' (-2) for feil svar. Et slikt svar vektes altså dobbelt så mye som et riktig svar, og trekker sumskåren mye mere ned enn et ikke-svar. Noe enklere er det for fjerde spillet (*tallforbindelser*), hvor feilsvar bare gir 0 poeng.

Vi ser uansett at den samlede sumskåren er sammensatt av ulike typer informasjon. Dette kan reise spørsmålet om hva er det egentlig som gir de observerte endringene over tid. Blir det ganske enkelt flere riktige svar med økende erfaring, eller er det antallet feil som reduseres? Eller blir barna mer forsiktige etter hvert, og unnlater å svare når de ikke er sikre på hva som er riktig respons til en oppgave? For å undersøke dette, kan vi se på fordelingen av responstyper på de tre tidspunktene.

### 3.3.1 Hund/katt

Figur 18 viser prosentandelene av fire slags responser fra *Hund og katt*-spillet. Vi ser umiddelbart at de største endringene her skjer i andelen *riktige svar* og andelen *ubesvarte*. Andelene *feilsvar* og *manglede data* er relativt små. Og de går nok også litt ned, men i klart mindre grad.

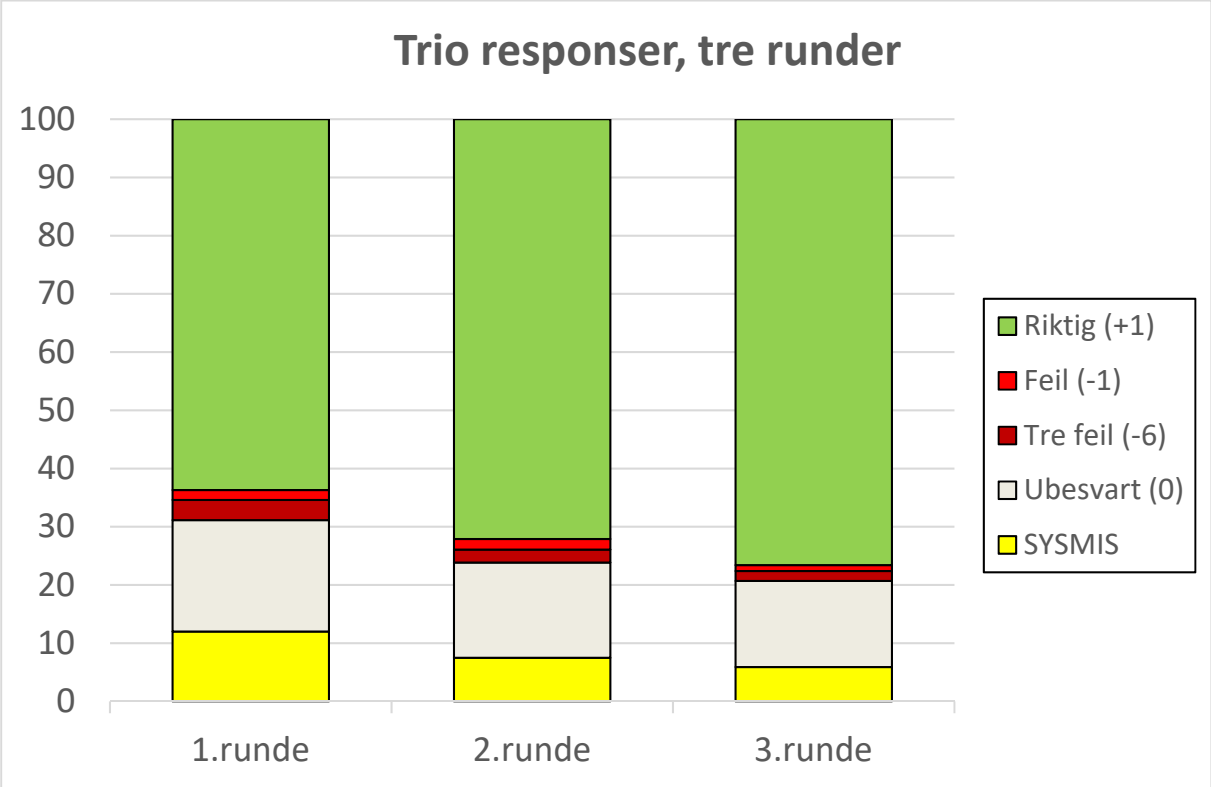



**Figur 18:** Hund/katt: Prosentandel av fire responstyper på tre tidspunkter

Trolig er den sentrale endringen her at det gis flere riktige svar etter hvert. Det kan også være grunn til å merke seg at denne økningen ikke har ført til særlig mange færre feil. Det er snarere andelen ubesvarte som har minket.

### 3.3.2 Trio

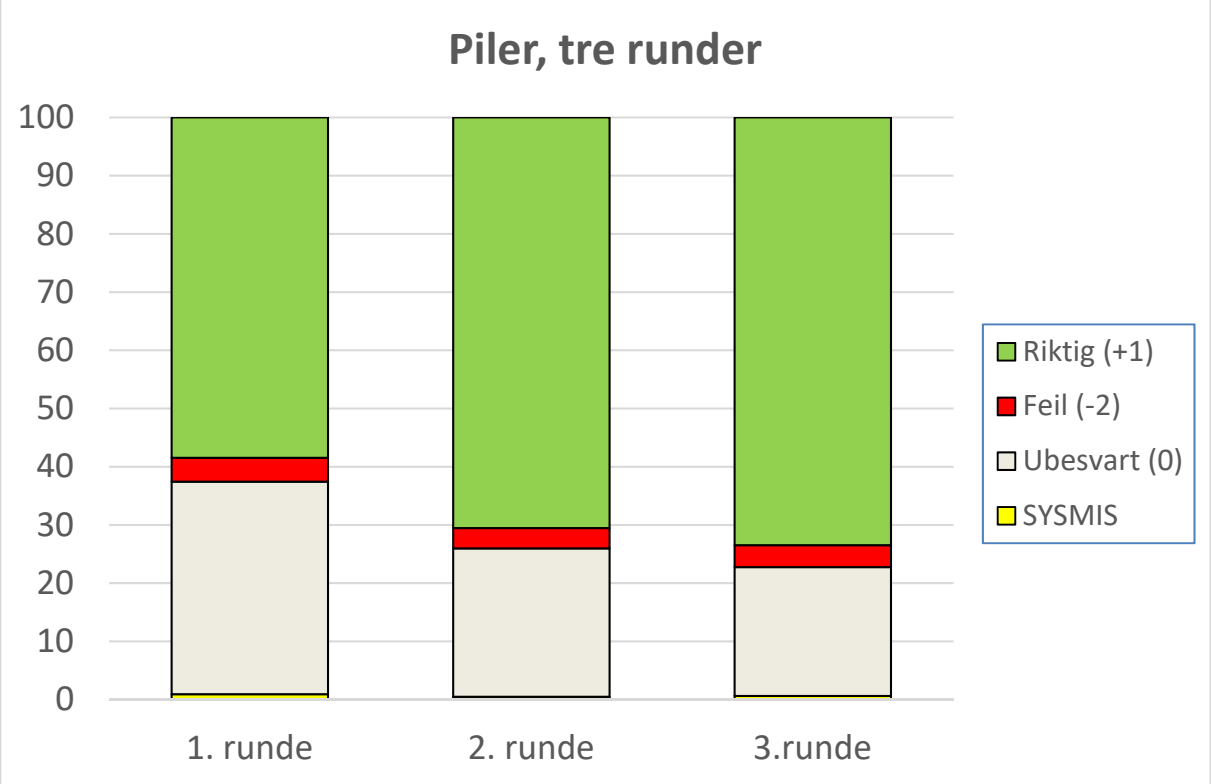
I figur 19 ser vi hvordan de ulike svarene i Trio-spillet endrer seg over tid. Også her er det trolig økningen i andelen *riktige svar* som er det mest påfallende. Det kan dessuten se ut til at denne økningen har gått på bekostning av både andelen *ubesvarte* og andelen *manglende data* (SYSMIS). Men også her viser det seg at andelen *feilsvar* er liten, og minker langt mindre enn økningen i andel riktige svar.




**Figur 19:** Trio: Prosentandel av fem responstyper på tre tidspunkter

### 3.3.3 Piler

Også for Piler-spillet synes økningen i andel riktig svar å være det sentrale, som vi ser i figur 20.




**Figur 20:** Piler: Prosentandel av fire responstyper på tre tidspunkter

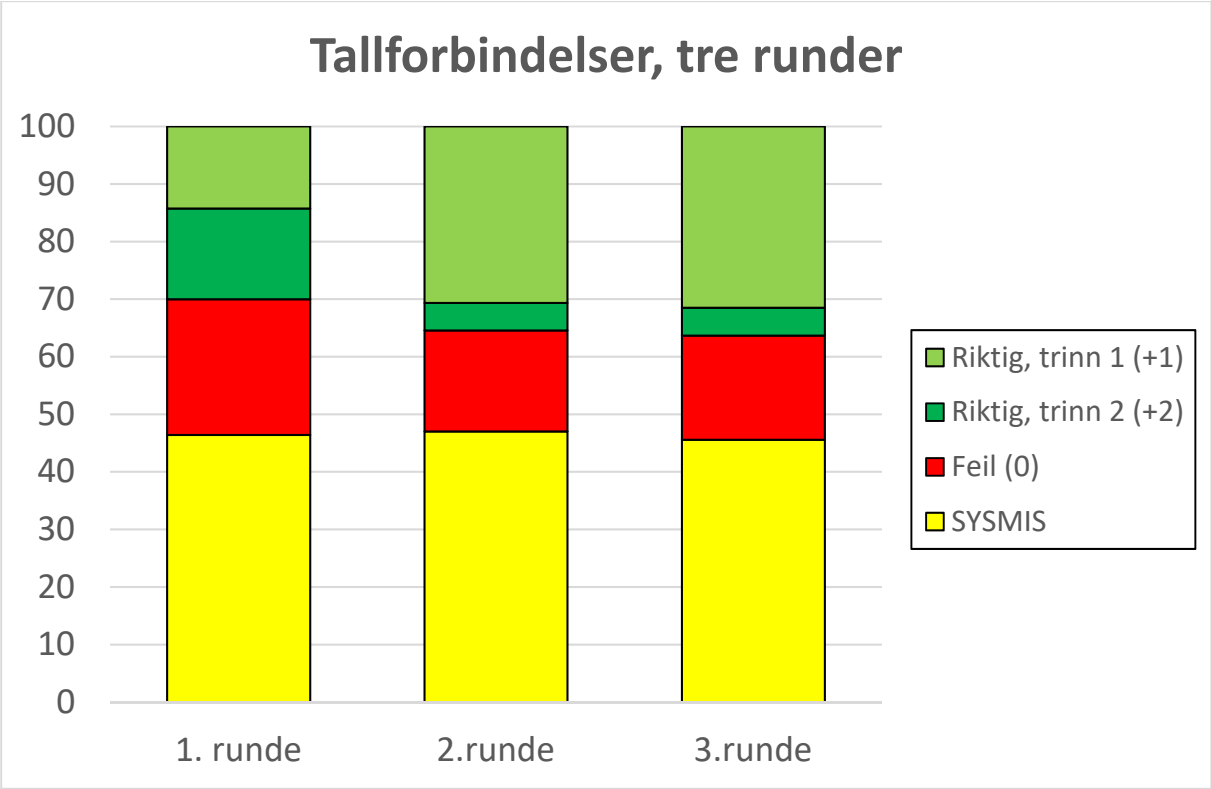
Og også her synes denne økningen å henge sammen med en fallende andel *ubesvarte*. Andelen *feilsvar* er liten, og ikke noen viktig del av det samlede mønsteret. Den gir derfor ingen forklaring på hvorfor andelen riktige har økt.

