FFI Forsvarets forskningsinstitutt

21/00737

FFI-RAPPORT

FFIs prediksjonsturnering

– datagrunnlag og foreløpige resultater

Alexander William Beadle

FFIs prediksjonsturnering
– datagrunnlag og foreløpige resultater

Alexander William Beadle

Forsvarets forskningsinstitutt (FFI) 13. januar 2022

Emneord
Prediksjon
Sikkerhetspolitikk
Forsvarspolitikk
Etterretning
Forsvarsplanlegging

FFI-rapport
21/00737

Prosjektnummer
1553

Elektronisk ISBN
978-82-464-3385-1

Engelsk tittel
FFI's forecasting tournament – dataset and preliminary results

Godkjennere
Sigurd Glærum, *forskningssjef*

Dokumentet er elektronisk godkjent og har derfor ikke håndskreven signatur.

Sammendrag

All etterretningsvurdering og forsvarsplanlegging utleder eller baserer seg på noen antagelser om hvordan Norges forsvars- og sikkerhetspolitiske omgivelser vil utvikle seg i årene som kommer. Vi har imidlertid visst lite om hvor godt prediksjonene som legges til grunn faktisk treffer.

Hensikten med FFIs prediksjonsturnering (2017–2020) var å måle hvor presist det er mulig å forutsi forsvars- og sikkerhetspolitiske utviklinger av relevans for Norge, og hva som kjennetegner personer som treffer bedre enn andre. Deltagerne ble bedt om å forutsi spørsmål som: Vil russiske militære fly krenke norsk luftrom det neste året? Hva blir utfallet av Brexit? Hvor mange NATO-land vil bruke 2 % av BNP på forsvar i 2024? Vil Trump vinne det neste presidentvalget? Hvis vi kan forutsi svarene på spørsmål som dette flere måneder og år i forveien, kan vi også være relativt sikre på retningen på utviklingen i Norges strategiske omgivelser fremover. I FFIs turnering ble det stilt 240 slike spørsmål om blant annet krig og konflikt, Russland, USA, Europa, økonomi og teknologi. Totalt ble det samlet inn 465 673 prediksjoner fra 1375 deltagere.

På den ene siden demonstrerer resultatene fra FFIs turnering hvor vanskelig det er å forutsi internasjonal politikk. Deltagerne sliter generelt med å treffe bedre enn tilfeldig gjetning, selv på kortsiktige spørsmål. Eksperter treffer bedre enn amatører, men forskjellene er små i praksis. Eksperter treffer heller ikke bedre på spørsmål innenfor sine egne fagområder enn eksperter med kompetanse på helt andre temaer. Kriteriene som vi normalt bruker til å avgjøre hvem vi skal høre på, som utdanningsnivå, relevant erfaring og spisskompetanse på aktuelle temaer, fremstår derfor som lite relevante i prediksjonssammenheng.

På den annen side viser resultatene at det er systematiske forskjeller i treffsikkerheten på individuelt nivå. Deltagernes prediksjonsevne korrelerer med en rekke individuelle egenskaper som kan kartlegges på forhånd, som kognitiv kontroll, tallforståelse, politisk kunnskapsnivå og grad av fordomsfri tenkning. Mange av disse egenskapene er trolig også overførbare til prediksjon i den virkelige verdenen fordi de samsvarer med forskning om hva som korrelerer med høyere prestasjonsevne på helt andre områder og i andre situasjoner. Det identifiseres også et sett med spesifikke teknikker forbundet med bedre treffsikkerhet, som det å lete etter informasjon fra flere kilder og å bruke metoder som grunnfrekvens, referanseklasser og ekstrapolasjon. Andre teknikker som ofte trekkes frem i fremtidsforskningen, ser derimot ikke ut til å ha betydning.

I stedet for å trekke et skarpt skille mellom prediksjonsturneringer og prediksjon i den virkelige verdenen, fremstår turneringer som et alternativ til dagens praksis i forsvars- og sikkerhetspolitiske analyser, der antagelser om den fremtidige utviklingen baseres på eksperter som ikke nødvendigvis har de riktige forutsetningene for å treffe best mulig. Det er nemlig mulig å bruke turneringer til å identifisere en gruppe deltagere som klarer å forutsi internasjonal politikk svært godt. De færreste av dem er profesjonelle eksperter. I stedet kjennetegnes de først og fremst av enda høyere scores på de individuelle egenskapene som korrelerer med bedre prediksjonsevne generelt og ved at de tenker på de riktige måtene når de predikerer.

Summary

The purpose of FFI's forecasting tournament (2017–2020) was to measure how accurate it is possible to predict political events and developments of relevance to Norwegian national security and what characterises people who are more accurate than others. The participants were given questions such as: Will Russian military aircraft violate Norwegian airspace within the next year? Will Russia conduct live fire exercises outside the Norwegian coast? What share of its GDP will Norway spend on defence? If it is possible to predict the outcome of questions such as these, we can also be relatively certain about the future development of key topics. FFI's tournament included 240 such questions about armed conflict, Russia, the US, Europe, economy and technology. In total, the dataset consists of 465,673 predictions from 1,375 participants.

FFI's tournament was inspired by the *Good Judgment Project* (GJP)'s tournament (2011–2015). In fact, FFI's participants have been measured on almost all of the same individual variables as in GJP. Thus, FFI's tournament can be used to re-examine key findings from GJP, based on a comparably sized dataset with a completely different set of participants and questions.

On the one hand, the results from FFI's tournaments find that the ability to predict international politics correlates with many of the same individual characteristics as in GJP, especially cognitive control, numeracy, knowledge, open-minded thinking and time used per question, but not with cognitive styles such as the need for cognitive closure or fox- vs. hedgehog-like thinking. However, these findings are nuanced through questionnaires with FFI's participants, which show that the specific cognitive styles participants used when the actually predicted were still important. In fact, specific approaches reflecting need for cognitive closure and the distinction between foxes and hedgehogs are both associated with lower accuracy in FFI's tournament, even though the participants' general scores on tests of these styles are not.

On the other hand, FFI's participants are significantly less accurate than GJP's. However, this gap is mainly due to differences in how the tournaments were organised. First, GJP's participants could update their forecasts every day until question closure, while FFI's could only make predictions during the first week after the questions were published. While the former way of forecasting is relevant to intelligence, the latter is more representative of how prediction is done during defence planning processes. Second, the accuracy of GJP's participants was improved through training and participation in collaborative teams, while FFI's participants predicted alone with no training. When these two differences are taken into account, the gap is greatly reduced.

Yet, the most important finding is that the best participants («superforecasters») were about equally accurate in both tournaments when based on the same time of prediction, even though all of GJP's superforecasters were both trained and part of collaborative teams. This raises a question of whether there is an «upper limit» of how accurate it is possible to predict politics and that this level of precision can be achieved simply by identifying the right people using forecasting tournaments. In fact, FFI's and GJP's superforecasters share a set of common characteristics, which makes it possible to identify them in advance.

Innhold

Forord

Denne rapporten presenterer de første resultatene fra FFIs prediksjonsturnering. Turneringen ble arrangert fra september 2017 til desember 2020 og ble gjennomført som en del av prosjektene «Globale trender og militære operasjoner» II (2016–2019) og III (2019–2022).

FFIs prediksjonsturnering var inspirert av en tidligere turnering, sponset av amerikansk etterretning fra 2011 til 2015, der fem akademiske lag konkurrerte om hvem som var best til å forutsi en rekke politiske hendelser. Vinnerlaget ble *Good Judgment Project* (GJP), ledet av professorene P. E. Tetlock og B. Mellers. Funnene fra GJP viste at det er mulig å forutsi hendelser relativt sikkert flere måneder frem i tid, at det er systematiske forskjeller i hvor gode individer er til å predikere og at det finnes noen enkeltpersoner som er spesielt gode («superforecastere»).

I FFIs turnering ønsket jeg å undersøke om funnene fra GJP også gjelder norske deltagere og spørsmål av betydning for Norge og Forsvaret spesielt. Resultatene er spesielt relevante for etterretnings- og forsvarsplanleggingsmiljøer samt fagfolk som arbeider med trender innenfor internasjonal politikk, krig og konflikt, teknologi, økonomi eller bestemte aktører og regioner.

Det endelige datagrunnlaget fra FFIs prediksjonsturnering vil bestå av omtrent like mange deltagere, spørsmål og prediksjoner som sammenlignbare studier fra GJP bygger på. Dette til tross for at GJP var en del av et forskningsprosjekt som kostet flere titalls millioner kroner, mens den samlede kostnaden for FFIs turnering var rundt fem millioner kroner. Kostnadsforskjellen skyldes blant annet at det i GJP ble gjennomført en rekke eksperimenter for å identifisere hvilke tiltak som kunne forbedre treffsikkerheten deltagerne underveis, mens det ikke ble gjort tilsvarende forsøk i FFIs turnering. Likevel treffer FFIs beste deltagere omtrent like godt som GJPs beste, uansett tiltak, fordi det mest effektive tiltaket er ganske enkelt å identifisere «de riktige personene». Disse deltagerne treffer også så godt som det synes å være mulig å forutsi forsvars- og sikkerhetspolitiske spørsmål. Bruk av turneringer for å identifisere norske superforecastere fremstår derfor som en nyttig metode for å forbedre treffsikkerheten også i den virkelige verdenen. Verktøyene som har blitt utviklet for å gjennomføre FFIs turnering kan gjenbrukes i andre deler av forsvarssektoren der det kan være aktuelt å måle eller forbedre treffsikkerheten.

FFIs turnering hadde aldri blitt noe av uten de hundrevis av deltagerne som har predikert hver måned i over tre år og besvart en rekke tester og spørreundersøkelser underveis. Turneringen er også et resultat av stor velvilje fra prosjektets forskningsledere (Sigurd Glærum og Alf Christian Hennum). Jeg har selv stått for gjennomføringen, spørsmålsgenereringen og forskningsdesignet, og er derfor ansvarlig for alle resultatene som presenteres. Jeg har imidlertid fått uvurderlig hjelp fra kollegaer ved FFI til kvalitetssikring av spørsmålene, utvikling av verktøyene og diskusjoner om resultatene, spesielt Vårin Alme, Sverre Diesen, Maria Fauske Fleischer, Halvor Kippe, Torbjørn Kveberg, Tobias Lillekvelland, Maria Lie Selle, Frank Steder og Kristian Åtland samt mine to doktorgradsveiledere Håvard Mokleiv Nygård og Jacob Aasland Ravndal.

Alexander W. Beadle,
Kjeller, 4. desember 2021

1 Innledning

Det er få områder hvor feilslåtte antagelser kan få større konsekvenser enn i forsvars- og sikkerhetspolitikken. I etterretningsvurderinger og forsvarsplanlegging er det helt nødvendig å gjøre analyser av det fremtidige trusselbildet, økonomiske trender og nye teknologiske muligheter. Samtidig er det umulig å vite helt sikkert hva som vil skje i fremtiden. Dette gjør beslutningsgrunnlaget svært avhengig av spesielt eksperters subjektive vurderinger av hva de tror vil skje. Problemet er at vi vet svært lite om hvor godt prediksjonene som gjøres faktisk treffer.

For å måle treffsikkerheten på forsvars- og sikkerhetspolitiske spørsmål har Forsvarets forskningsinstitutt (FFI) arrangert en treårig prediksjonsturnering (2017–2020). Denne rapporten presenterer det endelige datagrunnlaget og turneringens foreløpige svar på fire forskningsspørsmål:

1) Hvor presist er det mulig å predikere forsvars- og sikkerhetspolitiske utviklinger?

2) Er eksperter bedre til å predikere forsvars- og sikkerhetspolitiske utviklinger enn andre?

3) Finnes det individer som er bedre til å predikere forsvars- og sikkerhetspolitiske utviklinger enn andre?

4) Hva kjennetegner individene som er best til å predikere forsvars- og sikkerhetspolitikk?

Disse spørsmålene har tidligere blitt undersøkt i to amerikanske forskningsprosjekter – *Expert Political Judgment* (EPJ) og *Good Judgment Project* (GJP) – hvorav det siste er basert på resultatene fra en prediksjonsturnering som inspirerte FFI til å organisere sin egen. Det viktigste funnet fra disse prosjektene var at det er systematiske forskjeller i hvor gode enkeltpersoner er til å predikere. Mens utdanningsnivå og arbeidserfaring hadde lite å si, var de mest treffsikre kjennetegnet av høyere intelligens, større kunnskapsnivå og mer fordomsfrie måter å tenke på. Det fantes også en gruppe personer som traff svært godt og mye bedre enn resten («superforecastere»).

Det er imidlertid ikke gitt at funnene fra disse tidligere prosjektene er overførbare til en norsk forsvars- og sikkerhetspolitisk kontekst. Sammenhengene mellom individuelle egenskaper og treffsikkerhet kan være ulik for norske og amerikanske fagfolk. Spørsmålene som er blitt brukt til å måle treffsikkerheten i tidligere studier er også utviklet fra et amerikansk perspektiv. Det er ikke sikkert at spørsmål om temaer av størst betydning for norsk sikkerhet hadde gitt det samme resultatet. Videre har de fleste tidligere spørsmål hatt et relativt kortsiktig tidsperspektiv (rundt tre måneder), som gjør funnene mer relevante i etterretningssammenheng enn for forsvarsplanlegging, der perspektivet på langtidsplanene normalt er fire år.

Bakgrunnen for FFIs prediksjonsturnering var derfor å etterprøve tidligere funn med deltagere og spørsmål av større betydning for norske forsvars- og sikkerhetspolitikk. Totalt ble det stilt 240 spørsmål om hendelser og utviklinger basert på trender, aktører og regioner av særlig betydning for Norges strategiske omgivelser. Tidsperspektivet på spørsmålene varierte fra noen måneder til 3–4 år frem i tid, med et gjennomsnitt på rundt 1,5 år. Til sammen ble det samlet inn

465 673 prediksjoner fra 1375 deltagere, hvorav rundt halvparten arbeidet i forsvarssektoren, en tredel arbeidet med forsvars- og sikkerhetspolitiske spørsmål og rundt én av ti var eksperter brukt i media. Selv om turneringen omtales som FFIs, bestod den altså ikke bare av personer fra FFI, men et bredt spekter av deltagere med forskjellig bakgrunn og kompetanse.

I denne rapporten presenteres de første resultatene basert på de 150 første spørsmålene som hadde blitt avgjort ved utgangen av 2020. Dette foreløpige datagrunnlaget består av 274 764 sannsynlighetsestimater fra 833 deltagere som svarte på minst 20 % av de avgjorte spørsmålene.

Kapittel 2 beskriver hensikten og metoden til FFIs prediksjonsturnering. Her oppsummeres også de viktigste funnene fra de tidligere amerikanske forskningsprosjektene. Kapittel 3 inneholder en deskriptiv analyse av hele datagrunnlaget, inkludert hvem deltagerne var, hva slags spørsmål de fikk og antall prediksjoner som ble samlet inn. Kapittel 4 beskriver hvordan treffsikkerheten (rapportens avhengige variabel) og deltagernes bakgrunn, evner, kunnskapsnivå, tenkemåter og adferd i turneringen (de uavhengige variablene) har blitt målt.