### 3.3.4 Tallforbindelser

Det mest avvikende med dette svarmønsteret er den kanskje gjennomgående store andelen med *manglende data* (SYSMIS). Den skyldes nok at testprosedyren i dette spillet er noe forskjellig fra de tre første spillene. Men den synes stabil, og endrer seg lite over de tre rundene med datainnsamling.

Det er interessant å se at også i dette fjerde spillet er det en klar økning i andelen riktige svar. Kanskje er det også verd å merke seg at det er på de letteste oppgavene (Trinn 1 i vanskegrad) at økningen kommer, ikke blant de litt vanskeligere oppgavene (Trinn 2).

I dette spillet ser vi dessuten for første gang at økningen i antall riktige svar kan henge sammen med en liten nedgang i andelen feil.




**Figur 21:** Tallforbindelser: Prosentandel av fire responstyper på tre tidspunkter

### 3.3.5 Samlet vurdering av resultatene fra spillene over tre runder

*Samlet sett* synes det imidlertid klart at hovedtendensen er at barna gir flere riktige svar etter hvert, og at dette fører til færre ubesvarte, og ikke til færre feil. Når vi tidligere har sett at gjennomsnittet på de sammensatte sumskårene øker over tid, er det viktig forstå dette i lys av utviklingen av de ulike svarkategoriene. Det er f.eks. *ikke klart* at bedre gjennomsnittlige skårer skyldes færre feil med straffepoeng (-2).

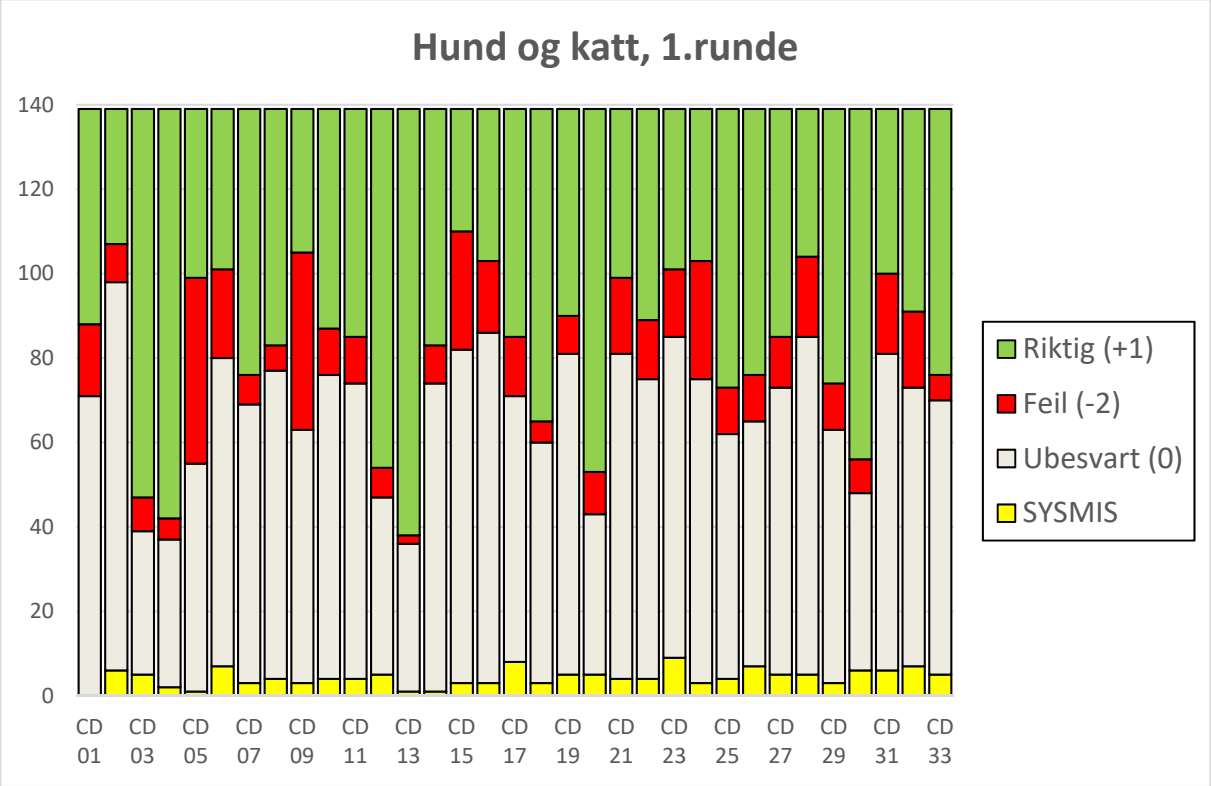
## 3.4 Forskjeller mellom oppgavene i hvert spill

Når skårene fra hver enkelt oppgave legges sammen til en samlet sumskåre, så betyr det at oppgavene gis samme vekt – enten dette gjøres bevisst og eksplisitt eller ikke. Dersom alle oppgavene er omtrent like vanskelige, er dette et relativt uproblematisk valg. Da spiller det liten rolle hvilke oppgaver barnet har klart og ikke, og det er da heller ingen grunn til å holde rede på dette.

Hvis man derimot *ikke* ønsker å forutsette at oppgavene (testleddene) er like vanskelige og kan behandles på samme måte, kan det hele bli mer komplisert.

### 3.4.1 Hund/katt

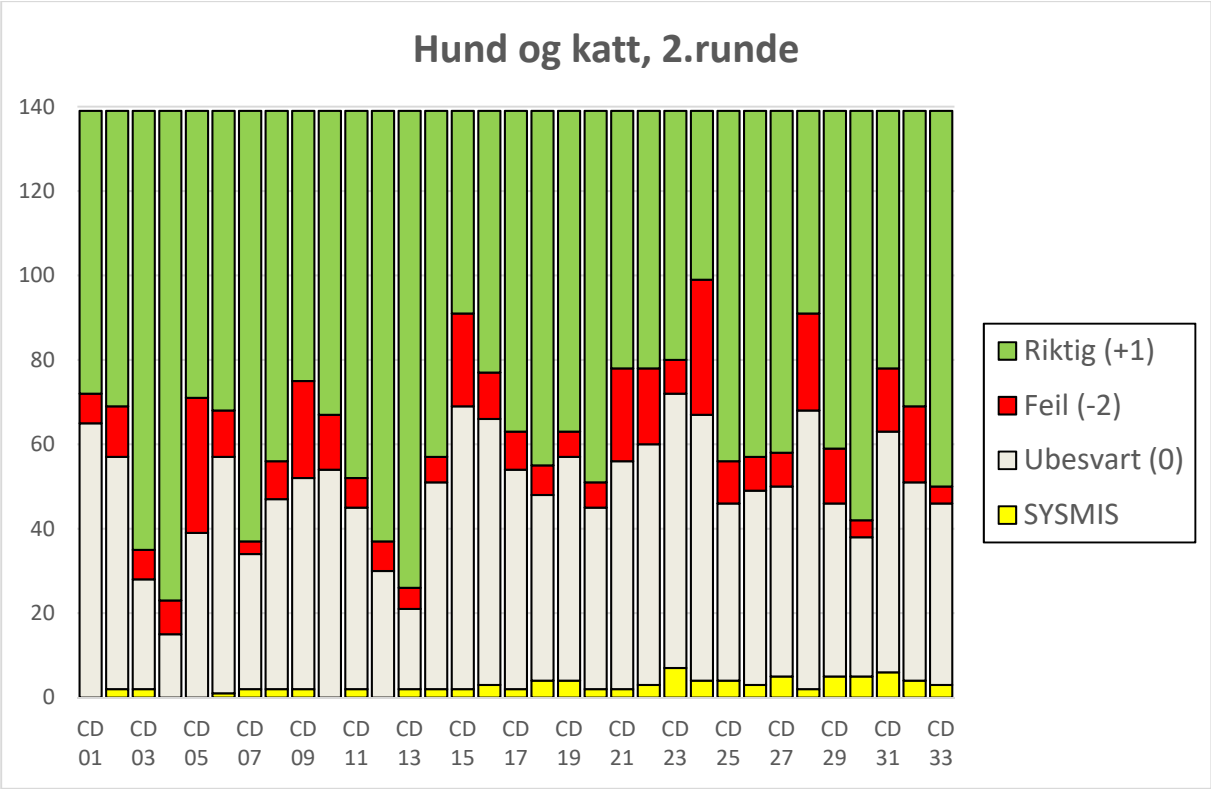
I dette spillet er det relativt begrensede forskjeller mellom de 33 oppgavene, som vi ser i figur 22. På samtlige oppgaver er det ganske mange av de 138 barna som ikke har svart på oppgaven i den første runden med datainnsamling. Men det er også mange som har gitt riktige svar. Andelen feilsvar er tydelig mindre, og det er svært få datapunkter som mangler.



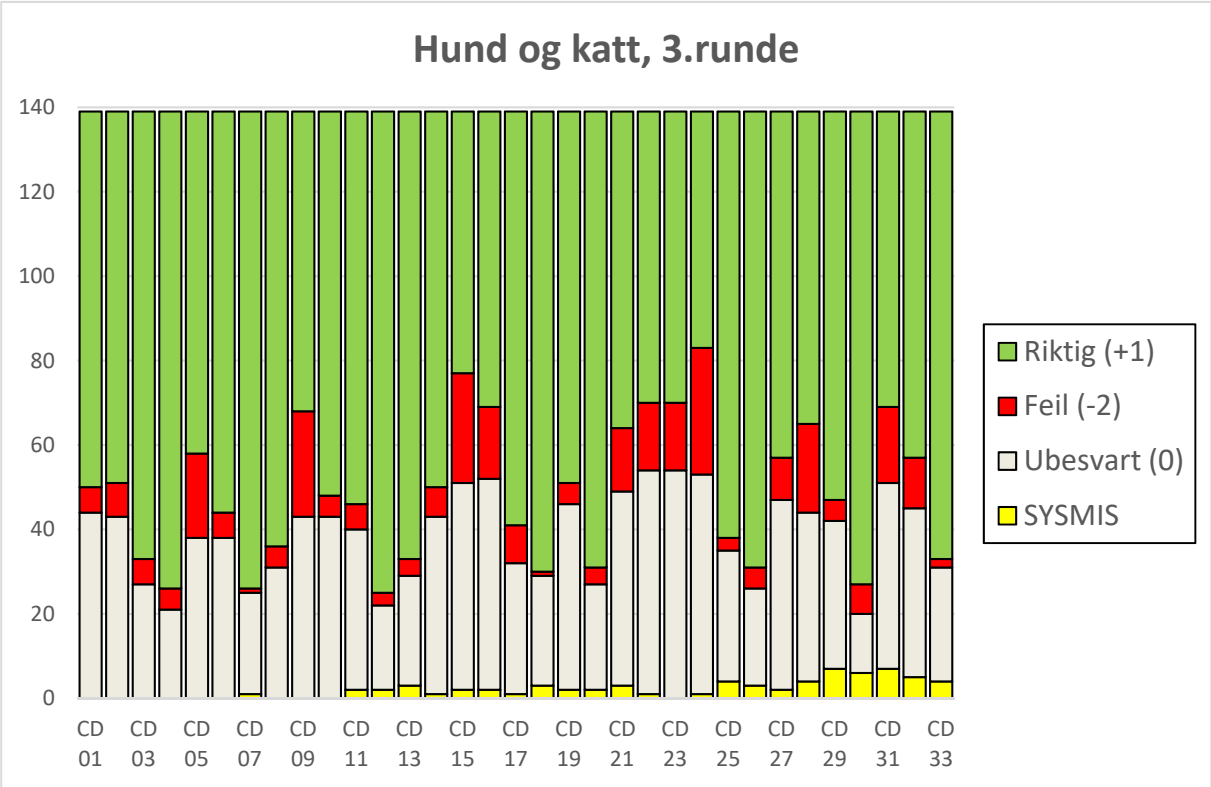

**Figur 22**: Svarmønstre på 33 oppgaver i Hund/katt-spillet, *første* runde med datainnsamling

Og som vi ser i de to neste figurene, er hovedmønsteret i svarene ganske likt i de to neste rundene med datainnsamling. Men vi finner ikke umiddelbart igjen tendensene fra figur 16. Der så vi klarere at mens andelen riktige svar øker over de tre gjentakelsene, så er det andelen ubesvarte som minker. Det blir ikke færre feil.

Men vi ser også at de 33 oppgavene ikke er like vanskelige. F.eks. gir oppgave 1 og 2 færre riktige svar enn oppgavene 3 og 4 på alle de tre gjentakelsene. Og oppgave 4 er konsekvent 'lettere' enn oppgave 3, da den alltid gir flere riktige svar. Slike konsekvente ulikheter er lette å finne i de tre figurene. Dette kan bety at forskjellen mellom oppgavene (eller 'testleddene') kan fortjene noe oppmerksomhet.




**Figur 23**: Svarmønstre på 33 oppgaver i Hund/katt-spillet, *andre* runde med datainnsamling



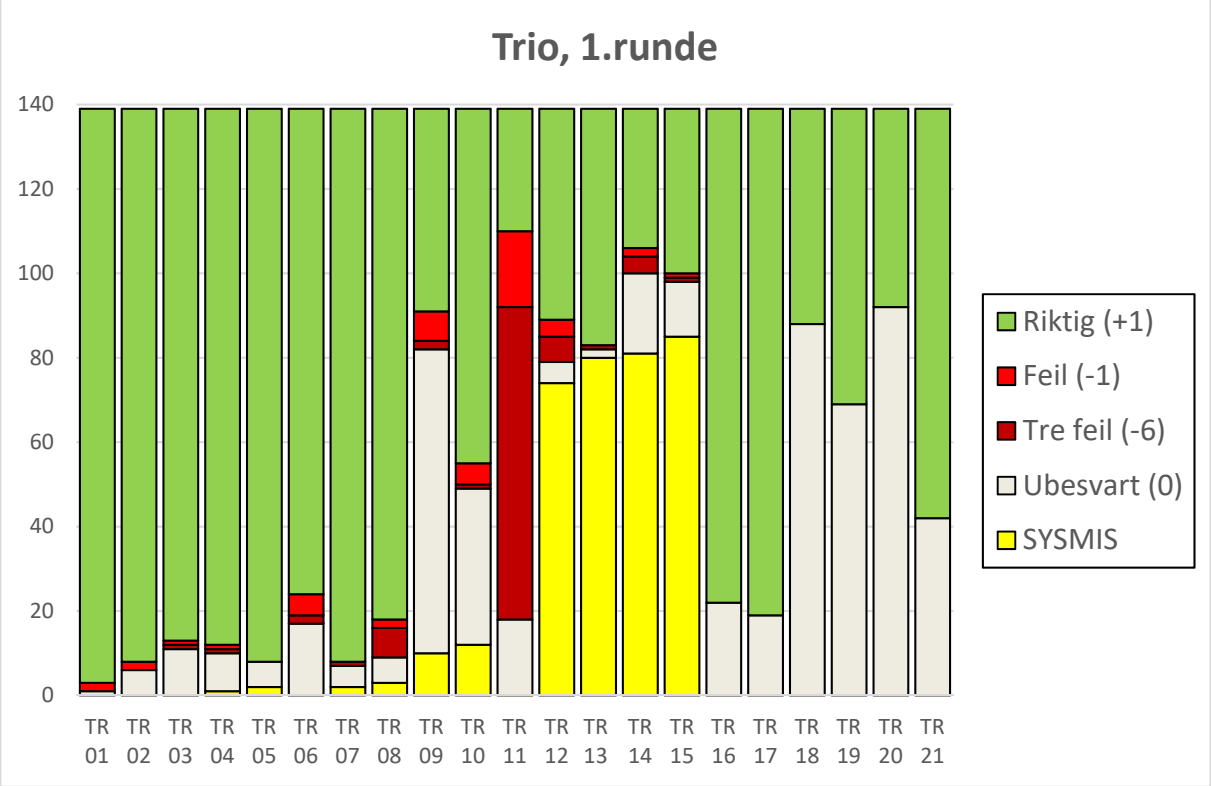

**Figur 24**: Svarmønstre på 33 oppgaver i Hund/katt-spillet, *tredje* runde med datainnsamling

### 3.4.2 Trio

Også dette spillet viser gjennomgående og tydelige forskjeller mellom oppgavene, som vi ser i figurene 25, 26 og 27. Noen oppgaver er lette, og gir mange riktige svar. Andre er vanskeligere, og skiller seg ut med mange ubesvarte eller ved at svar mangler (SYSMIS). Men vi ser også her at 'enkle' feil (-1) er sjeldne.

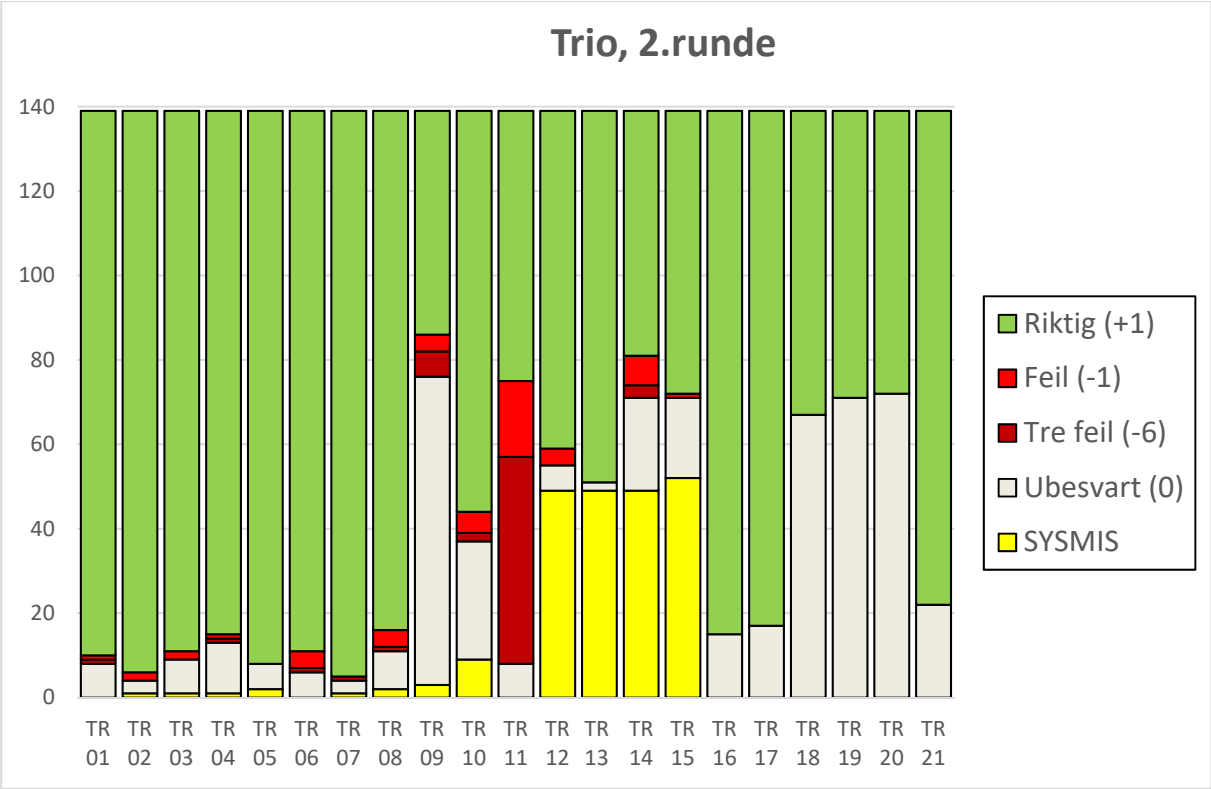
Ikke minst skiller oppgave 11 seg klart fra de øvrige, ved å gi en svært store andel responser med tre feil (-6). Dessuten er oppgavene 12 – 15 avvikende med svært mange manglende svar (SYSMIS). Disse forskjellene ser vi gjennom alle de tre rundene med datainnhenting. Men vi ser *samtidig* at andelen riktige svar øker på dette spørsmålet.

Det er også påfallende at oppgavene 16 – 21 gir *enten* riktige svar eller ubesvart. At oppgavene 1 – 8 er tydelig lettere (har stor andel riktige svar) enn de andre, viser også at forskjellene mellom de 21 oppgavene er betydelige.