Turneringen ble gjennomført som månedlige spørreundersøkelser med fem til syv spørsmål hver. Her ble deltagerne bedt om å oppgi hvor sannsynlig (i antall prosent) de mente hvert svaralternativ var. Treffsikkerheten ble først og fremst målt ved hjelp av Brier-score, som er et mål på evnen til å oppgi *høye* sannsynligheter til hendelser som *faktisk* skjer og *lave* sannsynligheter til de som *ikke* gjør det. Dette er spesielt relevant i forsvars- og sikkerhetspolitisk sammenheng, der hendelser er relativt unike, og de mest nyttige prediksjonene er dem som kan si at «dette vil skje» og «dette vil ikke skje». I tillegg ble deltagerne målt på treffprosent (hvor ofte de klarte å forutsi riktig utfall, uavhengig av sannsynligheter) og kalibrering (hvor sikre deltagerne var i forhold til hvor ofte de faktisk traff, som sier noe om hvor mye vi kan stole på prediksjonene).

Kapittel 5 presenterer og diskuterer de viktigste svarene på rapportens fire forskningsspørsmål:

- Deltagerne i FFIs turnering treffer generelt like dårlig som tilfeldig gjetning. Det vil si at de like så godt kunne fordelt sannsynlighetene helt likt på alle svaralternativer på alle spørsmål. De treffer også mye dårligere enn deltagerne i GJPs turnering. Gapet mellom de to turneringene skyldes imidlertid ikke forskjeller i deltagerne eller spørsmålene, men forskjeller i tidspunktene for når deltagerne kunne predikere og at det i GJP ble gjort tiltak for å forbedre treffsikkerheten underveis. Et overraskende funn er at det ikke finnes en sammenheng mellom spørsmålenes tidsperspektiv og treffsikkerheten i noen av turneringene, som tilsier at det ikke er vanskeligere å predikere jo lenger inn i fremtiden en ser. Samtidig gir funnene fra GJP grunn til å tro at deltagerne i FFIs turnering hadde truffet bedre, hvis de samme tiltakene hadde blitt gjennomført her. Disse tiltakene handlet først og fremst om å gi deltagerne opplæring i sannsynlighetstenkning og å sette dem sammen i grupper. Forskjellen mellom turneringene er også mindre når treffsikkerheten måles ved treffprosent, men FFIs deltagere er likevel altfor selvsikre.

- Ekspertene i FFIs turnering er bedre til å predikere enn amatørene, men forskjellene er små i praksis, uansett hvordan treffsikkerheten måles. I tråd med tidligere forskning fremstår tradisjonelle ekspertisekriterier, som utdanningsnivå, erfaring og kompetanse,

som irrelevante i prediksjonssammenheng. Det eneste kriteriet hvor det i FFIs turnering er en signifikant forskjell i treffsikkerheten handler om hvorvidt deltagerne hadde arbeidet med forsvars- og sikkerhetspolitiske spørsmål som en del av jobben sin eller ikke. Hvor lang erfaring eller hvor mye eller relevant spisskompetanse de har, spiller ingen rolle. Et overraskende funn er at eksperter ikke treffer bedre innenfor sitt eget fagområde enn fagfolk med kompetanse på helt andre temaer. I motsetning til tidligere funn er det ingen omvendt korrelasjon mellom treffsikkerheten og hvor mye ekspertene er brukt i media, som har vært et av de mest urovekkende funnene i eksisterende studier. Den største forskjellen på gruppenivå er mellom «akademikere» og «praktikere» i forsvarssektoren, der forskere treffer betydelig bedre enn offiserer. Alt i alt fremstår det imidlertid som lite hensiktsmessig å bruke formelle kriterier til å velge hvem vi skal høre på.

- Resultatene fra FFIs turnering bekrefter at det er systematiske forskjeller i individers evne til å forutsi forsvars- og sikkerhetspolitiske utviklinger. Resultatene gir også støtte til de fleste, men ikke alle, tidligere funn om hvilke individuelle egenskaper som henger sammen med bedre treffsikkerhet, spesielt høyere kognitiv kontroll, tallforståelse, politisk kunnskapsnivå og grad av fordomsfri tenkning. Dette understreker betydningen av det å bruke «de riktige folkene» til forsvars- og sikkerhetspolitiske analyser som krever prediksjon og muligheten for å bruke spesifikke tester for å identifisere dem på forhånd.

  Deltagernes treffsikkerhet korrelerer også med tiden de brukte på hvert spørsmål, men det viktigste er hva de brukte denne tiden på. I denne rapporten identifiseres det nemlig et sett med spesifikke måter å tenke på som kan bidra til bedre treffsikkerhet. Disse inkluderer ikke alle metodene som anbefales i fremtidsstudielitteraturen. Det viktigste er å unngå å bruke magefølelsen eller å velge det første som fremstår som mest sannsynlig, uansett hvilket grunnlag en selv tror at en har for å svare. De mest treffsikre tilnærmingene er å lete etter informasjon fra flere kilder, tenke over lignende historiske tilfeller med forskjellige utfall, ta utgangspunkt i dagens situasjon og bruke metoder som grunnfrekvens, referanseklasser og ekstrapolasjon for å fremskrive den videre utviklingen.

- Som i GJP finnes det en gruppe norske superforecastere i FFIs turnering. Disse skiller seg ikke bare ut ved å score enda høyere på de samme egenskapene som korrelerer med treffsikkerhet generelt, men de brukte også oftere de prediksjonsspesifikke metodene forbundet med bedre treffsikkerhet. Under like forhold treffer FFIs superforecastere faktisk like godt som GJPs superforecastere, til tross for at GJPs fikk alle tiltak som forbedret treffsikkerheten deres underveis. Den overraskende høye og like treffsikkerheten til superforecasterne i begge turneringer kan tyde på at det finnes en «øvre grense» for hvor godt det er mulig å predikere internasjonal politikk. Resultatene tyder også på at den største effekten av tiltak er å løfte treffsikkerheten til deltagerne generelt, mens det er mindre å hente for deltagere som i utgangspunktet treffer godt. Det enkleste og mest treffsikre tiltaket synes derfor å være å identifisere de aller beste superforecasterne.

For å forstå alle resultatene som beskrives i kapittel 5 forutsettes det en grunnleggende kjennskap til matematisk statistikk og statistiske metoder. Lesere som bare ønsker en oppsummering av de viktigste funnene kan hoppe rett til diskusjonene på slutten av hvert delkapittel.

Kapittel 6 diskuterer implikasjonene for prediksjon av internasjonal politikk i forbindelse med reelle trusselvurderinger, langtidsplaner og beslutninger. De fleste funnene om hvem som treffer best synes å ha stor overføringsverdi til den virkelige verdenen, fordi sammenhengene som identifiseres samsvarer med funn fra bedømmings- og beslutningspsykologien i helt andre situasjoner. Avslutningsvis argumenteres det for at prediksjonsturneringer kan være et mer treffsikkert alternativ til måten prediksjon gjøres i dag, basert på små fagmiljøer bestående av kun profesjonelle fagfolk som ikke nødvendigvis har de riktige forutsetningene for å treffe best mulig.

Vedlegg A inneholder alle testene som ble brukt til å måle deltagernes individuelle egenskaper. Dette inkluderer tester av evner som kognitiv kontroll, forståelse av tallkonsepter som sannsynlighet og kunnskap om internasjonal politikk, og av generelle tenkemåter, som aktiv fordomsfri tenkning, behov for kognitiv lukking og gleden av å engasjere seg i aktiviteter som krever tenkning. De fleste av disse testene ble oversatt til norsk i forbindelse med denne turneringen og kan gjenbrukes av andre.

Vedlegg B gir en deskriptiv analyse av det foreløpige datagrunnlaget, tilsvarende de som gjøres av hele datagrunnlaget i kapittel 3 og 4. Både spørsmålene og deltagerne som de foreløpige resultatene er basert på er imidlertid i stor grad representative for det komplette datagrunnlaget som vil foreligge når alle spørsmålene i turneringen er avgjort om noen år.

Alle 240 spørsmål som ble stilt i FFIs turnering, inkludert bakgrunnsinformasjon, svarkriterier og svaralternativer, er publisert i en egen FFI-rapport.[1] Idéen bak og metoden til turneringen er også beskrevet mer detaljert i en tidligere rapport.[2] Det samme er en grundigere gjennomgang av alle tiltakene som ble gjort for å forbedre treffsikkerheten til deltagerne i GJP.[3]

[1] Beadle, A. W. (2021), 'FFIs prediksjonsturnering – spørsmålskatalog', *FFI-rapport 21/00736* (Kjeller: FFI).
[2] Beadle, A. W. (2018), 'FFIs prediksjonsturnering – idé- og metodebeskrivelse', *FFI-rapport 18/00108* (Kjeller: FFI).
[3] Beadle, A. W. (2021), 'Hvordan forbedre treffsikkerheten til prediksjoner av internasjonal politikk? – en litteraturgjennomgang', *FFI-rapport 21/00735* (Kjeller: FFI).

2 Hensikt og metode

Dette kapittelet beskriver hensikten og metoden til FFIs prediksjonsturnering. Først diskuteres rollen til prediksjon i forbindelse med etterretning og forsvarsplanlegging. Deretter oppsummeres eksisterende forskning på prediksjon av internasjonal politikk og rapportens fire forskningsspørsmål i lys av denne. Til slutt forklares det hvordan FFIs turneringen ble gjennomført.

2.1 Prediksjon i forsvarssammenheng

Å predikere, eller forutsi, handler om å beskrive en fremtidig utvikling eller hendelse, gjerne ved hjelp av sannsynlighetsvurderinger. Selv om de aller fleste forsvars- og sikkerhetspolitiske studier innledes med beskrivelser av fremtiden som «usikker» eller «uforutsigbar», baserer anbefalingene seg ofte på antagelser og «prediksjoner» av hva som «sannsynligvis» vil skje.

I de norske sikkerhets- og etterretningstjenestenes åpne trusselvurderinger gjøres det ofte prediksjoner om hva som vil skje det neste året, for eksempel at det er lite sannsynlig med en fremforhandlet løsning mellom Russland og Ukraina, at antallet høyreekstreme terrorangrep i Vesten vil øke og at fredsprosessen i Afghanistan kan kollapse dersom de internasjonale styrkene trekkes ut før en politisk avtale mellom regjeringen og Taliban.[4] I analyser med lengre tidsperspektiv er det uvanlig med så tydelige prediksjoner, men de baseres vanligvis på flere antagelser om den fremtidige utviklingen likevel. Forsvarssektorens langtidsplaner innledes for eksempel med en rekke sikkerhetspolitiske rammer, som større betydning av nordområdene, fortsatt globalisering, videre russisk opprustning og økt spredning av masseødeleggelsesvåpen. Disse forventes å påvirke hvordan Forsvaret kan og bør utvikles for å håndtere fremtidige trusler best mulig.

Alle prediksjoner vil imidlertid være preget av usikkerhet. Til forskjell fra nåtiden og fortiden, finnes det ingen data om fremtiden som vi kan teste våre antagelser opp mot. Fremtiden eksisterer kun som tanker i våre hoder. Alle prediksjoner er derfor subjektive vurderinger av hva vi tror kan skje. Det er i tillegg psykologiske mekanismer som gjør at vi bommer mer enn nødvendig. Vi har blant annet en tendens til å velge bort informasjon som utfordrer våre eksisterende oppfatninger. Det var ikke først og fremst mangelen på etterretning som gjorde at en ikke klarte å forutse terrorangrepene 11. september 2001, men at den informasjonen som en hadde tilgjengelig ble feilaktig tolket.

Erkjennelsen av at våre prediksjoner ofte bommer, at endringer skjer raskere i dagens globaliserte omgivelser og at det kan være de mest overraskende hendelsene som får mest å si, har gjort at dagens fremtidsforskning og trusselvurderinger legger større vekt på analyser av alternative utviklingsløp, ikke bare fremskrivninger av dagens trender. I den sammenheng er det også andre evner enn treffsikkerhet som blir verdifulle, som forestillingsevne og teknikker for å tenke

[4] Eksempler hentet fra Politiets sikkerhetstjeneste (2021), *Nasjonal trusselvurdering 2021* og Etterretningstjenesten (2021), *Fokus 2021*.

«utenfor boksen». Anvendelse av scenarioer anses derfor som en «gullstandard» innenfor fremtidsmetoder, fordi de kan brukes til å utforske implikasjonene av flere mulige i stedet for bare én fremtid.[5]

Dette er spesielt viktig i forsvarsplanleggingen, der den viktigste hendelsen en skal planlegge for – en krig – er et svært sjeldent sosialt fenomen preget av stor uforutsigbarhet. Siden 2006 har FFI derfor brukt et sett med scenarioklasser til å identifisere gap i forsvarsstrukturen basert på det fremtidige trusselbildet.[6] Hensikten er å kartlegge *spekteret* av trusler Forsvaret må kunne håndtere de neste 10–25 årene, ikke forutsi én bestemt fremtid. FFIs scenarioklasser må likevel baseres på noen prediksjoner, fordi utvelgelsen av hvilke trusler som inkluderes er basert på eksperters vurderinger av hva som regnes som «mulig» i fremtiden. Okkupasjonstrusselen fra den kalde krigens dager er derfor ikke med i dagens scenarioer, fordi det er vurdert at «Russland ikke disponerer styrker – hverken i dag eller innenfor denne analysens tidshorisont – med kapasitet til å gjennomføre en invasjon av Norge».[7]

Scenarioklassene rangeres og revideres også ut fra hvor sannsynlige de anses å være basert på utviklingen i Norges sikkerhetspolitiske omgivelser. De minst alvorlige scenarioklassene, som terrorangrep og tvangsdiplomati, regnes som «utfordringer som kan oppstå til enhver tid», mens de to mest alvorlige klassene (strategiske overfall og begrensede angrep) ble i 2006 vurdert som «svært usannsynlige innenfor rammen av dagens sikkerhetspolitiske situasjon».[8] I 2014 ble det imidlertid vurdert at Russland «har utviklet en større evne og vilje til å anvende militærmakt på den måten som beskrives i FFIs mest alvorlige scenarioklasse».[9] Dette ble bekreftet kort tid etterpå da Russlands invaderte Ukraina. Scenariogrunnlaget ble revidert i 2018, der konklusjonen var at irregulære virkemidler, som dem brukt i Ukraina, var en reell sikkerhetsutfordring også for Norge.[10] I dag oppdateres FFIs scenarioer på nytt, basert på blant annet større usikkerhet rundt utsiktene til alliert støtte, utviklingen av kjernevåpen med lavere sprengkraft og en økende mengde sivil og militær infrastruktur i verdensrommet.