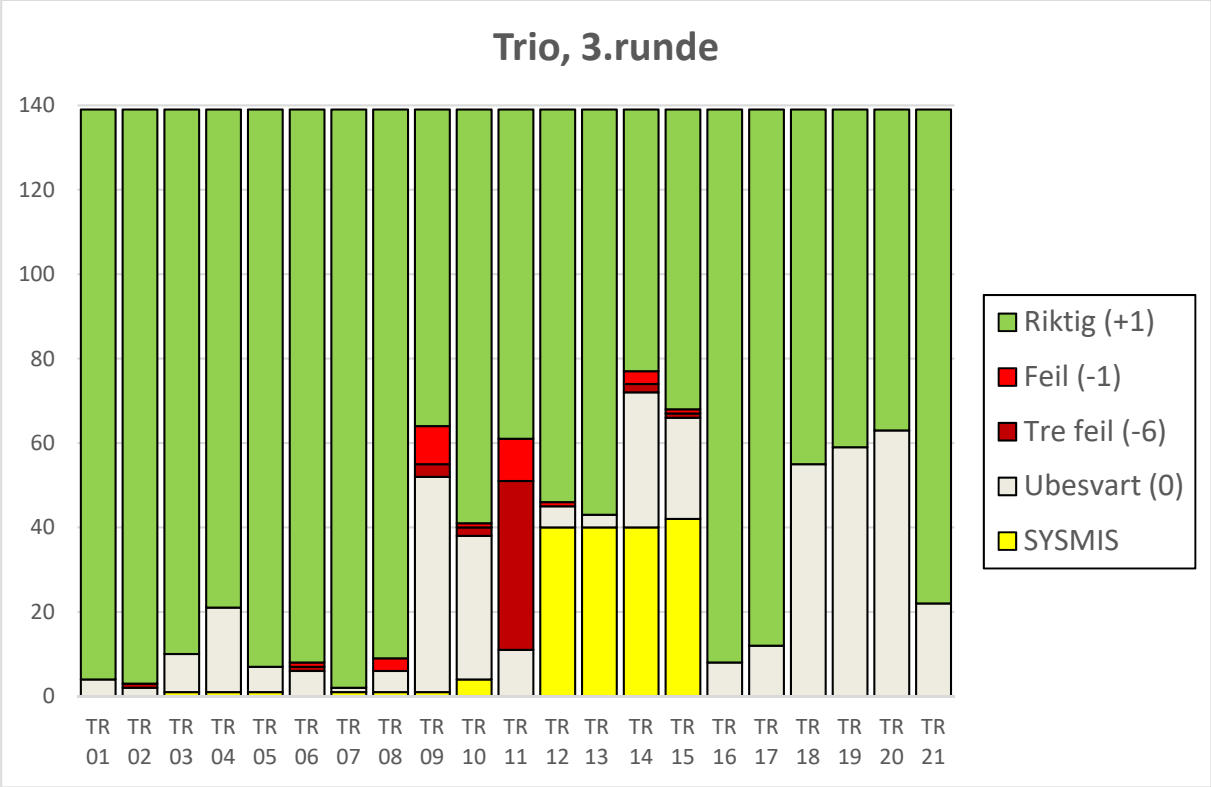



**Figur 25**: Svarmønstre på 21 oppgaver i Trio-spillet, *første* runde med datainnsamling

Også disse tre svarfordelingene bør sammenholdes mot utviklingen i gjennomsnittsforskjeller, slik den ble vist i figur 19. Det er nok både klart og korrekt at andelen korrekte svar øker over de tre gjentakelsene. Men det også tydelig at gjennomsnittsskårene forenkler og tilslører, siden forskjellene mellom oppgavene er betydelige. Med så store ulikheter mellom 'leddene' er det stor risiko for at den uforklarte eller uforståtte variansen i materialet bør undersøkes nærmere. Her er det mulig at viktige forhold ikke blir ivaretatt gjennom sumskårer og gjennomsnittstall.




**Figur 26**: Svarmønstre på 21 oppgaver i Trio-spillet, *andre* runde med datainnsamling

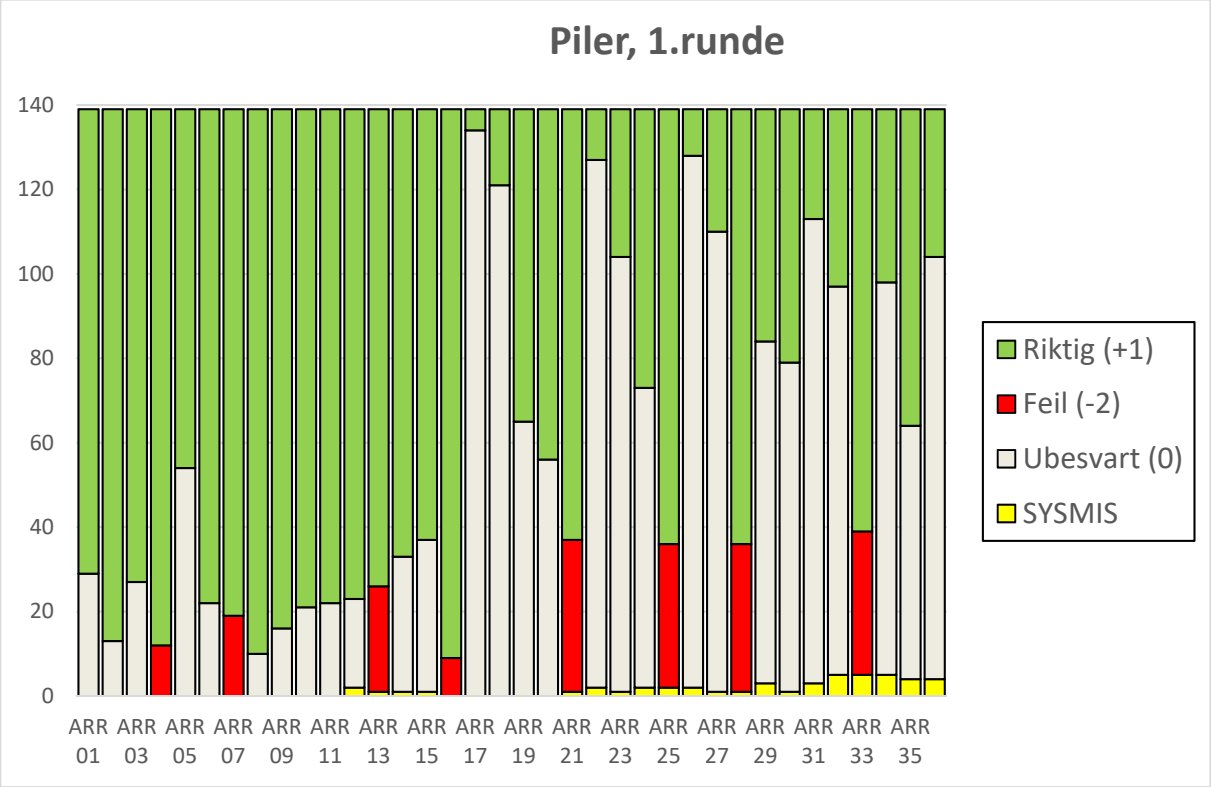



**Figur 27**: Svarmønstre på 21 oppgaver i Trio-spillet, *tredje* runde med datainnsamling

### 3.4.3 Piler

Når vi kommer til spillet med piler, er svarmønsteret igjen forholdsvis likt det vi så i hund/katt-spillet. Som figur 28 viser, så har de 36 oppgavene ikke samme vanskegrad. F. eks ser de første 16 oppgavene ut til å være lettere enn de påfølgende, mens det blir relativt mange ubesvarte oppgaver i den siste halvdelen av spillet. Direkte feilsvar er ikke bare ganske sjeldne, men de fordeler seg også på et svært begrenset antall oppgaver.

Vi ser dessuten et trappetrinn-mønster i den siste halvdelen av spillet. F.eks. har oppgave 17 svært få riktige svar, men så øker andelen gradvis i oppgavene 18 til og med 21. Et liknende mønster ser vi i oppgavene 22 – 25 og 26 – 28.