Det er derfor ikke mulig å trekke et klart skille mellom «prediksjon» og «kartlegging» av trusler. Målet i forsvarsplanleggingen kan ikke være å forutsi fremtiden, men å bomme så lite at det er mulig å tilpasse seg den når krigens realiteter avdekkes. Vi må erkjenne at mye av grunnlaget langtidsplaner, scenarioer og planverk bygger på i realiteten er basert på prediksjoner av hva vi tror vil skje. Studier har også vist vi ofte appliserer vår forståelse av dagens situasjon direkte på spørsmål om fremtiden på lang sikt. Det vi oppfatter som «mulig» i et 15–25 års perspektiv vil derfor alltid påvirkes av det vi tror i dag. Debatten om store forsvarsinvesteringer, som hvorvidt vi bør prioritere kampfly eller stridsvogner, domineres ofte av nylige hendelser – som Ukraina-

[5] Karlsen, J. E. og Øverland, E. F. (2010), *Carpe Futurum* (Oslo: Cappelen Damm).
[6] Johansen, I. (2006), 'Scenarioklasser i Forsvarsstudie 2007: En morfologisk analyse av sikkerhetspolitiske utfordringer mot Norge', *FFI-rapport 2006/02664* (Kjeller: FFI).
[7] Johansen (2006), 'Scenarioklasser i Forsvarsstudie 2007', s. 28.
[8] Johansen (2006), 'Scenarioklasser i Forsvarsstudie 2007', ss. 37–38.
[9] Bukkvoll, T., Glærum, S., Johansen, I., Diesen, S. og Lia, B. (2014), 'En gjennomgang av FFIs scenariogrunnlag for Forsvarets langtidsplanlegging', *FFI-rapport 2014/01154* (Kjeller: FFI). Begrenset.
[10] Åtland, K., Beadle, A., Diesen, S., Glærum, S., Mørkved, T., Nyhamar, T. og Stenersen, A. (2018), 'En gjennomgang av FFIs scenariogrunnlag for Forsvarets langtidsplanlegging, 2018', *FFI-rapport 18/00669* (Kjeller: FFI). Begrenset.

konflikten i 2014 – selv om dette er beslutninger med konsekvenser mange tiår fremover, der én konflikt ikke nødvendigvis vil være en god indikasjon på det langsiktige trusselbildet.

Det vi tror er sannsynlig på kort sikt kan derfor påvirke hva vi tror er mulig på lang sikt. I forsvarsplanlegging er det derfor spesielt viktig å vite hvor presist vi kan forutsi den sikkerhetspolitiske utviklingen og hvordan vi kan treffe best mulig på de spørsmålene som er predikerbare.

2.2 Tidligere forskning

Det teoretiske utgangspunktet for FFIs prediksjonsturnering er de samlede funnene fra to tidligere forskningsprosjekter. Det første var *Expert Political Judgment* (EPJ), som målte treffsikkerheten til profesjonelle eksperter på geopolitiske spørsmål med flere års tidsperspektiv. Det andre var *Good Judgment Project* (GJP), som identifiserte individuelle egenskaper som hang sammen med bedre treffsikkerhet. Til sammen representerer disse to prosjektene de mest systematiske studiene som er avgjort av individuelle prediksjoner av internasjonal politikk.

I det følgende oppsummeres det hvordan prosjektene ble gjennomført og de viktigste funnene.

2.2.1 *Expert Political Judgment* (EPJ)

Fra 1980-tallet og frem til 2003 samlet professor Philip E. Tetlock inn prediksjoner fra 284 politiske eksperter. Her ble ekspertene bedt om å anslå sannsynligheten (i prosent) til politiske, økonomiske og sikkerhetspolitiske utviklinger, både innenfor og utenfor sine egne fagområder. Dette arbeidet ble publisert i boken *Expert Political Judgment* (EPJ), som kom ut i 2005.[11]

Alle deltagerne i EPJ var «profesjonelle eksperter» som arbeidet med trender av betydning for stater, regioner eller verden generelt.[12] Ekspertene ble bedt om å gjøre én kortsiktig og én langsiktig prediksjon om utviklingen i 4 land, hvorav 2 lå innenfor og 2 utenfor deres eget kompetanseområde. For hvert land måtte de oppgi ett sannsynlighetsestimat for 3 ulike utfall på rundt 17 områder. Til sammen utgjorde dette rundt 140 spørsmål med 3 utfall hver per ekspert.[13] Ekspertene fikk imidlertid ikke de samme spørsmålene siden de hadde kompetanse på ulike land.

Spørsmålene dreide seg om fire temaer: 1) politisk styring og stabilitet, som valgresultater og kupp, 2) innenrikspolitisk og økonomisk utvikling, som BNP og rentenivåer, 3) forsvars- og sikkerhetspolitikk, som deltagelse i militære operasjoner og allianser, og 4) og forskjellige casestudier, som spredningen av masseødeleggelsesvåpen og maktskifter i tidligere kommunistland. De fleste spørsmålene ba ekspertene predikere 2, 5, 10 eller 20 år fremover, men da studien ble publisert var det bare noen få av spørsmålene som så 10 år eller lenger som hadde blitt avgjort.

[11] Tetlock, P. (2005), *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton: Princeton University Press). Denne oppsummering er basert på kapittel 2 og 3 som beskriver ekspertenes geopolitiske prediksjoner.
[12] For definisjonen av hvem som kvalifiserte som «ekspert» og mer informasjon om dem, se Tetlock (2005), *Expert Political Judgment*, s. 239ff.
[13] Antallet står ikke i Tetlock (2005), *Expert Political Judgment*, men ble opplyst gjennom korrespondanse 14.12.2020.

Resultatene var likevel nedslående. Ekspertene i EPJ klarte bare så vidt å slå tilfeldig gjetning, der en bare hadde fordelt sannsynligheten helt likt på alle utfall på alle spørsmål (f.eks. 50/50 % på et spørsmål med to utfall). Dette gav opphav til utsagnet som prosjektet ble mest kjent for: at eksperter er like dårlige til å predikere som en pilkastende ape med bind for øynene, der det er helt tilfeldig hvor godt man treffer.[14]

Tetlock fant likevel at det var systematiske individuelle forskjeller. Det var mulig å skille mellom to stereotyper eksperter – pinnsvin og rever – basert på hvordan de tenkte:[15]

- *Pinnsvinene* var kjennetegnet av at de kunne én eller to store emner eller teorier, som globalisering, maktbalanseprinsippet eller sivilisasjonskonflikt, som de appliserte på alle spørsmål (deduktiv resonnering). De plasserte komplekse problemer inn i årsak-virkningsforhold som de kjente fra før, mens det som ikke passet inn ble behandlet som irrelevant. Pinnsvinene var svært selvsikre i sine prediksjoner, og hadde lettere for å avvise motsigende synspunkter. De brukte gjerne ord som «dessuten», «og så videre» og «i tillegg til» for å trekke inn ytterligere argumenter for hvorfor de hadde rett, og skydde ikke ord som «umulig» eller «sikkert» i sine vurderinger av fremtiden. Gale prediksjoner ble bortforklart ved at de «bommet litt på tidspunktet», var «nesten riktige» eller at de ble avsporet av «uforutsigbare» hendelser.

- *Revene* var derimot kjennetegnet av at de kunne mange forskjellige, men ikke så store ting. De var skeptiske til store idéer om hvordan verden henger sammen og hvilke lover som egentlig gjaldt. I stedet anvendte de forskjellige analytiske tilnærminger avhengig av problemet som skulle løses (induktiv resonnering). De samlet mer informasjon fra mange kilder før de bestemte seg. I språket sitt brukte de oftere ord som «men», «imidlertid», «selv om» og «på den annen side». De snakket også om muligheter og sannsynligheter, ikke sikkerheter – og hadde lettere for å innrømme feil.

Av disse to typene var reveekspertene mye bedre til å predikere enn pinnsvinekspertene.[16] Pinnsvinene gjorde det faktisk ofte dårligere enn apen. Revene slo apen, men klarte likevel bare så vidt å slå enkle algoritmer som predikerte «ingen endring» eller «dagens endringstempo».

2.2.2 *Good Judgment Project* (GJP)

Fra 2011 til 2015 ble det gjennomført en stor prediksjonsturnering i USA. Turneringen ble arrangert av den føderale etaten *Intelligence Advanced Research Projects Activity* (IARPA), som sponser forskningsprosjekter som kan løse spesielt vanskelige utfordringer for amerikansk etterretning. Hensikten var å identifisere metoder som kunne øke treffsikkerheten i etterretnings-

[14] På engelsk: *dart-throwing chimpanzee*. For en diskusjon av metaforen, se forordet og s. 68 i Tetlock, P. og Gardner, D. (2015), *Superforecasting: The Art and Science of Prediction* (London: Random House Books).
[15] Se kapittel 3–6 i Tetlock (2005), *Expert Political Judgment*, for mer om de to typene eksperter, og Teltock og Gardner (2015), *Superforecasting*, ss. 68–73 for en kort oppsummering av de samme funnene.
[16] Tetlock og Gardner (2015), *Superforecasting*, s. 68ff.

analyser. Fem lag fra akademia og næringslivet konkurrerte om hvem som var best til å predikere rundt 500 spørsmål om internasjonal politikk, for eksempel: Vil Nord-Korea detonere et atomvåpen de neste tre månedene? Hvor raskt vil Kinas økonomi vokse det neste kvartalet?

Ett av lagene som deltok var *Good Judgment Project* (GJP), som ble etablert av tidligere nevnte Tetlock og kollegaen professor Barbara A. Mellers.[17] Deres tilnærming var å lage en intern turnering innad i prosjektet, der de gjennomførte eksperimenter for å teste ut hvilke tiltak som kunne øke deltagernes og dermed lagets aggregerte treffsikkerhet. GJPs interne turnering ble gjennomført på en egen nettportal der deltagerne måtte oppgi hvor sannsynlig (i prosent) de trodde ulike utfall var. Deltagerne kunne oppdatere sine prediksjoner helt frem til spørsmålet ble stengt. De kunne også velge hvilke spørsmål de skulle svare på, men ble oppfordret til å svare på så mange spørsmål som mulig. De konkurrerte med hverandre, enten alene eller på ulike lag.

Til forskjell fra EPJ var resultatene fra GJP svært oppløftende. Allerede etter to år traff de aggregerte prediksjonene fra GJP så godt at de fire andre lagene i IARPA-turneringen ble lagt ned.[18] De to siste årene av prosjektet ble derfor brukt til å optimalisere metodene som hadde vist seg å fungere for å oppnå høyest mulig aggregert treffsikkerhet.

GJP bekreftet funnet fra EPJ om at det er noen personer som er bedre til å predikere enn andre. Det var også en gruppe deltagere som var veldig mye bedre enn resten – som de kalte «*superforecastere*» – og som er det funnet prosjektet har blitt mest kjent for. I tillegg viste eksperimentene at det var mulig å forbedre personers treffsikkerhet gjennom relativt enkle tiltak. Vinneroppskriften til GJP var en kombinasjon av å rekruttere de riktige folkene, sette dem i grupper, gi dem opplæring i sannsynlighetstenkning og bruke algoritmer som la størst vekt på prediksjonene til deltagerne som hadde truffet best før og dem som baserte seg på nyere informasjon.

[17] For mer informasjon om GJP-prosjektet, se Tetlock og Gardner (2015), *Superforecasting*, ss. 16–20 og ss. 87–96. For et intervju med Tetlock, se 'How to Be Less Terrible at Predicting the Future', *Freakonomics*, 14. januar 2016. Ifølge Tetlock er arbeidsdelingen mellom Mellers og ham at hun gjør den dype forskningen, mens han tar seg av kommunikasjonsarbeidet. Mellers er også førsteforfatter på de fleste artiklene basert på resultatene fra GJP. Se foredragsrekken 'Edge Master Class 2015: A Short Course in Superforecasting', *Edge*, 17. aug.–21. sep. 2015, del 1.

[18] Da IARPA-turneringen ble lansert var målet å slå et uvektet snitt av alle prediksjonene med 20 % det første året, 30 % det andre året, 40 % det tredje året og 50 % det fjerde året. Se 'Edge Master Class 2015: A Short Course in Superforecasting', del 2. GJPs beste deltagere og beste algoritmer slo målet om 50 % allerede etter det første året, og de fortsatte å gjøre det de neste tre årene. GJP var det eneste laget som konsistent slo IARPAs mål for de første to årene, så lagene ble i stedet slått sammen etter dette. GJP fikk dermed mulighet til å rekruttere flinke deltagere fra andre lag.

2.3 Forskningsspørsmål

Hensikten med FFIs prediksjonsturnering var å måle treffsikkerheten til det norske forsvars- og sikkerhetspolitiske miljøet på spørsmål av relevans for etterretningsvurderinger og forsvarsplanleggingen. I denne rapporten er analysen av de foreløpige resultatene avgrenset til fire forskningsspørsmål som danner et utgangspunkt for å etterprøve funnene fra tidligere forskning.

2.3.1 Generell treffsikkerhet

Det første spørsmålet som analyseres i denne rapporten er: *Hvor presist er det mulig å predikere forsvars- og sikkerhetspolitiske utviklinger?*

Dette spørsmålet handler i første omgang om hvor nøyaktig det er mulig å predikere forsvars- og sikkerhetspolitikk. Hvis det ikke er mulig å slå enkle tilnærminger som tilfeldig gjetning, har det lite for seg å forsøke å predikere i det hele tatt. Gitt at det er mulig å predikere, er det i andre omgang særlig interessant å vite hvor langt frem i tid denne treffsikkerheten strekker seg.