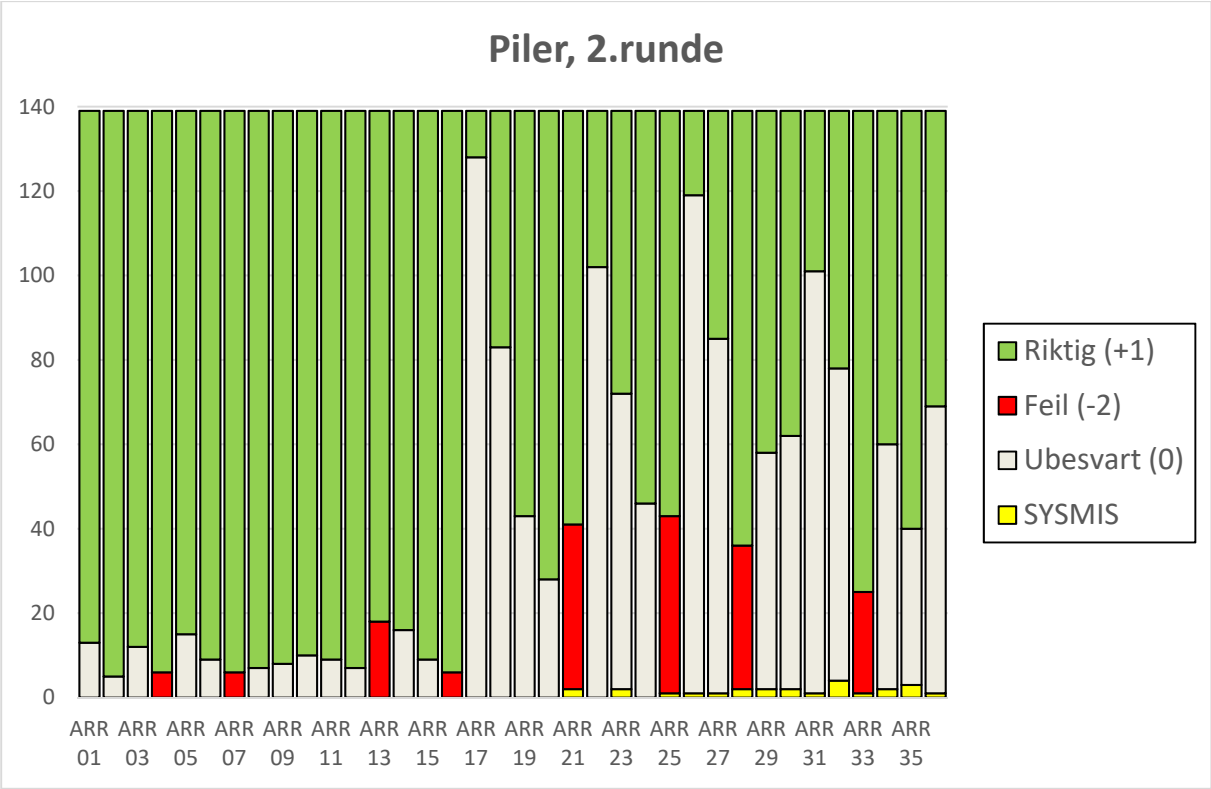



**Figur 28**: Svarmønstre på 36 oppgaver i Piler-spillet, *første* runde med datainnsamling

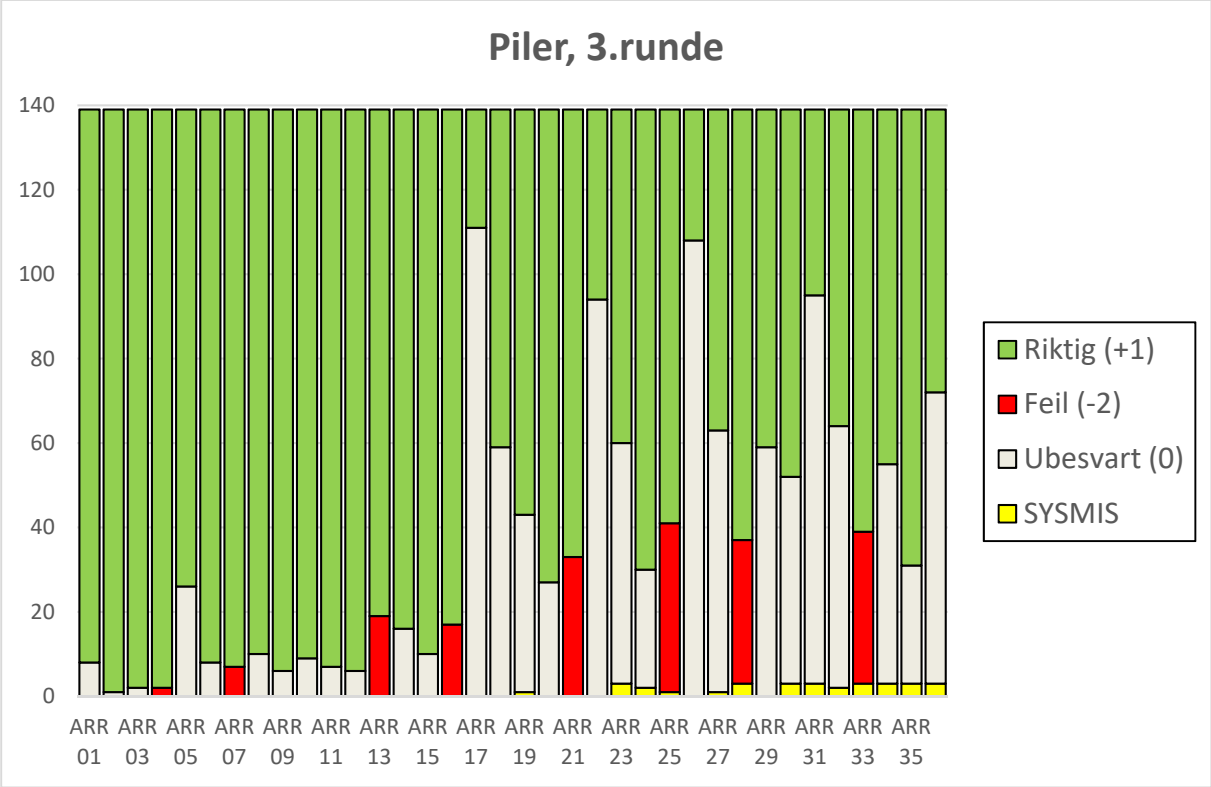
Igjen ser vi at disse mønstrene gjentar seg ved de to neste datainnsamlingene. De første oppgavene har flest riktige svar, og 'ubesvart' er også en vanlig respons. Feilsvar er ganske sjeldne, og det trappetrinn-liknende mønsteret er lett å finne igjen i figurene 29 og 30.

Samtidig er det mulig å se at andelen riktige svar øker litt fra første til andre trinn i datainnhentingen, og så videre fra andre til tredje – i klart samsvar med endringene som ble vist i figur 20.

De store forskjellene mellom leddene reiser altså minst like store spørsmål om uforstått varians som i hund/katt og trio-spillene. Hva går forskjellene mellom leddene ut på, og hvilken rolle har de spilt for svarfordelingene?




**Figur 29**: Svarmønstre på 36 oppgaver i Piler-spillet, *andre* runde med datainnsamling

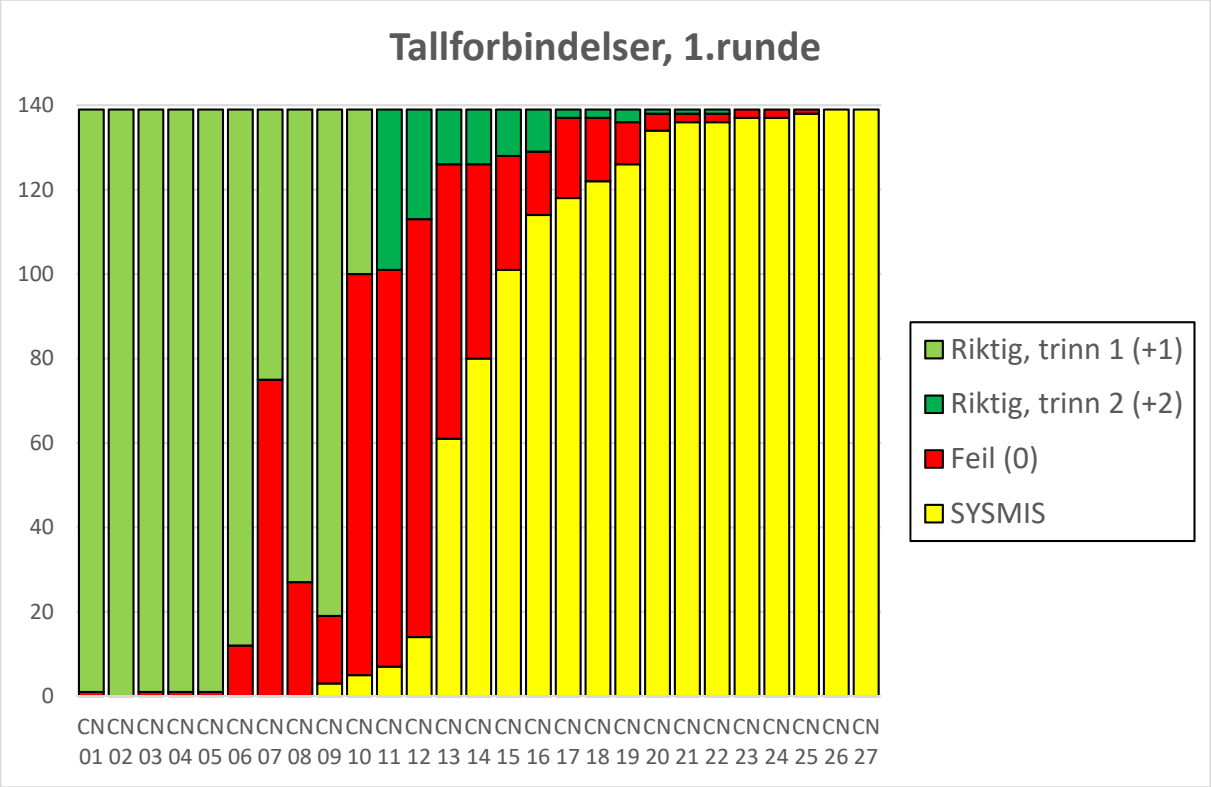



**Figur 30**: Svarmønstre på 36 oppgaver i Piler-spillet, *tredje* runde med datainnsamling

### 3.4.4 Tallforbindelser

Svarmønsteret fra dette spillet skiller seg særlig ut med at svært mange responser mangler (SYSMIS). Det har også tydelig flere feilsvar enn det vi har sett tidligere. Vi bør også merke oss at det her opereres med to grader av 'riktighet', som gir ulik uttelling i poeng. Det skyldes at testen her eksplisitt bruker oppgaver med tre ulike vanskegrader, og at ingen har klart noen av de aller vanskeligste.

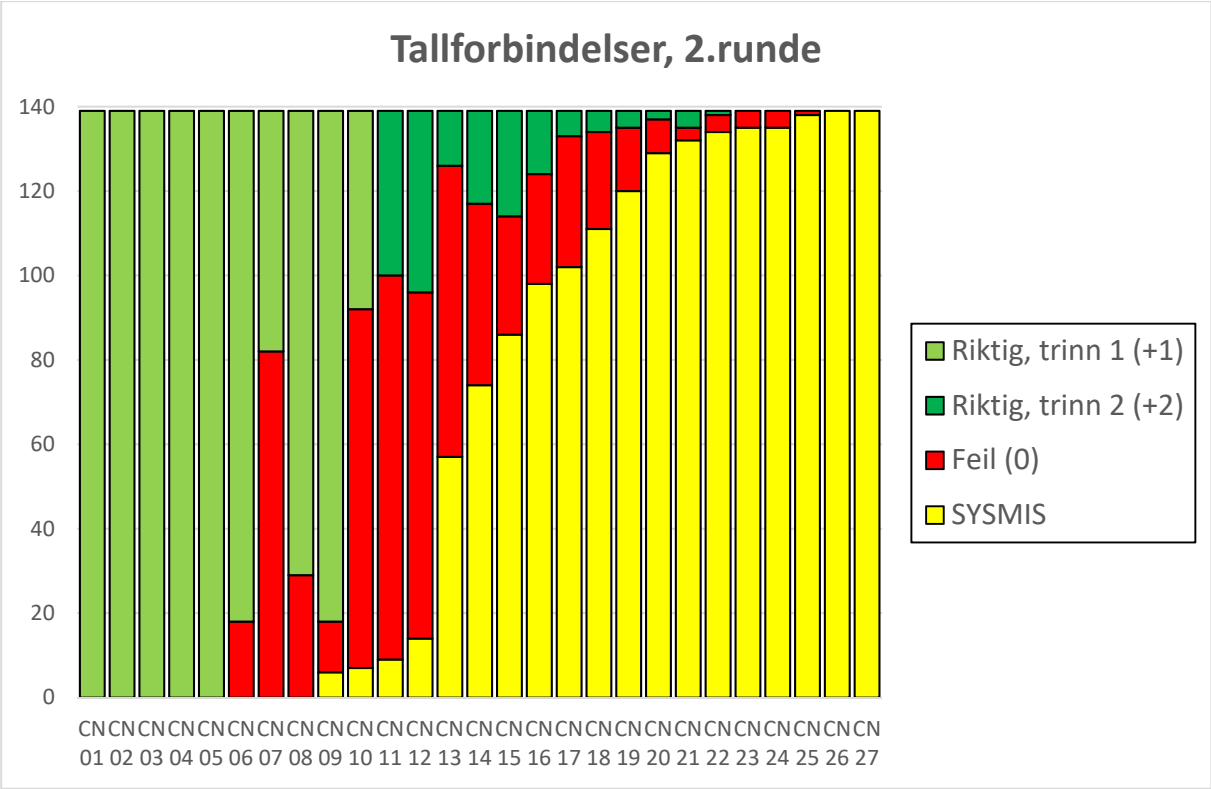
Men figur 31 viser uansett at det er svært store forskjeller mellom leddene. Fordelingen av responstyper varierer mye mellom oppgavene. Vanskegraden her går over hele skalaene fra lett (nesten alle svarer riktig) til vanskelig (ingen svarer riktig).



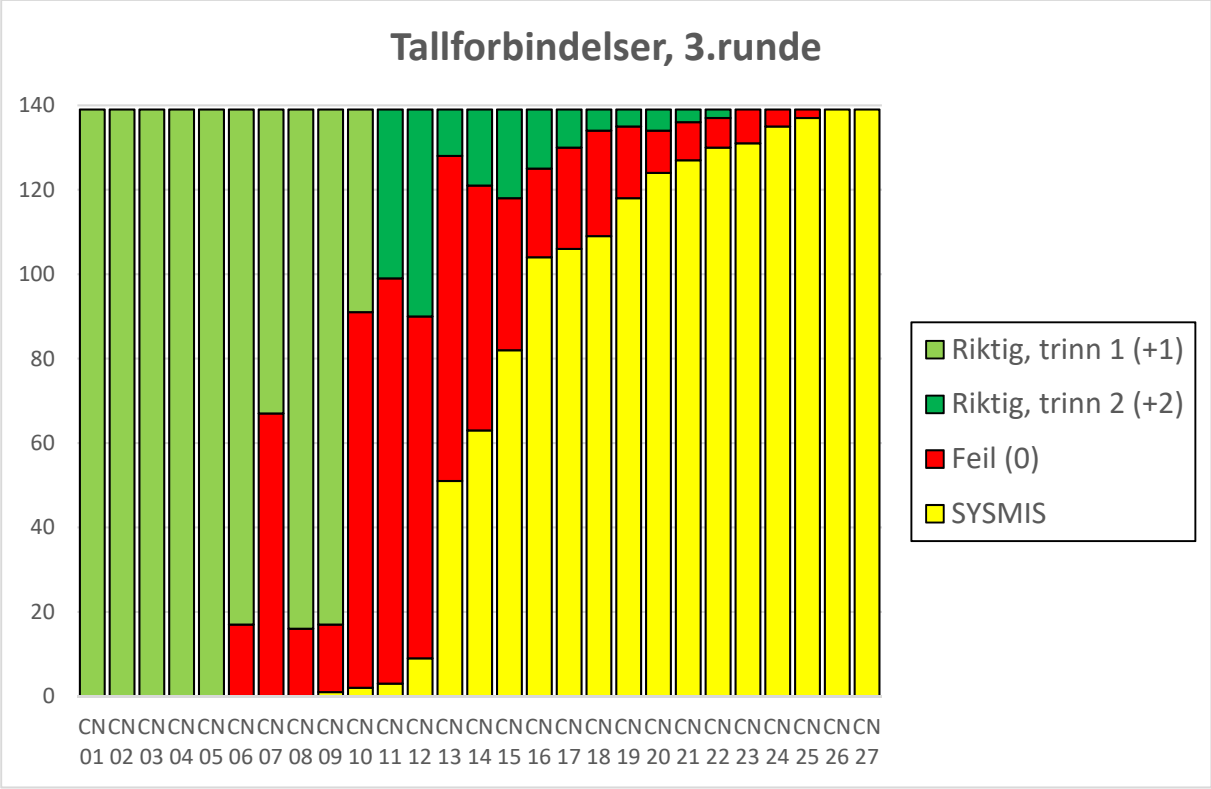

**Figur 31**: Svarmønstre på 27 oppgaver i Tallforbindelser-spillet, *første* runde med datainnsamling

Også her er det relevant å sammenligne mot de sammenlagte tallene i figur 21. Også i figurene 31 – 33 er det mulig å se at andelen korrekte svar øker endel, både fra første til andre og fra andre til tredje datainnsamling. Kanskje er det også mulig å oppfatte at andelen feilsvar går noe ned.

Det mest påfallende er at forskjellen mellom de 27 oppgavene er betydelig. Derfor blir mye informasjon borte også her, når vi helt enkelt slår sammen skårene fra alle oppgaver til et samlende tall. Spørsmålet om 'undertrykket' varians blir minst like sterkt her som i de tre foregående spillene; hva kan forskjellene mellom oppgavene fortelle oss? Kunne vi fått vite noe mer om utvikling av eksekutive funksjoner hos barn dersom vi hadde brukt informasjonen som ligger i disse forskjellene?




**Figur 32**: Svarmønstre på 27 oppgaver i Piler-spillet, *andre* runde med datainnsamling




**Figur 33**: Svarmønstre på 27 oppgaver i Piler-spillet, *tredje* runde med datainnsamling

# 4. ET MULIG ALTERNATIV

Siden det er knyttet noen spørsmål til målene fra de fire skalaene i Yellow Red, kan det være nyttig å vurdere om det kan finnes relevante alternativer for noen av prosedyrene der. Det er ønskelig å bedre både reliabilitet og validitet på data fra de fire spillene, selv om man ikke umiddelbart ser hvordan dette kunne være mulig. Men det er neppe ønskelig å røre ved selve gangen i de fire spillene eller rekkefølgen på oppgavene. Mye av dette er 'innbakt' i automatiserte prosedyrer, slik at evt. endringer vil være både kostnads- og tidkrevende. Det er dessuten samlet inn svært mye data med det eksisterende opplegget, og det er viktig å sikre at gamle og nye data kan sammenlignes. Også dette hensynet tilsier at endringer i spillene vil være lite hensiktsmessig.

Men det kan være noe enklere å se nærmere på hvordan responsene blir *kodet*. Her vil det trolig være litt lettere å prøve ut alternativer. En evt. omlegging av algoritmer og programvare for selve kodingen kan nok gjøres uten å endre prosedyrene for spillene og datainnhentingen. Vi har derfor sett på en mulighet for å kode eller klassifisere rådata fra responsene i spillene. Den innebærer en mye enklere koding, hvor man bare skiller mellom de riktige svarene og alle andre responser.

Andre omkodingsmuligheter finnes selvsagt, som f.eks. Rasch-skalering. Den tar utgangspunkt i at oppgavene i et spill ikke er like vanskelige, og kan derfor gir høyere skåre for løsning av vanskeligere oppgaver. Det innebærer imidlertid en betydelig mer komplisert koding, og vil ikke bli drøftet videre i denne rapporten.

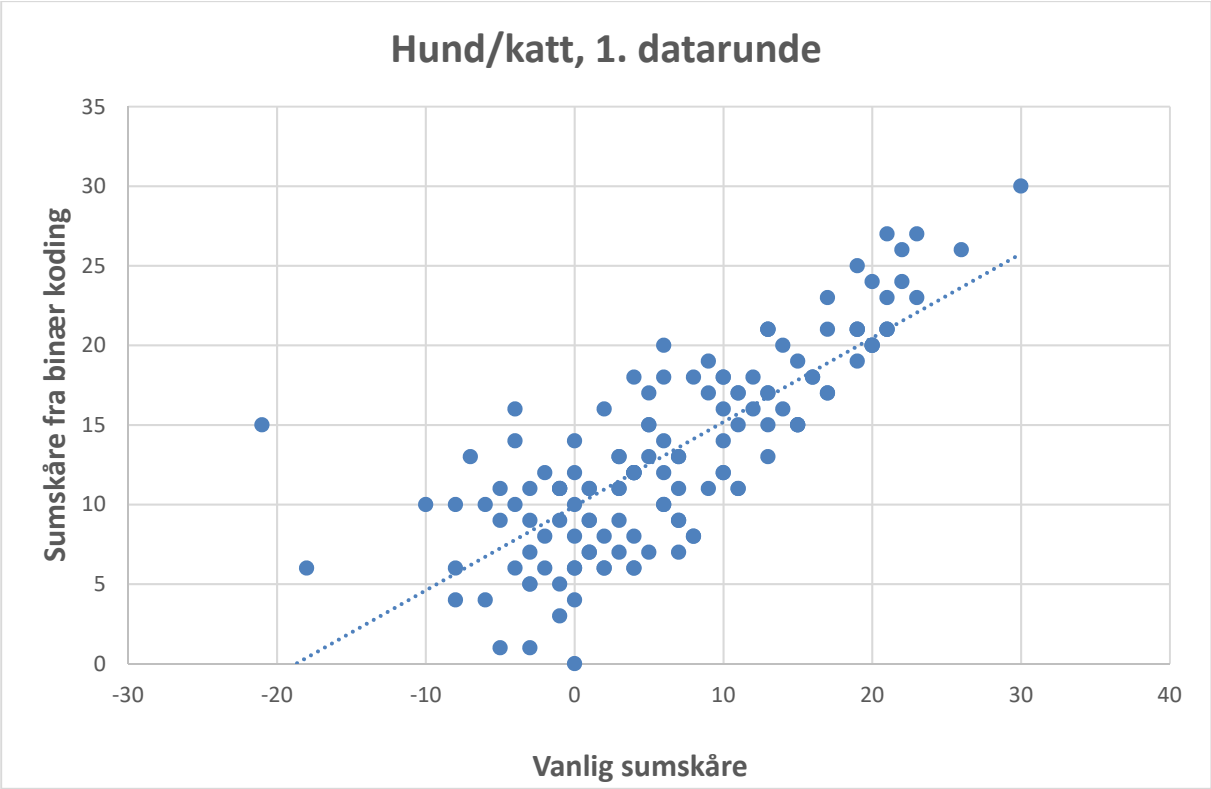
Men omkoding til en binær skåre innebærer som nevnt bare å skille mellom de riktige svarene og alle de andre. Om man da gir 1 poeng for riktige responser og 0 poeng for alle andre (inkludert manglende svar), så kan det tenkes å gi noen fordeler.

Den viktigste fordelen er at man *ikke* vil miste informasjonen fra alle som har fått notert 'manglende svar' på en oppgave. Dette problemet gjelder særlig for Tallforbindelser-spillet, som vi så i resultatkapitlet.