I EPJ nærmet treffsikkerheten til de politiske ekspertene seg tilfeldig gjetning på spørsmål som så 3–5 år frem i tid.[19] Ifølge Tetlock var en del av forklaringen at spørsmålenes tidsperspektiv lå utenfor det som syntes å være mulig å predikere.[20] I GJP derimot, var tidsperspektivet mye kortere. Her var det gjennomsnittlige tidsperspektivet bare litt over 100 dager.[21] Her traff deltagerne mye bedre, og den beste algoritmen slo tilfeldig gjetning 86 % av gangene.[22]

I FFIs turnering har hensikten vært å måle treffsikkerheten på spørsmål med et lengre og mer relevant tidsperspektiv for forsvars- og sikkerhetspolitiske analyser i den virkelige verdenen. Dette tidsperspektivet styres først og fremst av selve prosessene som analysen skal støtte, ikke av hvor langt det faktisk er mulig å predikere. For eksempel publiserer Etterretningstjenesten årlige trusselvurderinger. Utviklingen av Forsvaret styres etter langtidsplaner som normalt gjelder for fire år av gangen. Samtidig krever de største strukturvalgene og materiellinvesteringer i Forsvaret gjerne et tidsperspektiv på 15–25 år. De fleste forsvars- og sikkerhetspolitiske studiene som skal støtte forsvarsplanleggingen definerer imidlertid sjeldent tidsperspektivet som legges til grunn og selve prediksjonene, som at moderniseringen av det russiske forsvaret «trolig vil fortsette i mange år fremover».[23] Av denne grunnen har det også vist seg å være svært vanskelig å vurdere treffsikkerheten til tidligere langtidsplaners sikkerhetspolitiske beskrivelser.[24]

Hvis det viser seg at forsvars- og sikkerhetspolitiske utviklinger ikke er mulig å forutsi lenger enn et par år fremover, er det liten grunn til å tro at lignende prediksjoner vil treffe noe bedre på

[19] Tetlock og Gardner (2015), *Superforecasting*, s. 5.
[20] Tetlock og Gardner (2015), *Superforecasting*, s. 244.
[21] Tidsperspektivet til spørsmålene i GJP varierer mellom rundt 100 og 130 dager avhengig av kildene som brukes.
[22] Tetlock, P., Mellers, B., Rohrbaugh, N. og Chen, E. (2014), 'Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate', *Current Directions in Psychological Science*, 23:4, ss. 290–295, s. 291.
[23] Eksempel hentet fra Ekspertgruppen for Forsvaret av Norge (2015), *Et felles løft* (Forsvarsdepartementet) s. 18.
[24] Bæk, S. (2019), 'Forsvarets tidligere langtidsplaner – hvor godt har de sikkerhetspolitiske beskrivelsene truffet?', *FFI-notat 19/01609* (Kjeller: FFI).

lengre sikt. Det vil bare ta mye lengre tid å få svar på. De fleste spørsmålene i FFIs turnering har derfor et tidsperspektiv på 6, 12, 24 eller 36 måneder fremover. Det gjennomsnittlige tidsperspektivet var 521 dager, altså fire ganger så langt som spørsmålene i GJP, men samtidig innenfor de 3–5 årene som EPJ viste at det var mulig å treffe bedre enn tilfeldig gjetning.

2.3.2 Ekspertise

Det andre spørsmålet som analyseres i denne rapporten er: *Er eksperter bedre til å predikere forsvars- og sikkerhetspolitiske utviklinger enn andre?*

Forsvarsplanlegging baseres i stor grad på eksperters vurderinger av dagens trender. Eksperter brukes ofte i utvalg og kommisjoner som skal støtte forsvarsplanleggingen. Selekteringen er som regel basert på formell kompetanse og kjennskap til aktuelle fagområder. Antagelsen er at eksperter har bedre forutsetninger enn andre for å kunne vurdere fremtidig utvikling. Her er ikke prediksjonsevne det eneste viktige, men beskrivelser av fremtiden utgjør ofte en sentral del.

Det var derfor et overraskende funn at utdannings- og erfaringsnivået til ekspertene i EPJ hadde lite å si for hvor godt de traff.[25] Hvorvidt de hadde doktorgrad, politisk erfaring, tilgang til gradert informasjon, relevant ekspertise eller mange års erfaring hadde ikke betydning. Tvert imot var de mest brukte ekspertene de dårligste til å predikere. De mest selvsikre ekspertene var også dårligere enn andre fagfolk til å predikere innenfor sine egne fagområder. Også i GJP kom en overraskende stor andel av superforecasterne fra andre disipliner enn statsvitenskap, spesielt fysikk, biologi og programvareutvikling.[26] Den eksisterende forskningen tilsier derfor at relevant spisskompetanse ikke har noe å si for prediksjonsevnen på spørsmål om det samme temaet.

En generell forklaring på hvorfor eksperters treffsikkerhet ikke alltid er god, er at betydningen av ekspertise avtar når den iboende usikkerheten ved fenomenet som predikeres øker.[27] Internasjonal politikk anses for å være et slikt fagområde, der det finnes få sikre teorier og kausalsammenhengene ofte er svært komplekse. Forsvars- og sikkerhetspolitikk er om mulig beheftet med enda større usikkerhet, fordi krig er både et sjeldent og et uforutsigbart sosialt fenomen. Det kan derfor tenkes at betydningen av ekspertise er enda mindre på spørsmål om forsvars- og sikkerhetspolitikk enn den var på de mer generelle politiske spørsmålene som ble stilt i EPJ og GJP. I denne rapporten vil derfor betydningen av høyere utdanning, relevant erfaring, eksperters bruk i media og spesifikke kompetanse på forsvars- og sikkerhetspolitiske spørsmål etterprøves.

Hverken EPJ eller GJP har imidlertid målt treffsikkerheten til personer uten *noe* høyere utdanning. Alle deltagerne i begge prosjektene var høyt utdannede. EPJ bestod utelukkende av profesjonelle eksperter, der nesten alle hadde mastergrad. I GJP var minstekravet utdanning på bachelorgradsnivå, og over halvparten hadde utdanning på masternivå. Analogien til den pilkastende apen kan imidlertid gi inntrykk at «hvem som helst» kan treffe like godt som eksperter, men dette er altså ikke undersøkt. I FFIs turnering deltok det personer uten høyere utdanning

[25] Tetlock (2005), *Expert Political Judgment*, s. 54.
[26] 'Edge Master Class 2015: A Short Course in Superforecasting', del 2.
[27] For mer om dette, se det nye forordet i Tetlock, P. E. (2017), *Expert Political Judgment: How Good Is It? How Can We Know?* (New Jersey: Princeton University Press).

eller forsvars- og sikkerhetspolitisk erfaring, som også gjør det mulig å undersøke hvordan eksperter treffer sammenlignet med forsvars- og sikkerhetspolitisk interesserte amatører.

2.3.3 Individuelle egenskaper

Det tredje spørsmålet som analyseres i denne rapporten er: *Finnes det individer som er bedre til å predikere forsvars- og sikkerhetspolitiske utviklinger enn andre?*

Et gjennomgående funn i både EPJ og GJP var at det er systematiske forskjeller mellom hvor godt individer klarer å predikere, og at disse i stor grad handler om deres kognitive evner og måter å tenke på. I begge prosjektene ble det gjennomført bivariate korrelasjonsanalyser for å undersøke sammenhengene mellom treffsikkerheten og en rekke individuelle egenskaper som studier fra kognitiv psykologi har identifisert som potensielt relevante for individers prediksjonsevne. Høyere treffsikkerhet hang i disse prosjektene sammen med følgende variabler:[28]

- **Kognitive evner.** I GJP korrelerte høyere kognitive evner, inkludert abstrakt resonneringsevne, kognitiv kontroll og tallforståelse, med bedre treffsikkerhet.

- **Kunnskapsnivå.** Høyere score på tester av generell kunnskap om internasjonal politikk korrelerte også med bedre treffsikkerhet i GJP. Det samme gjorde vokabular, men sammenhengen var ikke like sterk som med politisk kunnskap.

- **Fordomsfri tenkning.** Høyere score på aktiv fordomsfri tenkning korrelerte også med høyere treffsikkerhet i GJP, men denne variabelen var mindre viktig enn kognitive evner og politisk kunnskapsnivå. I motsetning til EPJ, der ekspertenes behov for kognitiv lukking og i hvor stor grad de identifiserte seg selv som «pinnsvin» korrelerte med dårligere treffsikkerhet, var dette ikke tilfellet i GJP.

- **Innsats.** Foruten hva deltagerne kan og hvordan de tenker, handlet treffsikkerheten i GJP også om innsatsen deres i selve turneringen. Mer spesifikt korrelerte antallet prediksjoner per spørsmål og tiden deltagerne brukte per spørsmål sterkere med treffsikkerheten enn noen andre uavhengige variabler. Hvor mange spørsmål deltagerne svarte på, hang derimot ikke sammen med prediksjonsevnen.

Sammenhengene mellom treffsikkerhet og individuelle egenskaper har imidlertid ikke blitt etterprøvd i andre forskningsprosjekter. For å gjøre dette er deltagerne i FFIs turnering målt på de samme individuelle variablene som deltagerne ble i EPJ og GJP. Hvis deltagerne i FFIs og GJPs turneringer ligner mye på hverandre, gir dette også større grunn til å anta at de er representative for den typen personer som deltar i slike prediksjonsaktiviteter. Det betyr i så fall at vi kan forvente å finne de samme sammenhengene i andre, lignende turneringer.

[28] Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E. og Tetlock, P. (2015), 'The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics', *Journal of Experiment Psychology: Applied*, 21:1, ss. 1106–1115.

2.3.4 Norske superforecastere

Det fjerde spørsmålet som analyseres her er: *Hva kjennetegner individene som er best til å predikere forsvars- og sikkerhetspolitikk?*

Det mest kjente funnet fra GJP var at det fantes en gruppe personer som var svært god til å forutsi fremtidige hendelser.[29] Disse ble kalt «superforecastere» og bestod av de 2 % beste deltagerne i løpet av et år. De første superforecasterne ble identifisert etter at det første året var over. 70 % av superforecasterne forble imidlertid blant de 2 % beste etter det andre året, som tyder på en svært liten sannsynlighet for at prediksjonsevnen deres bare var tilfeldig. Identifisering av
superforecastere og algoritmer som optimaliserte prediksjonene deres var de tiltakene som bidro mest til å øke treffsikkerheten i GJP. Superforecasterne traff også 30 % bedre enn et prediksjonsmarked med amerikanske etterretningsanalytikere med tilgang til gradert informasjon.[30]

Foruten at de scoret høyere på alle testene av kognitive evne, kunnskapsnivå, fordomsfri tenkning og deltagelse i turneringen, skilte superforecasterne seg i tillegg ut på tre andre områder:

- **Motivert av å vinne.** På spørsmål om motivasjonen for å delta oppgav superforecasterne et høyere ønske enn andre deltagere om å havne blant de beste. Denne variabelen korrelerte også med treffsikkerhet, men mindre enn de fleste andre variabler.

- **Probabilistisk tilnærming til fremtiden.** Deltagerne i GJP scoret generelt lavere enn den amerikanske gjennomsnittsbefolkningen på tro på en gudommelig orden, der hva som skjer i fremtiden er bestemt av skjebnen, og høyere på en mer vitenskapelig og probabilistisk tilnærming, der fremtidige hendelser vurderes ut fra sannsynligheter og tilfeldigheter. Her scoret superforecasterne enda litt høyere enn resten av deltagerne.

- **Oppgavespesifikke ferdigheter.** Superforecasterne scoret også bedre på mål av ferdigheter knyttet til prediksjon spesielt. De var mer sensitive for variasjoner i omfanget på spørsmålet, som ulike tidsperspektiver og geografisk utstrekning. De var mer finkornede i sine prediksjoner. Det vil si at de brøt sannsynlighetsskalaen fra 0 % til 100 % ned i flere distinksjoner enn andre. De brukte f.eks. 22 %, 24 % og 26 % i stedet for å svare 25 % i alle tre tilfeller. Denne nedbrytningen bidro også til høyere treffsikkerhet.

I sum konkluderte GJP med at superforecasterne hadde en distinkt profil sammenlignet med de nest beste og resten av deltagerne. I denne rapporten vil det derfor undersøkes om det finnes en tilsvarende gruppe deltagere i FFIs turnering – norske superforecastere – som skiller seg fra resten av deltagerne basert på de samme karaktertrekkene som i GJP.

[29] Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L. og Tetlock, P. (2015), 'Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions', *Perspectives on Psychological Science*, 10:3, ss. 267–281.
[30] Tetlock, P. E., Mellers, B. A. og Scobilic, J. P. (2017), 'Bringing probability judgments into policy debates via forecasting tournaments', *Science*, 355:6324, ss. 481–483.

2.4 Prediksjonsturnering

Metodisk var FFIs prediksjonsturnering en blanding av EPJ og GJP. I likhet med EPJ var ambisjonen å måle hvor presist det er mulig å predikere og hvem som er bedre enn andre. Det ble derfor ikke gjennomført eksperimenter for å identifisere hvilke tiltak som kunne forbedre prediksjonsevnen underveis, slik som i GJP. I likhet med GJP ble innsamling av prediksjoner gjort ved å arrangere en turnering, der deltagerne konkurrerte om å levere de beste sannsynlighetsestimatene. Bakgrunnen for dette er at turneringer er én av relativt få måter som finnes for å samle inn mange prediksjoner fra et stort antall personer over tid.[31]

En viktig forskjell fra EPJ og GJP er at FFIs turnering var åpen for alle som ønsket å delta. Deltagere registrerte seg på turneringens nettside eller gjennom skjemaer på graderte nettverk i Forsvaret.[32] Alle måtte oppgi kjønn, alder, utdanningsnivå, tilhørighet til forsvarssektoren, forsvars- og sikkerhetspolitisk arbeidserfaring, hvilke temaer, aktører og regioner de i så fall hadde kompetanse innenfor, og hvorvidt de hadde blitt brukt som eksperter i media. I tillegg ble alle bedt om vurdere sin interesse for og tro på egen evne til å forutsi forsvars- og sikkerhetspolitikk.

Hver måned fikk alle registrerte deltagere tilsendt rundt fem til syv spørsmål. Spørsmålene tok utgangspunkt i temaer av relevans for norsk forsvars- og sikkerhetspolitikk, som: «Vil russiske militære fly ville krenke norsk luftrom det neste året?» og «Hvor mange NATO-land vil bruke minst 2 % av BNP på forsvar i 2020?». Som i EPJ og GJP ble deltagerne bedt om å oppgi hvor sannsynlig (i antall prosent) de trodde hendelsen eller de forskjellige svaralternativene var. Hvert spørsmål ble innledet med en kort beskrivelse av den siste utviklingen og eventuelle historiske data, f.eks. antall ganger russiske fly har krenket norsk luftrom.

Måten deltagerne i FFIs turnering leverte sine prediksjoner på, skilte seg fra GJPs turnering på ett vesentlig punkt. GJP bestod av en online portal, der deltagerne logget seg på og valgte hvilke spørsmål de ville svare på. Her kunne de også oppdatere prediksjonene sine så mange ganger de ville inntil spørsmålet ble avgjort, noe deltagerne også ble oppfordret til å gjøre.[33] I FFIs turnering besvarte deltagerne spørsmålene gjennom spørreundersøkelser tilsendt på mail. Deltagerne kunne i likhet med GJP velge hvilke spørsmål de ville svare på, men fikk i motsetning til GJP ikke mulighet til oppdatere prediksjonene sine senere. I stedet hadde deltagerne én uke til å besvare de månedlige spørsmålsrundene. I løpet av denne uken kunne deltagerne endre svarene sine, men det var bare den sist registrerte prediksjonene som ble stående.

Etter at hvert spørsmål ble avgjort, fikk alle deltagerne som hadde predikert en e-post med plassering og score på det aktuelle spørsmålet, pluss sammenlagt plassering og score basert på alle spørsmålene som hadde blitt avgjort så langt. På nettsiden ble det publisert topplister med de 20 beste deltagerne på hvert spørsmål og de 20 beste deltagerne sammenlagt per år og i turneringen

[31] De vanligste alternativet er prediksjonsmarkeder. For en sammenligning av treffsikkerheten ved bruk av prediksjonsturneringer og -markeder basert på eksperimenter i GJP, se Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L. og Mellers, B. (2017), 'Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls', *Management Science*, 63:3, ss. 587–900.
[32] For nettsiden til FFIs prediksjonsturnering, se https://prediksjonsturnering.ffi.no/.
[33] Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters', s. 268.

som helhet. Ved slutten av hvert år ble det også kåret vinnere innenfor ti forskjellige kategorier (f.eks. beste deltager, ekspert og offiser), som fikk et krus med tilhørende tittel som premie.

Mens GJP gjennomførte alle testene av kognitive evner, kunnskapsnivå og tenkemåter under selve registreringen, ble disse variablene målt gjennom fire spørreundersøkelser fordelt på det andre og tredje året i FFIs turnering. Terskelen for deltagelse ble ansett som potensielt for høy, hvis alle som ønsket å delta måtte bruke flere timer før de kunne begynne å predikere, og det ble ansett som viktigere å ha flere deltagere enn at alle måtte ta alle testene. Kognitive evner anses å være relativt stabile individuelle egenskaper, så tidspunktet deltagerne tok testene på kan antas ikke å ha betydning for resultatene. Til forskjell fra GJP var alle testene også frivillige, men antallet respondenter var omtrent det samme som på de månedlige spørsmålsrundene (500–600).

Etter at turneringen var avsluttet, men før de endelige resultatene var klare, ble det gjennomført en siste spørreundersøkelse, der deltagerne ble bedt om å beskrive med egne ord hvordan de tenkte når de svarte på spørsmålene. Hensikten var å samle mer detaljert informasjon om deltagernes adferd i turneringen, utover de rent kvantitative variasjonene som ble målt i GJP.

3 Datagrunnlag

I FFIs prediksjonsturnering ble det samlet inn 464 342 prediksjoner fra 1375 personer på 240 spørsmål. Dette kapittelet gir en deskriptiv analyse av det komplette datasettet.

Først beskrives spørsmålene som ble stilt, inkludert temaer, typer spørsmål og tidsperspektiver. Deretter beskrives det hvem deltagerne var, inkludert ekspertise innenfor forsvars- og sikkerhetspolitikk og variasjoner i deltagermassen ved forskjellige minstekrav. Til slutt beskrives antallet prediksjoner som ble samlet inn, basert på ulike definisjoner av en prediksjon.

I hvert delkapittel sammenlignes FFIs turnering med datagrunnlagene til EPJ og GJP. Hensikten er å kunne sammenligne størrelsene til, og variasjonene i, de tre datasettene som kan brukes til å besvare de fire forskningsspørsmålene beskrevet i forrige kapittel. Den deskriptive analysen av EPJ er basert på boken om prosjektet og korrespondanse med P. Tetlock. Analysen av GJP er basert på bøker og et tjuetalls akademiske artikler som er skrevet om resultatene, det fullstendige datasettet, replikasjonsdatasettene til relevante studier, et eget datasett med detaljert informasjon om alle spørsmålene som ble stilt og korrespondanse med B. Mellers. Denne rapportens analyser er basert på deltagerne og spørsmålene i replikasjonsdatasettene, som samsvarer mest med antallene analysert i de tilhørende artiklene. Prediksjonene som brukes til å beregne treffsikkerheten deres er derimot basert på det fullstendige datasettet, som inkluderer noen flere prediksjoner enn dem som er brukt til å beregne scorene som er oppgitt i replikasjonsdatasettene.

Oppsummert består FFIs datagrunnlag av flere spørsmål enn i EPJ og et antall som er sammenlignbart med GJP. FFIs turnering har en større andel spørsmål av relevans for norsk forsvars- og sikkerhetspolitikk, særlig om krig, konflikt og militære operasjoner, Russland og USA. Det deltok like mange eksperter i FFIs turnering som i EPJ. FFIs datagrunnlag har like mange eller omtrent halvparten av deltagerne i GJP, avhengig av hvilken av de to meste relevante studiene det sammenlignes med. Det var imidlertid en mye høyere andel av deltagerne i FFIs turnering som deltok alle årene enn i GJP, som gir et bedre grunnlag for å studere variasjoner i treffsikkerheten over tid. Hvis vi bare er interessert i hvor presist det er mulig å predikere forsvars- og sikkerhetspolitiske spørsmål, uten at dette er forsøkt påvirket underveis, er antallet deltagere som vi kan bruke til å måle dette langt høyere i FFIs turnering enn i alle GJPs studier.

FFIs turnering består også av både færre og flere prediksjoner enn GJP, avhengig av hvordan disse telles. Spørsmålene i FFIs turnering hadde flere svaralternativer per spørsmål. Dette gir mer nyanserte prediksjoner og er mer realistiske i forbindelse med støtte til beslutningssammenheng. FFIs datagrunnlag har også et lenger tidsperspektiv, både på spørsmålene som ble stilt og på prediksjonene som ble samlet inn.

3.1 Spørsmål

3.1.1 Antall

I FFIs turnering ble det stilt 240 spørsmål, fordelt på 40 månedlige runder fra 2017 til 2020. Av disse ble 88 spørsmål stilt det første året (sept. 2017–des. 2018), 65 spørsmål det andre året (jan. 2019–des. 2019) og 87 spørsmål det tredje året (jan. 2020–des. 2020). Rundene bestod normalt av 5–7 spørsmål, men fem av dem var «spesialrunder» med opptil 10 spørsmål hver. Til sammen varte FFIs turnering like mange måneder som GJP, fordi de fire turneringsårene der var tilpasset den akademiske kalenderen og varte bare 10 måneder hver, altså 40 måneder.

Det totale antallet spørsmål i FFIs turnering (240) var i utgangspunktet bare halvparten av spørsmålene som ble avgjort i løpet hele GJP (488).[34] De to studiene fra GJP som det er mest aktuelt å sammenligne FFIs resultater med, baserte seg imidlertid på færre spørsmål enn dette:

- Den første er en artikkel fra 2015, som analyserte sammenhenger mellom treffsikkerhet og deltagernes individuelle variasjoner generelt.[35] Det er her de tidligere nevnte funnene om evner, tenkemåter og innsats som korrelerte med treffsikkerhet er hentet fra (se underkapittel 2.2.3). Selve artikkelen baserte seg på 199 spørsmål fra de to første årene av GJP, mens datasettene inkluderer litt flere, men rundt 200 spørsmål.[36] Denne studien av individuelle variasjoner omtales derfor heretter som «GJP200», selv om replikasjonsdatasettet som denne rapporten tar utgangspunkt i inkluderer noen flere spørsmål (211).

- Den andre er også en artikkel fra 2015, men denne studerte bare «superforecasterne».[37] Det er her de ytterligere kjennetegnene på de beste deltagerne er hentet fra (se underkapittel 2.2.4). Selve artikkelen baserte seg på rundt 350 spørsmål fra de tre første årene av GJP.[38] Denne studien omtales derfor heretter som «GJP350», selv om replikasjonsdatasettet som denne rapporten tar utgangspunkt i inkluderer litt færre spørsmål (347).

[34] I det fullstendige datasettet til GJP finnes det totalt 614 spørsmål, men dette inkluderer mange som utgikk (*voided*) eller er irrelevante, fordi de handlet om andre temaer som Tour de France og amerikansk fotball. I et eget datasett med alle de relevante spørsmålene (tilsendt fra Mellers i feb. 2020), finnes det totalt 488 fra alle fire årene av GJP. Her er det registrert hhv. 212 og 365 spørsmål fra de to og tre første årene. Av disse er det hhv. 211 og 347 spørsmål som er inkludert i replikasjonsdatasettene til de sammenlignbare studiene som er basert på de to og tre første årene.
[35] Mellers mfl. (2015), 'The Psychology of Intelligence Analysis', heretter også omtalt som «GJP200-artikkelen».
[36] Det fullstendige datasettet inkluderer 234 spørsmål fra de to første årene, mens spørsmålsdatasettet fra Mellers inkluderer 212. Replikasjonsdatasettet til GJP200 inkluderer 211 spørsmål, og det er derfor disse som brukes her.
[37] Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters', heretter også omtalt som «GJP350-artikkelen».
[38] Ifølge GJP350-artikkelen ble det stilt «over 100 spørsmål» hvert de to første årene og «rundt 150 spørsmål» det tredje året, men det er ikke oppgitt et eksakt antall spørsmål som studien baserer seg på. I det fullstendige datasettet fra GJP finnes det 402 spørsmål fra de tre første årene, men replikasjonsdatasettet til GJP350 inkluderer hhv. 102, 109 og 136 spørsmål for hvert av de tre første årene, som gir 347 spørsmål totalt. Dette er omtrent like mange som i en annen studie basert på de tre første årene, som inkluderte 344 spørsmål. Se Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J. og Tenneye, E. R. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', *Management Science*, 63:11, ss. 3552–3565.

Det er ikke publisert tilsvarende studier av individuelle variasjoner eller superforecasterne basert på spørsmålene fra alle de fire årene av GJP. Studiene basert på det fjerde turneringsåret har i stedet fokusert på mer avgrensede problemstillinger, der bare noen typer spørsmål har vært relevante eller omhandlet eksperimenter som kun ble gjennomført i visse perioder av turneringen. GJPs andre studier er derfor stort sett basert på rundt 100–150 spørsmål hver.[39]

FFIs komplette datagrunnlag, som heretter omtales som «FFI240», består altså av litt flere og en del færre spørsmål enn de to mest relevante datasettene til GJP med hhv. 200 og 350 spørsmål. I EPJ svarte ekspertene på rundt 140 spørsmål hver, heretter omtalt som «EPJ140», men de fikk forskjellige spørsmål, og det finnes ingen oversikt over spørsmålene som ble stilt. Resten av den deskriptive analysen av spørsmålene er derfor avgrenset til FFIs og GJPs datasett.

3.1.2 Tema

For å kunne måle treffsikkerheten til deltagerne som forsøkte å predikere dem var det et felles kriterium i både FFIs og GJPs turneringer at spørsmålene måtte være *falsifiserbare*.[40] Det vil si at det må være mulig å slå fast om en hendelse faktisk skjer eller ikke. Dette forutsetter gjensidig utelukkende svaralternativer, tydelig definerte svarkriterier og et sluttidspunkt. Ta for eksempel spørsmålet: «Vil Putin stille som kandidat i Russlands presidentvalg i 2018?». For at svaret skulle bli «ja», kunne svarkriteriene være at Putin måtte formelt registreres som presidentkandidat innen fristen, mens annonseringer om at han plana å stille ikke var nok. Spørsmålet ville senest bli avgjort etter at det russiske presidentvalget hadde blitt avholdt i løpet av 2018.

Disse kriteriene gjør samtidig at de store spørsmålene, som hvilken utenrikspolitisk linje Russland vil føre fremover, ikke er mulig å stille i prediksjonsturneringer. Det er imidlertid mulig å lage klynger av falsifiserbare spørsmål, som til sammen kan være indikatorer på den generelle utviklingsretningen. Eksempler kan være antall russiske krenkelser av norsk luftrom, om Russland vil intervenere militært andre steder i verden og hvorvidt landet vil inngå en fredsavtale med Ukraina de neste årene. For å sikre mest mulig relevante prediksjoner, er det derfor viktig at disse små spørsmålene dekker forskjellige aspekter av informasjon av betydning for de store.

[39] Se Merkle, E., Steyvers, M, Mellers, B. og Tetlock, P. (2016), 'Item Response Models of Probability Judgments: Application to a Geopolitical Forecasting Tournament', *Decision*, 3:1, ss. 1–19; Merkle, E., Steyvers, M., Mellers, B. og Tetlock, P. (2017), 'A neglected dimension of good forecasting judgment: The questions we choose also matter', *International Journal of Forecasting*, 33:4, ss. 817–832, og Chang, W., Atanasov, P., Patil, S., Mellers, B. og Tetlock, P. (2017), 'Accountability and adaptive performance under uncertainty: A long-term view', *Judgment and decision making*, 12:6, ss. 610–626, som er basert på hhv. 133, 157 og 135 spørsmål.

[40] I EPJ og GJP er dette kriteriet omtalt som klarsynt-testen (*clairvoyance test*), fordi spørsmålene må være så tydelige at en klarsynt som kan se inn i fremtiden, vil kunne fortelle hva utfallet ble basert på de formulerte kriteriene.

Den første forskjellen mellom FFIs og GJPs turneringer gikk nettopp på innholdet i spørsmålene deltagerne fikk. Spørsmålene i FFIs turnering var basert på trender, aktører og regioner som tidligere fremtidsstudier har identifisert som relevante for norsk forsvars- og sikkerhetspolitikk,[41] mens spørsmålene i GJP ble laget av og skulle være relevante for amerikansk etterretning.[42]

Tabell 3.1 viser hvordan spørsmålene i FFI240, GJP200 og GJP350 fordelte seg på 18 tematiske kategorier. Hvert spørsmål kan tilhøre flere kategorier samtidig. For eksempel vil et spørsmål om oljeprisen bare registreres innenfor «økonomi», mens et spørsmål om NATO vil intervenere militært i Syria registreres i «krig, konflikt og militære operasjoner», «NATO» og «Midtøsten og Nord-Afrika». Ved å registrere spørsmålene innenfor flere kategorier som passer reduseres betydningen av subjektive vurderinger av hvilken kategori som passer «best». I stedet viser analysen hvilke kategorier som var de vanligste i hvert datasett.[43]

I FFIs turnering var krig, konflikt og militære operasjoner (35 %) den klart vanligste kategorien, etterfulgt av Russland, Europa, USA og økonomi (17–23 %). Dette utvalget reflekterer de viktigste aktørene og områdene i norsk forsvars- og sikkerhetspolitikk fra 2017 til 2020. Kategoriene med færrest spørsmål var Sentral-Asia, Afrika sør for Sahara, Mellom- og Sør-Amerika.

I GJP var Midtøsten og Nord-Afrika (33–38 %) den klart vanligste kategorien, etterfulgt av krig, konflikt og militære operasjoner, økonomi og Øst-Asia (16–20 %).[44] Dette skyldes antageligvis at det styrende kriteriet for spørsmålene var betydningen for amerikansk nasjonal sikkerhet.[45] Spørsmålskategoriene reflekterer således de viktigste fokusområdene for amerikansk etterretning i perioden turneringen ble gjennomført fra 2011 til 2015.

Den største forskjellen mellom spørsmålene i de to turneringene er antallet spørsmål om de viktigste aktørene for norsk sikkerhet. I GJP fikk deltagerne over dobbelt så mange spørsmål om nesten alle andre regioner enn Russland. GJP stilte heller nesten ingen spørsmål om USA eller NATO. Forklaringen på det lave antallet spørsmål om USA var at den amerikansk etterretningen bevisst unnlot å stille spørsmål om amerikansk innenrikspolitikk.[46] De få spørsmålene som omhandlet USA var derfor utelukkende av utenrikspolitisk karakter. FFIs turnering stilte mange spørsmål om USA, inkludert amerikansk innenrikspolitikk, som utfall av kongressvalg og beslutninger tatt under Trump-administrasjonen.