Men det er også ulemper knyttet til en slik enkel omkoding. Den viktigste er selvsagt at man mister informasjon som kan være viktig. F.eks. blir forskjellen mellom (ett eller flere) feilsvar og 'ikke-svar' borte, sammen med skillet mellom 'vanlige' og 'vanskelige' oppgaver i spillet om tallforbindelser. Det prinsipielle spørsmålet blir da om den variansen som skyldes disse forskjellene i hovedsak bidrar til å måle de eksekutive funksjonene vi er ute etter, eller om den bidrar mest til en generell 'støy' eller feilvarians – som tilslører det vi ønsker å måle.

Kanskje er det mulig å danne seg et bilde av disse spørsmålene ved å gjøre den nevnte omkodingen, og så se på forholdet mellom disse nye variablene og de 'normale' variablene fra hvert spill. Vi kan da se på korrelasjonen mellom de to målene først. De varierer imidlertid en god del: For *Hund/katt* er korrelasjonen på 0.79, for *Trio* 0.96, for *Piler* 0.93, og for *Tallforbindelser* 0.55. For *Trio* og *Piler* er det altså et svært høyt samsvar mellom de to kodingsformene, og det synes klart at de to variablene i all hovedsak måler det samme. Men for *Hund/katt* og *Tallforbindelser* er det mer sannsynlig at samsvaret er så svakt at det er interessant – og derfor noe vi bør forsøke å forstå.

Scatterplot av disse fire relasjonene kan kanskje gi noen idéer om dette. I figur 34 ser vi forholdet mellom sumskårer basert på henholdsvis 'normal' koding og binær (løst/ikke løst) koding av oppgavene i Hund/katt-spillet i første runde med datainnsamling.

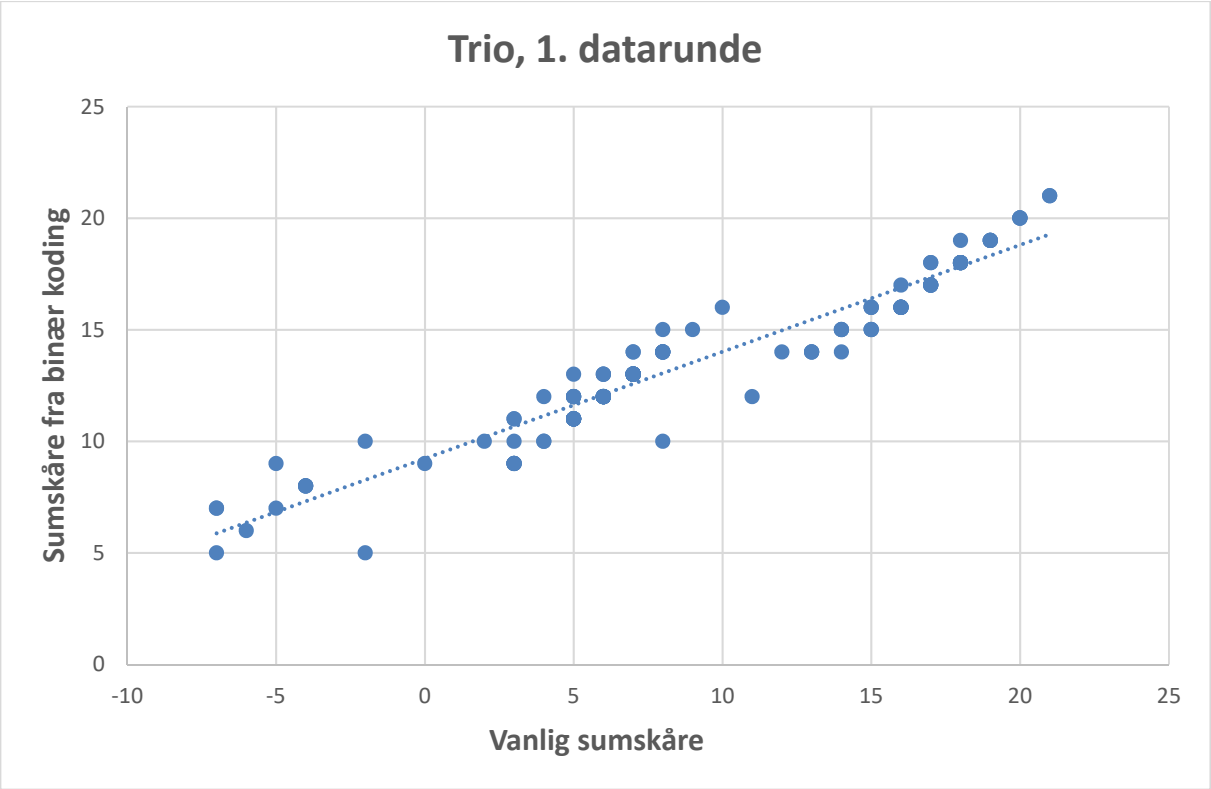



**Figur 34:** Sumskårer fra 'Normale' og binære sumskårer, Hund/katt-spillet i 1. datarunde

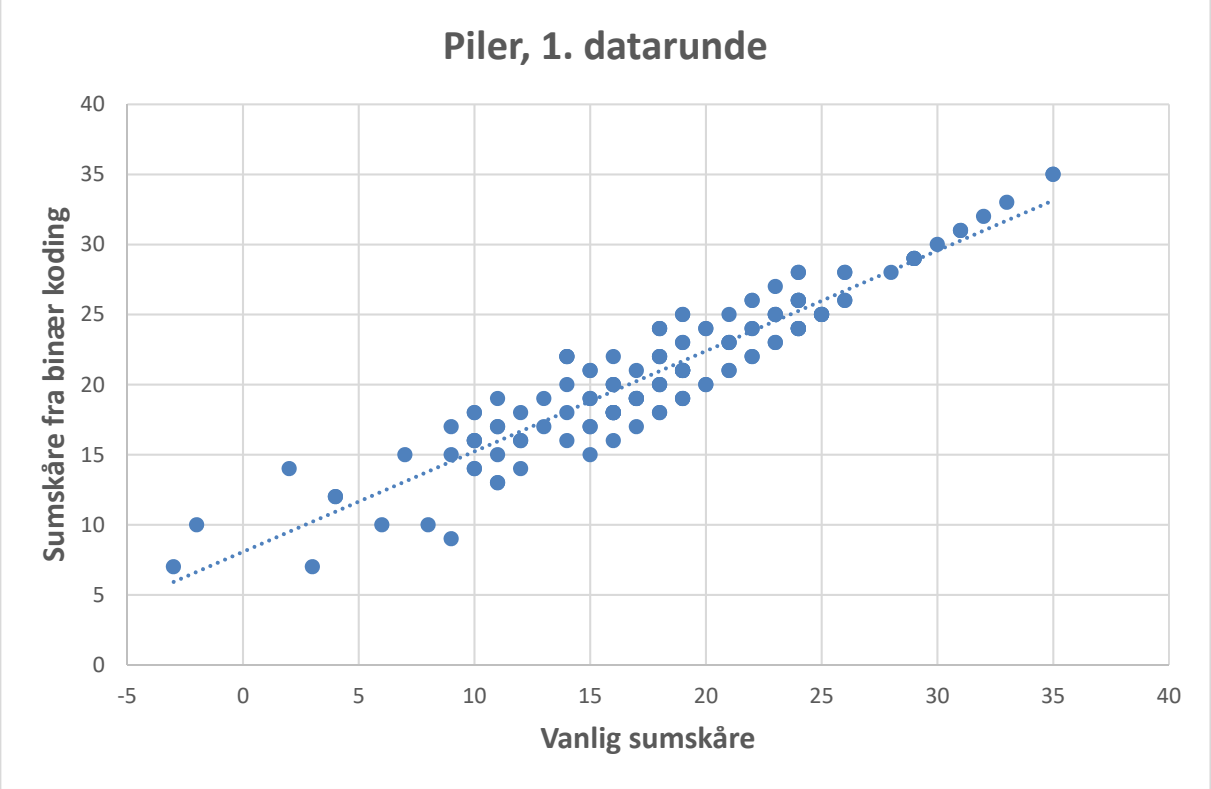
Her ser vi for det første at plottet gjenspeiler den klart signifikante korrelasjonen ($r = 0{,}79$; $p < 0{,}001$) mellom de to sumskårene med ulikt opphav. Observasjonene viser vel en rimelig grad av samling langs trendlinjen, og det virker sannsynlig at de to variablene langt på vei måler det samme.

Likevel er også avvikene betydelige, og viser at de to variablene slett ikke er identiske. Dersom de to skalaene måler det samme, så blir de i det minste påvirket av litt ulike former for 'støy' eller forstyrrelser. Her kan vi derfor ikke utelukke at det kan være noe ulike fordeler og ulemper knyttet til de to ulike formene for uthenting av data fra spillet.

Noe enklere bilder får vi i figurene 35 og 36 på neste side. Både for Trio- og Piler-spillene er samsvaret mellom de to skalaene svært høy, og observasjonene ligger langt tettere samlet rundt trendlinjen enn de gjorde for Hund/katt-spillet. Det ser derfor ut til at den 'normale' og den binære skalaen ikke bare måler det samme i hovedsak i disse to spillene, men også at det neppe er særlig store ulikheter i den 'støyen' som er knyttet til de to målene. Da er det heller ikke særlig sannsynlig at det kan finnes grunner til å foretrekke ett av målene eller skalaene framfor det andre.