[41] Beadle, A. W. og Diesen, S. (2015), 'Globale trender mot 2040 – implikasjoner for Forsvarets rolle og relevans', *FFI-rapport 2015/01452* (Kjeller: FFI), og Beadle, A. W., Diesen, S., Nyhamar, T. og Bostad, E. K. (2019), 'Globale trender mot 2040 – et oppdatert fremtidsbilde', *FFI-rapport 19/00045* (Kjeller: FFI).
[42] Moore mfl. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', ss. 3555.
[43] I FFIs turnering ble spørsmålene i snitt gruppert innenfor flere kategorier per spørsmål (2,1) enn i GJPs (1,6). Dette tilsier at innholdet i FFIs spørsmål favnet noe bredere enn GJPs, men forskjellen er ikke veldig stor.
[44] Denne fordelingen var også svært stabil over tid, selv når alle 488 spørsmål inkluderes. Mot slutten økte andelen spørsmål om Russland, antageligvis pga. Ukraina-konflikten i 2014, men endte likevel bare opp på 7 % etter fire år.
[45] Chang, W., Chen, E., Mellers, B. og Tetlock, P. (2016), 'Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments', *Judgment and Decision Making*, 11:5, ss. 509–526, s. 512.
[46] Mellers, B., Tetlock, P. og Arkes, H. R. (2019), 'Forecasting tournaments, epistemic humility and attitude depolarization', *Cognition*, 188, ss. 19–26, s. 22.

Kategori	FFI240	GJP200	GJP350
Internasjonal politikk generelt (f.eks. FNs sikkerhetsråd, internasjonale avtaler, pandemier, klimaendringer)	19 (7,9 %)	9 (4,3 %)	20 (5,8 %)
Krig, konflikt og militære operasjoner (f.eks. militære øvelser og angrep, våpentester, fredsforhandlinger)	**83 (34,6 %)**	**40 (19,0 %)**	**68 (19,5 %)**
Teknologi (f.eks. digitale angrep, tester av ny teknologi, kryptovaluta)	29 (12,1 %)	8 (3,8 %)	18 (5,2 %)
Økonomi (f.eks. BNP, oljepris, våpensalg, handelsavtaler, sanksjoner)	**41 (17,1 %)**	**35 (16,6 %)**	**64 (18,4 %)**
Terrorisme (f.eks. omfang, taktikk)	15 (6,2 %)	0 (0 %)	0 (0%)
Norsk innenrikspolitikk (f.eks. valgresultat, innvandring, vedtak)	38 (15,8 %)	0 (0 %)	0 (0%)
Norsk utenrikspolitikk (f.eks. deltagelse i multilaterale avtaler, bilaterale forhold)	26 (10,8 %)	0 (0 %)	0 (0%)
Russland	**54 (22,5 %)**	5 (2,4 %)	16 (4,6 %)
Norden	9 (3,8 %)	0 (0 %)	0 (0 %)
Europa	**49 (20,4 %)**	**44 (20,9 %)**	**66 (19,0 %)**
NATO	16 (6,7 %)	1 (0,5 %)	1 (0,3 %)
USA	**45 (18,8 %)**	8 (3,8 %)	20 (5,8 %)
Mellom- og Sør-Amerika	2 (0,8 %)	11 (5,2 %)	24 (6,9 %)
Midtøsten og Nord-Afrika	31 (12,9 %)	**81 (38,4 %)**	**115 (33,1 %)**
Afrika sør for Sahara	1 (0,4 %)	24 (11,4 %)	32 (9,2 %)
Øst-Asia	19 (7,9 %)	**34 (16,1 %)**	**64 (18,4 %)**
Sentral-Asia	0 (0 %)	0 (0 %)	1 (0,3 %)
Sør- og Sørøst-Asia	4 (1,7 %)	15 (7,1 %)	32 (9,2 %)
Annet (ingen politisk relevans)	0 (0 %)	2 (0,9 %)	3 (0,9 %)

Tabell 3.1 Kategorisering av spørsmålene i FFIs og GJPs turneringer. De fem vanligste kategoriene innenfor hvert studie er uthevet med fet skrift.

3.1.3 Typer

Den andre forskjellen var at det ble stilt langt flere spørsmål med flere svaralternativer i FFIs turnering sammenlignet med GJP. I FFIs turnering kunne deltagerne få tre typer spørsmål:

- **Binære spørsmål.** Her er det bare ett riktig og ett galt svar, som «ja» og «nei». Eksempel: «Vil Putin stille som kandidat i Russlands presidentvalg i 2018?». Hvis du svarer at det er 80 % sannsynlig for at Putin stiller, betyr det samtidig at du tror det er 20 % sannsynlighet for at Putin *ikke* gjør det.

- **Kategoriske spørsmål.** Her finnes det bare ett riktig, men flere gale svar. Eksempel: «Fra hvilket parti vil forsvarsministeren komme fra etter stortingsvalget i 2017?», og følgende svaralternativer: R, SV, Ap, MDG, Sp, V, KrF, H, FrP. Her må det oppgis sannsynlighetsestimater for hvert svaralternativ og summen av dem må bli 100 %.

- **Ordinale spørsmål.** Her finnes det bare ett riktig svar, men noen gale svar er riktigere enn andre. Eksempel: «Hvor mange vil bli drept i islamistiske terrorangrep i Europa i 2017?», og følgende svaralternativer: 0–49, 50–99, 100–149, 150–199 og 200+. Som ved kategoriske spørsmål må det oppgis estimater for alle svaralternativene, men hvis det riktige ble 50–59, anses 0–49 og 100–149 som riktigere svar enn 150–199 og 200+.

I GJP fikk deltagerne i tillegg «betingede» binære spørsmål, der deltagerne ble bedt om å vurdere sannsynligheten for at den samme hendelsen skulle skje, gitt ulike forutsetninger. Deltagerne ble for eksempel spurt: «Vil Nord-Korea gjennomføre en ny prøvesprengning innen mars 2014?», men bedt om å oppgi hvor sannsynlig dette var hvis: a) FN innfører flere sanksjoner før dette tidspunktet og b) FN ikke innfører flere sanksjoner. Her kunne de for eksempel oppgi 90 % sannsynlighet for en prøvesprengning hvis FN innførte flere sanksjoner og 60 % hvis ikke. Treffsikkerheten ble bare beregnet for den betingelsen som inntraff.

FFIs turnering inkluderte også delvis betingede spørsmål, som for eksempel: «Vil Russland varsle minst én ny øvelse utenfor norskekysten det neste halvåret, og vil det gjennomføres skarpskyting?». Disse spørsmålet ble imidlertid behandlet som kategoriske, og deltagerne ble derfor bedt om å fordele sannsynlighetene på alle utfall samlet sett, f.eks. «ingen øvelse» (40 %), «øvelse uten skarpskyting» (40 %) og «øvelse med skarpskyting» (20 %).

Tabell 3.2 viser fordelingen av alle spørsmålstypene i FFI240, GJP200 og GJP350. Sammenlignet med GJP hadde FFIs turnering en mye større andel kategoriske og ordinale spørsmål. Nesten alle spørsmålene i GJP var vanlige eller betingede binære (80–85 %) mot bare en tredel av FFIs (31 %). Derimot var halvparten av FFIs spørsmål ordinale (47 %) og en femtedel kategoriske (22 %) mot bare en tiendedel av hver type i GJP (hhv. 8–13 % og 7 %).

Spørsmålstype	FFI240	GJP200	GJP350
Binære, betingede binære	74 (30,8 %)	148 (70,1 %), 32 (15,2 %)	206 (59,4 %), 72 (20,7 %)
Kategoriske	53 (22,1 %)	15 (7,1 %)	24 (6,9 %)
Ordinale	113 (47,1 %)	16 (7,6 %)	45 (13,0 %)

Tabell 3.2 Typer spørsmål i FFIs og GJPs turneringer.[47]

Den høyere andelen ordinale og kategoriske spørsmål betød også at FFIs turnering hadde flere svaralternativer per spørsmål. Tabell 3.3 viser hvor mange spørsmål med forskjellige antall svaralternativer det var i FFIs og GJPs turneringer. I snitt hadde hvert spørsmål i FFIs turnering 3,5 svaralternativer, mens GJP-utvalgene hadde ca. 2,3 alternativer per spørsmål.

Antall svaralternativer	FFI240	GJP200	GJP350
2	74 (30,8 %)	180 (85,3 %)	278 (80,1 %)
3	38 (15,8 %)	10 (4,7 %)	28 (8,1 %)
4	81 (33,8 %)	16 (7,6 %)	31 (8,9 %)
5	39 (16,2 %)	5 (2,4 %)	10 (2,9 %)
6	5 (2,1 %)	0	0
7	0	0	0
8	2 (0,8 %)	0	0
9	1 (0,4 %)	0	0

Tabell 3.3 Antall spørsmål per antall svaralternativer i FFIs og GJPs turneringer.

Det var ikke et bevisst valg å stille en så høy andel ordinale og kategoriske spørsmål i FFIs turnering. Spørsmål med flere alternativer er imidlertid mer interessante enn binære i analysesammenheng, fordi nyanser er viktig for politisk og militær beslutningstagere. Det er f.eks. mer interessant å vite *hvor mange* eller *hvilke* land som oppnår NATOs 2 %-mål innen 2024 enn *hvorvidt* et bestemt antall klarer det. Sammen med et større fokus på regioner, aktører og temaer av betydning for Norge, er FFIs datagrunnlag er derfor bedre egnet enn GJPs til å si noe om treffsikkerheten til fagfolk som skal støtte forsvars- og sikkerhetspolitiske beslutninger.

[47] Fordelingen av GJPs spørsmål er basert på datasettenes egen kategorisering av spørsmål som vanlige/betingende binære, binære med mer enn to svaralternativer (her kalt «kategoriske») eller ordinale. Dette er bare analysert i Moore mfl. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', som baserer seg på 344 spørsmål, men beskrivelsen er svært lik fordelingen basert på replikasjonsdatasettets 347 spørsmål i denne rapporten.

3.1.4 Tidsperspektiv

Den tredje forskjellen mellom FFIs og GJPs turneringer var spørsmålenes tidsperspektiv. Tidsperspektivet er her definert som antallet dager fra et spørsmål ble stilt til datoen det *senest* ville blitt avgjort. Noen spørsmål ble avgjort før denne sluttdatoen, hvis hendelsen det ble spurt om inntraff før tiden gikk ut, men det er irrelevant i denne sammenhengen, siden det var spørsmålets sluttdato deltagerne måtte forholde seg til når de skulle predikere.

I GJP var det gjennomsnittlige tidsperspektivet rundt 130 dager.[48] Bakgrunnen var at IARPA, som finansierte GJP, ikke ønsket å kaste bort millioner av dollar på et prosjekt som forsøkte å predikere lenger enn det som var mulig. Det relativt korte tidsperspektiv gjør samtidig funnene fra GJP mer relevant for etterretningsanalyser og trusselvurderinger enn for forsvars- og sikkerhetspolitiske studier som skal støtte forsvarsplanleggingen med et tidsperspektiv på flere år.

I FFIs turnering var det gjennomsnittlige tidsperspektivet 521 dager, altså fire ganger så langt som i GJP, men samtidig innenfor mulighetsrommet på 3–5 år som EPJ hadde vist at det var mulig å treffe bedre enn tilfeldig gjetning. Dette gjør det mulig å utforske hvorvidt GJPs funn om individuelle variasjoner i treffsikkerheten fortsatt gjelder på spørsmål med lengre tidsperspektiver. Tabell 3.4 viser hvordan spørsmålene fordelte seg på forskjellige tidsperspektiver i de to turneringene. Mens deltagerne i FFIs turnering stort sett ble bedt om å predikere 6, 12, 18 og 24 måneder fremover, ble deltagerne GJP knapt bedt om å forutse noe lenger enn ett år.

Tidsperspektiv	**FFI240**	**GJP200**	**GJP350**
Opptil 6 måneder (under 183 dager)	41 (17,1 %)	156 (73,9 %)	266 (76,7 %)
6–12 måneder (183–365 dager)	62 (25,8 %)	46 (21,8 %)	70 (20,2 %)
12–18 måneder (366–548 dager)	62 (26,3 %)	8 (3,8 %)	9 (2,6 %)
18–24 måneder (549–731 dager)	28 (11,7 %)	1 (0,5 %)	2 (0,6 %)
24–30 måneder (732–914 dager)	18 (7,5 %)	0 (0 %)	0 (0 %)
30–36 måneder (915–1097 dager)	13 (5,4 %)	0 (0 %)	0 (0 %)
Over 36 måneder (over 1097 dager)	15 (6,3 %)	0 (0 %)	0 (0 %)

Tabell 3.4 Tidsperspektiver på spørsmål i FFIs og GJPs turneringer.

I tillegg hadde også prediksjonene i FFIs turnering et lengre tidsperspektiv enn i GJPs. I GJP kunne nemlig deltagerne selv velge når de skulle svare på et spørsmål, og de kunne oppdatere sine prediksjoner hele tiden frem til spørsmålet ble avgjort. Det betyr at i praksis hadde mange

[48] Tidsperspektivet på spørsmålene i GJP200 og GJP350 var hhv. 133 og 126 dager. I GJPs artikler er det imidlertid benyttet en litt annen definisjon av tidsperspektiv. Her er tidsperspektivet basert på antallet dager fra spørsmålet ble åpnet til det *faktisk* ble avgjort, som kunne være før sluttdatoen. Derfor er tidsperspektivene som oppgis i denne rapporten litt høyere enn i GJPs artikler. I praksis er det imidlertid liten forskjell på disse måtene å måle tidsperspektivet på. Basert på replikasjonsdatasettene til GJP200 og GJP350 innebærer GJPs definisjon gjennomsnittlige tidsperspektiver på hhv. 120 og 113 dager – sammenlignet med de 133 og 126 dagene som følger av FFIs definisjon.

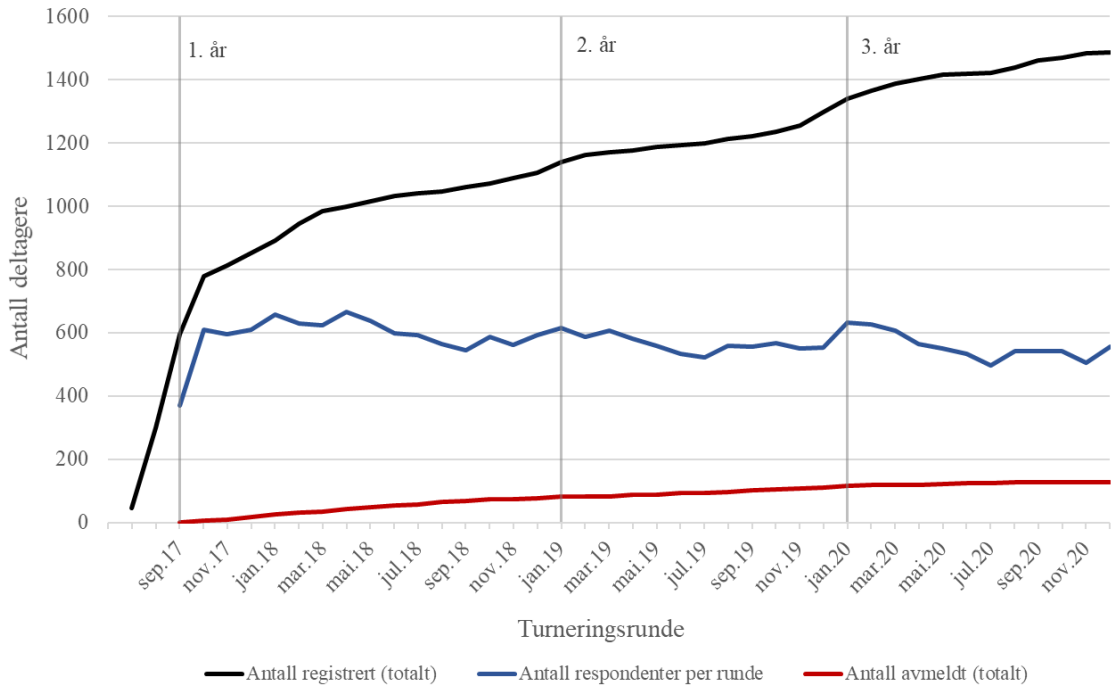
av prediksjonene et enda kortere tidsperspektiv enn de rundt 130 dagene de først ble bedt om å forutse. I FFIs turnering fikk deltagerne derimot bare mulighet til å besvare spørsmålene i løpet av den første uken etter at de ble publisert, ikke oppdatere prediksjonene underveis. Prediksjonenes tidsperspektiv var derfor alltid like lange som på spørsmålenes. FFIs turnering gir derfor et mer presist utgangspunkt for å måle sammenhengene mellom treffsikkerhet og tidsperspektiv.

Siden tidsperspektivet på spørsmålene i FFIs turnering er betydelig lengre enn i GJP, vil det ta noen år før alle spørsmålene er avgjort. Det er bare de 150 første av alle 240 spørsmålene som var avgjort ved utgangen av 2020 som analyseres i denne rapporten. Ved utgangen av 2021 vil imidlertid 200 (83 %) av spørsmålene vil være avgjort og rundt 220 (93 %) av spørsmålene ved utgangen av 2022.

3.2 Deltagere

3.2.1 Antall

Det var totalt 1485 personer som registrerte seg for å delta i FFIs turnering fra registreringen åpnet i juni 2017 til turneringen ble avsluttet i desember 2020 (se figur 3.1).



Figur 3.1 Deltagelse i FFIs prediksjonsturnering.[49]

[49] Avmeldte inkluderer alle deltagere som svarte på minst ett spørsmål, men som meldte seg selv av eller sluttet å svare over flere måneder og dermed ble slettet fra mottakerlisten. Tidspunktet for avmelding er da basert på den siste runden deltageren svarte, selv om de meldte seg av senere enn dette. I tillegg var det 101 deltagere som registrerte seg, men aldri svarte på noen spørsmål. Disse er inkludert i antallet registrerte deltagere og slettet etter hvert, men regnes ikke blant de avmeldte, siden de aldri deltok i utgangspunktet.

De aller fleste registrerte seg enten rett før eller rett etter den første spørsmålsrunden i september 2017 (sort graf). Dette samsvarer i tid med forsøkene på å rekruttere deltagere gjennom personlige invitasjoner til medlemmer av relevante fagmiljøer, kronikker, sosiale media og plakater på utdanningsinstitusjoner, som var omtrent på samme måte deltagerne ble rekruttert i GJP.[50] Etter det første året ble det ikke gjort nye rekrutteringsforsøk, men antallet registrerte fortsatte å stige. Det var også flere nye deltagere som registrerte seg enn eksisterende som meldte seg av (rød graf), som betyr at det totale antallet personer som mottok spørsmålsrundene økte gradvis.

Antallet respondenter på de månedlige spørsmålsrundene holdt seg imidlertid stabilt gjennom hele turneringen, med et snitt på 573 deltagere per runde (blå graf). Siden antallet registrerte steg uten en tilsvarende økning i respondenter, sank imidlertid svarprosenten utover i turneringen, fra et snitt på 65 % det første året til 43 % det tredje året.[51] For turneringen som helhet var den gjennomsnittlige svarprosenten 54 % blant alle deltagere som mottok spørsmålsrundene.

Av alle 1485 personene som registrerte seg for å delta, var det 1375 (93 %) som svarte på minst ett spørsmål i løpet av turneringen. De resterende 110 personene (7 %) registrerte seg, men deltok aldri. Av de 1375 deltagerne som deltok, var det 857 (62 %) som svarte på minst 20 % av spørsmålene, som var minstekravet som ble satt for å få en sammenlagt plassering. Hensikten med minimumskravet var å motivere deltagerne til å svare på flere spørsmål og å unngå at personer som traff spektakulært på bare noen få spørsmål kunne vinne turneringen.

Til sammenligning består det fullstendige datasettet til GJP av totalt 9185 deltagere som svarte på minst ett spørsmål i løpet av de tre første årene, men dette antallet er et misvisende mål på deltagelsen.[52] Det var nemlig bare 3002 av dem som er registrert med minst ett spørsmål ett av de to første årene, mens det tredje året kom det 6183 nye deltagere etter at GJP kunne rekruttere deltagere fra de andre lagene i IARPA-turneringen. I tillegg svarte de fleste deltagerne på bare noen få spørsmål, og en liten andel av dem oppfylte minstekravene i noen av studiene fra GJP.

GJP200-artikkelen om sammenhenger mellom treffsikkerhet og individuelle variasjoner er basert på 743 deltagere.[53] Her var minstekravet at deltagerne måtte å ha deltatt begge de to første årene og svart på minst 30 spørsmål til sammen. Dette kravet tilsvarer rundt 15 % av de 199 spørsmålene som ble stilt. GJP200s replikasjonsdatasett inkluderer noen flere deltagere (801) og spørsmål (211) enn artikkelen, og det er disse deltagerne og spørsmålene som analyseres i denne rapporten.[54] Til sammenligning var det 925 deltagere som oppfylte det samme minstekravet om å ha svart på minst 15 % av de 240 spørsmålene i FFIs turnering, altså flere enn i GJP200.

[50] Moore mfl. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', ss. 3555.
[51] Beregnet ved å dele antall respondenter per runde på antall registrerte, minus antall avmeldte på samme tidspunkt.
[52] Her er det tatt utgangspunkt i alle deltagere som er registrert med minst én prediksjon i det fullstendige datasettet på et av spørsmålene som ble publisert i løpet av de tre første årene. Hvilket år et spørsmål ble stilt er basert på hvilket år de ble kategorisert innenfor i det egne datasettet med alle relevante spørsmål fra GJP.
[53] Mellers mfl. (2015), 'The Psychology of Intelligence Analysis', s. 6.
[54] I det fullstendige datasettet er 3002 deltagere registrert med minst ett spørsmål i løpet av de to første årene. Replikasjonsdatasettet til GJP200 består derimot av 802 deltagere, som bare inkluderer deltagere som oppfylte artikkelens

GJP350-artikkelen om superforecasterne hadde et minstekrav om at deltagere måtte ha svart på minst 25 spørsmål i ett av de tre første årene.[55] Artikkelen oppgir ikke hvor mange den er basert på, men replikasjonsdatasettet inneholder 1751 deltagere som oppfyller dette minstekravet, og dette ligger svært nært antallet som fremgår av artikkelens korrelasjonsanalyse.[56] Siden det i snitt ble stilt rundt 115 spørsmål per år, tilsvarer dette kravet rundt 20 % av spørsmålene i ett av turneringsårene.[57] Til sammenligning var det 1087 deltagere som oppfylte det tilsvarende kriteriet i FFIs turnering, altså en del færre enn i GJP.

I EPJ var minstekravet strengere. Denne studien inkluderte bare de 284 ekspertene som hadde svart på minst 50 % av spørsmålene.[58] I GJP200 og GJP350 var det hhv. 508 og 345 deltagere som oppfylte dette kravet. I FFI240 var det 534 deltagere som svarte på minst 50 % av spørsmålene, altså nesten dobbelt så mange som i EPJ og minst like mange som i GJP.

Tabell 3.5 sammenligner antallene deltagere som oppfyller de forskjellige minstekravene som ble brukt i hver studie. Antallene er ikke direkte sammenlignbare, fordi hvert år i GJP ble regnet som en egen turnering, der det ble gjort justeringer og rekruttert flere deltagere før oppstarten av det neste. I FFIs turnering var det ingen pause mellom årene. Her representerte turneringsårene bare tidspunkter for kåring av vinnere og publisering av sammenlagte resultater. Sammenligningen gir likevel et inntrykk av den relative størrelsen på deltagermassene i alle studiene når de sammenlignes på likest mulig grunnlag. Tallene uthevet med fet skrift viser antallene deltagere basert på studienes egne minstekrav, mens tallene som ikke er uthevede viser antallene deltagere som ville oppfylt minstekravene gitt samme krav som ble brukt i de øvrige datasettene.

Tabell 3.5 viser også at minstekravet i FFIs turnering er strengere enn det som ble satt i GJP200 og GJP350. FFIs datagrunnlag er fortsatt større enn GJP200s uansett minstekrav, mens gapet mellom FFIs og GJP350s datasett reduseres betraktelig når GJPs datasett baseres på det samme kravet til deltagelse som i FFIs.

krav om å ha svart på minst 30 spørsmål til sammen. Én av disse er ikke registrert med prediksjoner i det fullstendige datasettet. Replikasjonsdatasettet som analyseres i denne rapporten består derfor av 801 deltagere.

[55] I GJP350-artikkelen formuleres minstekravet som om at deltagerne måtte ha svart på minst 25 spørsmål i *alle* de tre årene: «*To increase the reliability of performance estimates, we only included forecasters in both comparison groups if they had made forecasts for at least 25 questions in each tournament*.» Dette stemmer imidlertid ikke med replikasjonsdatasettet, der det bare er 270 deltagere som oppfyller dette kravet. Antallet deltagere som svarte på minst 25 spørsmål i *ett* av de tre årene, var 1751, som er tilnærmet identisk med det som diskuteres i artikkelen.

[56] I det fullstendige datasettet er 9185 deltagere registrert med minst ett spørsmålet i løpet av de tre første årene. Replikasjonsdatasettet til GJP350 består derimot av 2389 deltagere. Dette inkluderer bare deltagere som svarte på minst 10 spørsmål ett av årene. Av disse var det 1752 som oppfylte minstekravet om å ha svart på minst 25 spørsmål ett av årene. Én av disse er ikke registrert med prediksjoner i det fullstendige datasettet. Replikasjonsdatasettet som analyseres i denne rapporten består derfor av 1751 deltagere. Dette antallet ligger nært *df*-verdien på 1774 som er rapportert i korrelasjonsanalysen i Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters', s. 275, og som er en verdi som tilsvarer omtrent hvor mange deltagere utvalget som analysen baserer seg på.

[57] Basert på GJP350s replikasjonsdatasett med hhv. 102, 109 og 136 spørsmål det første, andre og tredje året.

[58] Korrespondanse med Philip Tetlock, 14.12.2020.

Minstekrav til deltagelse	FFI240[59]	GJP200[60]	GJP350[61]	EPJ140
Minst 15 % av alle spørsmål (GJP200)	925	**801**	1 236	-
Minst 20 % av spørsmål ett år (GJP350)	1 087	790	**1 751**	-
Minst 20 % av alle spørsmål (FFI)	**857**	770	1 040	-
Minst 50 % av alle spørsmål (EPJ)	534	508	345	**284**

Tabell 3.5 Antall deltagere, gitt ulike minstekrav.

I tillegg er det en betydelig grad av kontinuitet i hvem deltagerne i FFIs turnering var sammenlignet med deltagerne i GJP. Av de 857 deltagerne som oppfylte minstekravet i FFI240 var det hele 600 (70 %) som deltok alle tre årene. I GJP200 deltok alle 801 deltagere begge årene, siden dette var ett av minstekravene for å bli inkludert i studien, men denne studien baserte seg bare på to år, ikke tre som FFIs. I GJP350 var det derimot bare 402 (23 %) av 1751 deltagere som deltok alle tre årene. Det var med andre ord omtrent dobbelt så mange og en tre ganger så stor andel deltagere som deltok gjennom hele FFIs turnering sammenlignet med den største av GJPs studier. FFIs datagrunnlag gir således et enda bedre grunnlag enn GJPs for å måle konsistensen i personers treffsikkerhet over tid.

Denne relativt stabile deltagelsen i FFIs turnering er også overraskende, fordi at ingen av deltagerne fikk betalt for å delta. I GJP fikk derimot deltagerne et gavekort til en verdi av $150 hvis de leverte minst 25 prediksjoner det første året, $250 dollar hvis de oppfylte dette kravet det andre året, og ytterligere $100 hvis deltagerne deltok begge årene.[62] I FFIs turnering var den eneste belønningen status og krus til sammenlagtvinnere. Den stabile deltagelsen blir enda mer overraskende med tanke på at halvparten av deltagerne også måtte ha vært klar over at de befant seg på den nedre halvdel av resultatlisten, siden det bare var disse 857 som ble rangert og at de fikk vite sin sammenlagte plassering etter hvert spørsmål som ble avgjort.

[59] FFIs eget minstekrav var at deltagerne måtte ha svart på minst 20 % av alle spørsmål, som tilsvarer minst 48 av 240. GJP200s krav om minst 15 % av alle spørsmål tilsvarer minst 36. GJP350s krav om minst 20 % av spørsmålene ett av årene innebærer at deltagerne må ha svart på minst 18 av 88 spørsmål det første året, 13 av 65 spørsmål det andre året eller 18 av 87 spørsmål det tredje året. EPJs krav om minst 50 % av alle spørsmål tilsvarer minst 120.
[60] GJP200-artikkelens eget krav var minst 30 av 199 spørsmål over to år. Her inkluderes 801 deltagerne som oppfylte dette, selv om replikasjonsdatasettet de er hentet fra inneholdt 211 spørsmål (og en tilsvarende andel på 15 % derfor utgjør minst 32 spørsmål). GJP350s krav om minst 20 % av spørsmålene ett av årene innebærer at deltagerne i GJP200 må ha svart på minst 21 av 102 spørsmål det første året eller 22 av 109 det andre året i replikasjonsdatasettet. FFIs krav om minst 20 % av alle spørsmål tilsvarer minst 43 av 211. Minst 50 % av alle spørsmål tilsvarer minst 106.
[61] GJP350-artikkelens eget krav var minst 20 av spørsmålene ett av årene. Dette utgjorde 20 % av 102 spørsmål det første året, 18 % av 109 det andre året og 15 % av 136 det tredje året. GJP350s krav er derfor rundet av minst 20 % av spørsmålene ett av årene når antallet deltagere i de andre datasettene er beregnet. GJP200s krav om å ha svart på minst 15 % av alle spørsmål tilsvarer minst 53 av alle 347 spørsmål i replikasjonsdatasettet. FFIs krav om minst 20 % av alle spørsmål tilsvarer minst 70. EPJs krav om minst 50 % av alle spørsmål tilsvarer minst 174.
[62] Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters', s. 269.