**Figur 35:** Sumskårer fra 'Normale' og binære sumskårer, Trio-spillet i 1. datarunde

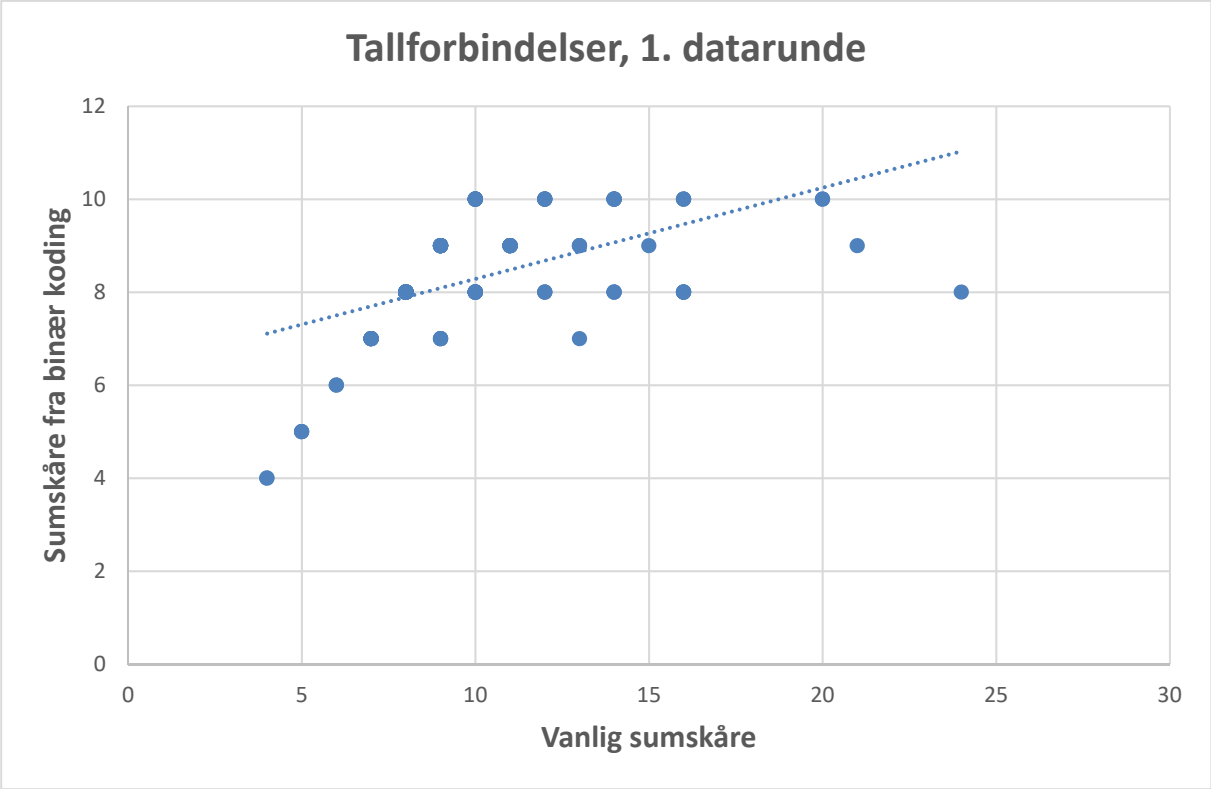



**Figur 36:** Sumskårer fra 'Normale' og binære sumskårer, Piler-spillet i 1. datarunde

Når vi kommer til spillet med tallforbindelser, er det imidlertid ikke fullt så enkelt. Her er korrelasjonen mellom 'normal' og binær sumskala tydelig lavere (0,55), selv om den fortsatt er klart og tydelig statistisk signifikant ($p < 0{,}001$).

Men i figur 37 ser vi at observasjonene ikke er så godt samlet rundt trendlinjen. Mange observasjoner har et tydelig avvik, og kanskje gjelder dette særlig de høyeste og de laveste skårene på den vanlige sumskåren. Med litt velvilje kan det faktisk se ut til at plottet antyder et kurvilineært forhold. Det er nok ingen overraskelse at en stigende tendens i lave, vanlige skårer henger sammen med økende verdier på den binært baserte skalaen. Men at en stigende tendens i de høyeste vanlige skårene kan se ut til å henge sammen med en *fallende* tendens på den andre skalaen, er neppe som ventet.

Trolig bør vi se litt nærmere på dette, for f.eks. å finne ut om dette er en følge av tilfeldig varierende støy, eller om det er et signal om noe mer substansielt – og som foreløpig ikke er forstått. Et naturlig første skritt er da å se om disse sammenhengene gjentar seg ved 2. og 3. datainnsamling.




**Figur 37:** Sumskårer fra 'Normale' og binære sumskårer, Tallforbindelser-spillet i 1. datarunde

# 5. KORT SAMMENFATNING

Både erfaringene og datamaterialet fra pilotprosjektet er forholdsvis omfattende, og gir et visst grunnlag for vurdering av Yellow Red-applikasjonen.

En viktig side ved *Yellow Red* som testprosedyre, er at den virker motiverende for de fleste barna. Den presenteres (og oppleves) som et sett med spennende spill, som styrkes av at nettbrettet blir brukt. Det er likevel klart at situasjonen ikke passer like godt for alle barn, og at barnas reaksjoner og tilpasninger varierer en god del. Noe av denne variasjonen må forventes å gjenspeiles i barnas prestasjoner, som rimelig kan være. I likhet med andre tester vil også resultatene fra *Yellow Red* derfor være påvirket av andre forhold enn barnas eksekutive funksjoner, selv om det bare er disse man ønsker å vurdere.

Vi ser også at selv om *Yellow Red* nok opprinnelig ble utviklet for bruk på enkeltpersoner, lar den seg tilpasse til bruk i mindre grupper. I vårt materiale er barn testet i begrensede grupper, med inntil ni i hver gruppe. Så langt har dette vist seg gjennomførbart, og har gitt tillitvekkende data. Det synes imidlertid klart at nøye utformede instruksjoner og erfaren ledelse er viktige forutsetninger for de gode resultatene.

Selve prosedyren for spillene er 'innebygd' i et program for nettbrettene. Det innebærer forholdsvis komplekse relasjoner mellom stimuluspresentasjon, innhenting av informasjon og måling av reaksjonstid. Programmet må forstås som Chile-teamets intellektuelle eiendom. Det kan dessuten hverken «åpnes» eller modifiseres uten tilgang til betydelig datakunnskap. I hovedsak må Yellow Red-programmet derfor brukes uten særlige endringer.

En litt større frihetsgrad kan imidlertid finnes i prosedyrene for skåring og koding av responsene. Dette er ikke 'innebygget' i programmet i samme grad, slik at det er mulig å arbeide videre med den informasjonen programmet gir om flere sider ved de responsene som er gitt. Som antydet i forrige kapittel kan dette gi anledning til en viss videreutvikling av Yellow Red.[1]

[1] Siden dette ble skrevet, har faktisk prof. Rosas og hans team gjennomført en omlegging av skåringsalgoritmene for *Yellow Red (Rosas-Días, R., Espinoza, V., Santa-Cruz, C., & Martínez, C. (2022, Mai 2022). The Yellow Red Test. Preliminary Results of the Chilean standardization process [Power Point presentation].* Endringen synes å bygge på noen av våre spørsmål om omkoding til enklere skårer, og teamet gir tillitvekkende informasjon om følgene for både reliabilitet og validitet i et chilensk materiale. De reviderte prosedyrene for skåring vil nok derfor bli tatt i bruk i vårt fortsatte arbeid med *Yellow Red*.

# 6. LITTERATURLISTE


Andersen, P. N., Klausen, M. E., & Skogli, E. W. (2019). Art of Learning -- An Art-Based Intervention Aimed at Improving Children's Executive Functions. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01769

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037-2078.

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*, 135-168. https://doi.org/10.1146/annurev-psych-113011-143750

Fallmyr, Ø., & Egeland, J. (2011). Psykometriske egenskaper for den norske versjonen av Behavior Inventory of Executive Function (BRIEF). *Tidsskrift for Norsk Psykologforening*, *48*, 339-343.

Garolera, M. (2019). *YellowRed International Results* [Internal report]. Escuela de Psicologia de la Pontificia Universidad Católica de Chile.

Hundevadt, M. O., & Klausen, M. E. (2019). *Kan kunst være nøkkel for utvikling av eksekutive funksjoner hos barn? Avsluttende rapport for forskningspiloten "Kunsten å lære"*.

Lehto, J. E., Juujärvi, P., Kooistra, L., & Pulkkinen, L. (2003). Dimensions of executive functioning: Evidence from children *British Journal of Developmental Psychology*, *21*, 59-80.

Margolis, J. L., Nussbaum, M., Rodriguez, P., & Rosas, R. (2006). Methodology for evaluating a novel education technology: a case study of handheld video games in Chile *Computers & Education*, *46*, 174-191. https://doi.org/10.1016/j.compedu.2004.07.007

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, Design, and Analysis. An Integrated Approach*. Lawrence Erlbaum Associates.

Rosas-Días, R., Espinoza, V., & Garolera, M. (2019). *Intercultural evidence of a Tablet based executive functions test for children between 7 to 10 years* EARLI (European Association for Research on Learning and Instruction) conference, Aachen.

Rosas-Días, R., Espinoza, V., Santa-Cruz, C., & Martínez, C. (2022, Mai 2022). *The Yellow Red Test. Preliminary Results of the Chilean standardization process* [Power Point presentation].

Rosas, R., Espinoza, V., & Garolera, M. (2020). Evidencia intercultural de un test basado en Tablet para medir las functiones ejecutivas de niños entre 6 y 10 años: resultados preliminares. *Papeles de Investigatión (CEDETI, Escuela de Psicologia de la Pontifica Universidad Católica de Chile)*, *2020*(12).

Rosas, R., Espinoza, V., Garolera, M., & San-Martin, P. (2017). Executive Functions at the start of kindergarten: are they good predictors of academic performance at the end of year one? *Studies in Psychology (Estudias de Psicología)*, *38*(2), 451-472.

Rosas, R., Nussbaum, M., Cumsille, P., Marianov, V., Correa, M., Flores, P., Grau, V., Lagos, F., López, X., López, V., Rodriguez, P., & Salinas, M. (2003). Beyond Nintendo: design and assessment of educational video games for first and second grade students. *Computers & Education*, *40*(2003), 71-94.

Sørensen, L., & Hysing, M. (2014). Måleegenskaper ved den norske foreldreversjonen av Behavior Rating Inventory of Executive Function (BRIEF). *PsykTestBarn*, 2-6.

Tenorio, M., Arango, P., Aparicio, A., & Rosas, R. (2014). TENI: A comprehensive battery for cognitive assessment based on games and technology. *Child Neuropsychology*. https://doi.org/10.1080/09297049.2014.977241

Wright, A., & Diamond, A. (2014). An effect of inhibitory load in children while keeping working memory load constant. *Frontiers in Psychology*, *5*(213), 1-9. https://doi.org/10.3389/fpsyg.2014.00213

Høgskolen
i Innlandet

Yellow Red-testen tar sikte på å måle noen sider ved barns eksekutive funksjoner. Den er utformet som fire dataspill for bruk på Android-nettbrett. Selv om spillene opprinnelig var laget for bruk på ett enkelt barn, kan den administreres i mindre grupper. De fleste barn lar seg engasjere i spillene, og gjennomfører som planlagt.

Resultatene fra en pilotundersøkelse virker i all hovedsak tillitvekkende. Svarfordelingene samsvarer f.eks. både med rimelige forventninger om normalfordeling, og med vanlige endringer og modning hos barn.

Testen er rettighetsbelagt. Den er tilgjengelig som en nedlastbar app, men krever passord for å la seg bruke. På kort sikt er det derfor neppe hensiktsmessig å søke etter forbedringer av testprosedyren. Kanskje bør man heller se nærmere på de nåværende reglene for skåring av responsene. Forsøk med alternative kodingsalgoritmer kan muligens trekke ut annen relevant og interessant informasjon.