3.2.2 Kjønn, alder og utdanning

Et annet sentralt spørsmål ved sammenligninger av studiene er hvordan deltagermassene ser ut. Hvis to studier baserer seg på veldig forskjellige deltagere kan avvikende funn skyldes utvalgene som sammenlignes. Hvis utvalgene er relativt like er det derimot bedre grunnlag for å kunne bekrefte eller avkrefte sammenhengene som eventuelt observeres.

Tabell 3.6 viser kjønnsfordelingen, snittalderen og utdanningsnivået til deltagerne i FFI240, GJP og EPJ. Her, og i alle påfølgende tabeller, legges studienes egne minstekrav til grunn for utvalget som analyseres, siden det er disse deltagerne de respektive studiene baseres på. Det vil si at det er de 857 deltagerne som svarte på minst 20 % av spørsmålene som ligger til grunn for FFIs turnering. Forskjellen mellom deltagerne de to og tre første årene av GJP er så små at det bare er GJP350s 1751 deltagere som analyseres her. Disse inkluderer uansett de fleste deltagerne fra de to første årene. EPJ140 inkluderer alle 284 eksperter som svarte på minst 50 % av spørsmålene. I FFIs turnering ble deltagerne bedt om å oppgi hvor mange års høyere utdanning de hadde, som utgjør alternativene i tabellen, mens i EPJ og GJP ble deltagerne bedt om å oppgi type utdanning. Her er derfor utdanningsnivåene oppgitt i EPJ og GJP tilpasset FFIs kategorier.

Sammenlignbare variabler	**FFI240[63]**	**GJP350[64]**	**EPJ140[65]**
Kjønn	90 % menn	84 % menn	76 % menn
Alder (snitt)	40 år	40 år	43 år
Ingen høyere utdanning	75 (9 %)	12 (1 %)	-
1–3 års høyere utdanning	194 (23 %)	544 (31 %)	11 (4 %)
4–5 års høyere utdanning	302 (35 %)	710 (41 %)	124 (44 %)
Over 5 års høyere utdanning	286 (33 %)	483 (28 %)	148 (52 %)

Tabell 3.6 Direkte sammenlignbare variabler i EPJ, GJP og FFIs turnering.

[63] Gjennomsnittsalderen er basert på deltagernes alder i 2017, som var da de fleste deltagerne registrerte seg.
[64] Kjønn og utdanningsnivå er basert på GJP350s datasett, mens snittalderen er fra artikkelen. Kjønnsfordelingen samsvarer med artikkelen (84 % menn), men datasettet inneholder ikke nøyaktig alder, bare aldersgrupper. Aldersgruppene med flest deltagere er 25–29 (404), 30–34 (388) og 35–39 år (222), som tilsier et lavere snitt enn 40 år, som oppgitt i artikkelen. I GJP200 fikk deltagerne valgene mellom utdanningsnivåene: ingen bachelor, bachelor, master eller doktorgrad, som her kategoriseres som hhv. ingen, 1–3, 4–5 og over 5 års høyere utdanning. Siden mange av deltagerne i GJP200 oppgav utdanningsnivået på nytt i GJP350, er det den siste opplysningen som legges til grunn her. I GJP350 fikk deltagerne valgene: ingen videregående skole, videregående skole, *associate degree*, bachelor, master, profesjonsutdanning/-doktorgrad eller annen doktorgrad (inkl. ph.d.). Her regnes ikke *assosiate degree* som høyere utdanning, siden dette er en mellomting mellom videregående skole og bachelornivå. Doktorgrad inkluderer alle med prof.utdanning/prof.doktorgrad, fordi de fleste som oppgav dette i GJP350 oppgav doktorgrad i GJP200. Andelen med doktorgrad er derfor i realiteten litt mindre, men ikke snakk om mer enn et par titalls deltagere. Analysen er avgrenset til de 1748 av 1751 deltagere med data på utdanningsnivået. I GJP350-artikkelen er det oppgitt at 64 % hadde utdanning på mastergradsnivå, som ligger nært andelen på 69 % basert på datasettene. For artikkelens tall på kjønn, alder og utdanningsnivå, se Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters', s. 269.
[65] Antall er beregnet ut fra 284 eksperter og andelene oppgitt i Tetlock (2005), *Expert Political Judgment*, s. 239ff. I EPJ hadde 96 % masternivå og 52 % doktorgrad, som her kategoriseres som hhv. 4–5 og over 5 års høyere utdanning. De siste 4 % regnes som bachelornivå, siden det antas at alle ekspertene hadde minst ett års høyere utdanning.

Felles for alle datagrunnlagene er at den store majoriteten av deltagerne var menn, gjennomsnittsalderen var rundt 40 år og de fleste var høyt utdannede. Det er bare noen små forskjeller i utdanningsnivåene. Over to tredeler av deltagerne i både FFIs og GJPs turneringer hadde minst 4–5 års høyere utdanning. Utdanningsnivået var ikke overraskende enda høyere i EPJ, siden studien bare inkluderte personer som var «profesjonelle eksperter», dvs. at de levde av å gi råd om politiske eller økonomiske utviklinger. FFIs turnering inneholder en liten gruppe deltagere uten noe høyere utdanning, som var helt fraværende i GJP, siden kravet for å delta var utdanning på minst bachelorgradsnivå.

Oppsummert er likevel deltagermassene i FFIs turnering, GJP og EPJ svært like basert på disse direkte sammenlignbare variablene. Det er også slående hvor lav kvinneandelen var i alle studiene. En mulig forklaring er at andelen kvinner allerede i utgangspunktet er lav i de fleste miljøene deltagerne ble rekruttert fra. For eksempel er kvinneandelen i Forsvaret under 13 %,[66] mens kvinner bare utgjør 20 % av forskerne ved FFI.[67]

3.2.3 Ekspertise

I FFIs turnering ble det også samlet inn informasjon om deltagernes ekspertise, inkludert hvor mange som har forsvars- og sikkerhetspolitisk arbeidserfaring, hvilke områder de i så fall hadde kompetanse innenfor, og hvor mange av dem som hadde blitt intervjuet som eksperter i media. Tabell 3.7 sammenligner derfor ekspertene i FFIs turnering med ekspertene i EPJ. GJP er ikke med her, fordi det ikke finnes informasjon om deltagernes ekspertise i noen av datasettene.[68]

FFIs turnering skiller seg først og fremst ut ved at hele 405 (47 %) av deltagerne arbeider innenfor forsvarssektoren. Av disse var 23 % militære, hvorav den største gruppen var offiserer opp t.o.m. oberstløytnantnivå (OF1–4), og 15 % var forskere fra ulike institusjoner i forsvarssektoren, hovedsakelig FFI og Forsvarets høgskole, inkludert underliggende krigsskoler.

Av de resterende 452 (53 %) deltagerne som ikke arbeidet innenfor forsvarssektoren, kom de fleste fra faglig, vitenskapelig og teknisk tjenesteyting, offentlig administrasjon, informasjon og kommunikasjon eller var pensjonerte/ikke yrkesaktive. Av disse deltagerne som ikke arbeidet innenfor forsvarssektoren i dag, var det likevel rundt en tredel som hadde gjort det tidligere. Det betyr at totalt 61 % av deltagerne hadde arbeidserfaring fra forsvarssektoren, og kan derfor antas å ha en viss grad av kjennskap til spørsmål av forsvars- og sikkerhetspolitisk art.

[66] 'Setter mål om 30 prosent kvinner', *Forsvarets Forum*, 4. juni 2019.
[67] Strand, K. R., Gisnås, H. og Eggereide, B. (2019), 'Utvikling i sentrale HR-parametere i forsvarssektoren – et dypdykk hos Forsvarsbygg og Forsvarets forskningsinstitutt', *FFI-rapport 19/01599* (Kjeller: FFI), s. 30.
[68] Den eneste tilleggsinformasjonen som finnes om deltagerne i GJPs datasett er antall timer de jobbet per uke, deres erfaring med og interesse for trading, og antall timer per uke de brukte på internasjonale nyheter og politikk.

	FFI240	**EPJ**[69]
Bransje/ sektor (andel av deltagere)	405 (47 %) forsvarssektoren, hvorav de fleste var: - 165 (19 %) offiserer - 126 (15 %) forskere - 38 (4 %) befal/grenader/konstabel 101 (12 %) faglig, vitenskap. og tekn. tjenester 61 (7 %) offentlig administrasjon 58 (7 %) informasjon og kommunikasjon 50 (6 %) ikke yrkesaktiv/pensjonert 181 (21 %) fordelt på resterende (opptil 4 % hver)	116 (41 %) akademia 74 (26 %) staten 48 (17 %) tenketanker og stiftelser 23 (8 %) internasjonale organisasjoner 23 (8 %) privat sektor (inkludert media)
Eksperter (andel av deltagere)	267 (31 %) arbeidet eller hadde arbeidet med forsvars- og sikkerhetspolitikk som en del av jobben (10 års arbeidserfaring i gjennomsnitt	Alle 284 var profesjonelle eksperter (12 års relevant erfaring i gjennomsnitt)
Ekspertise-områder (andel av eksperter)	134 (50 %) krig, konflikt og militære operasjoner 89 (33 %) internasjonal politikk 83 (31 %) NATO 74 (28 %) teknologi 70 (26 %) Russland 54 (20 %) USA 51 (19 %) norsk utenrikspolitikk 50 (19 %) Midtøsten og Nord-Afrika 49 (18 %) terrorisme 48 (18 %) Europa 47 (18 %) Norden 44 (17 %) økonomi 27 (10 %) norsk innenrikspolitikk Alle resterende regioner (opptil 23/10 % hver)	116 (41 %) områdestudier 68 (24 %) internasjonal relasjoner 34 (12 %) økonomi 31 (11 %) nasjonal sikkerhet og rustningskontroll 26 (9 %) journalistikk 6 (2 %) diplomati 3 (1 %) internasjonal rett
Intervjuet i media (andel av eksperter)	75 (28 %) med relevant arbeidserfaring oppgav at de hadde blitt intervjuet som eksperter i media om forsvars- og sikkerhetspolitiske spørsmål 60 (23 %) sitert eller omtalt minst 10 ganger	173 (61 %) intervjuet av minst ett stort medium 60 (21 %) intervjuet minst 10 ganger

Tabell 3.7 Sammenligninger av ekspertise.

I EPJ var derimot alle de 284 deltagerne profesjonelle eksperter, definert som en person som levde av å gi råd om politiske eller økonomiske utviklinger. I FFIs turnering var det 267 deltagere (31 %) som oppgav at de arbeidet eller hadde arbeidet med forsvars- og sikkerhetspolitiske spørsmål som en del av jobben sin, altså omtrent like mange «eksperter» som i EPJ. I snitt hadde ekspertene i FFIs turnering 10 års arbeidserfaring, omtrent like lenge som i EPJ (12 år). FFIs 267 eksperter inkluderer imidlertid bare deltagere som arbeidet spesifikt med forsvars- og

[69] Tetlock (2005), *Expert Political Judgment*, s. 239ff. Antall er beregnet ut ifra andelene oppgitt for 284 eksperter.

sikkerhetspolitiske spørsmål, ikke politikk generelt, slik som i EPJ. Det er derfor mulig at det egentlig finnes flere politiske eksperter i FFIs datagrunnlag, dersom den samme, bredere definisjonen til EPJ hadde blitt lagt til grunn.

Det sterkere fokuset på forsvars- og sikkerhetspolitiske spørsmål reflekteres også i hvilke områder ekspertene hadde kompetanse innenfor. Av ekspertene i FFIs turnering var det flest som oppgav ekspertise innenfor krig, konflikt og militære operasjoner, internasjonal politikk, NATO, teknologi og Russland. I EPJ kom de fleste ekspertene fra områdestudier, uten at det oppgis hvilke, og nest flest fra internasjonale relasjoner. Andelene deltagere med kompetanse innenfor ulike ekspertiseområder er imidlertid ikke direkte sammenlignbare, siden ekspertiseområdene i EPJ behandles som gjensidig utelukkende kategorier, mens i FFIs turnering kunne deltagerne oppgi ekspertise på flere områder og dermed kan telle innenfor flere kategorier.

Et siste relevant aspekt ved ekspertene er deres opptredener i media, siden ett av hovedfunnene i EPJ var at ekspertene som ble mest brukt i media også var de dårligste til å predikere. Av EPJs 284 eksperter hadde 173 (61 %) blitt intervjuet av minst ett stort medium og 60 eksperter (21 %) blitt intervjuet minst 10 ganger. I FFIs turnering var det bare 75 deltagere (28 % av ekspertene) som hadde blitt intervjuet, men igjen inkluderer dette bare eksperter som ble intervjuet i media om spesifikt forsvars- og sikkerhetspolitiske spørsmål. Av disse er det 60 deltagere som er sitert eller omtalt minst 10 ganger i løpet av det siste tiåret, basert på treff i den digitale arkivtjenesten Atekst, som inneholder internettkilder og artikler fra de største norske papiravisene, fagbladene og magasinene.[70] Det er altså nøyaktig like mange deltagere i FFIs turnering som i EPJ som er sitert eller omtalt minst 10 ganger, men det er ikke skilt mellom store og små medier i FFIs kartlegging, og det var langt flere eksperter i EPJ som var sitert færre ganger.

Et vesentlig forbehold når det gjelder resultatene fra prediksjonsturneringer som FFIs og GJPs, er at deltagerne ikke er representative for de generelle gruppene de kommer fra. De består av personer som på eget initiativ har registrert seg og deltatt i en turnering. Det er derfor ikke mulig å si noe om forskjellene mellom kjønn, alder, utdanning eller forsvars- og fagmiljøene generelt; bare om forskjellene mellom disse gruppene i FFIs turnering. Dette er et forbehold som også gjelder GJP og EPJ. Samtidig er størrelsen på hele populasjonen av personer som arbeider med forsvars- og sikkerhetspolitiske spørsmål i Norge ukjent, men den er i alle fall mye mindre enn i USA, slik at deltagerne i FFIs turnering utgjør en større andel av den reelle populasjonen.

I det nye forordet til en oppdatert utgave av EPJ-boken svarer Tetlock på flere innvendinger som har kommet mot funnene derfra.[71] En av dem er at det ikke er personene hvis prediksjoner virkelig betyr noe som deltar i prediksjonsundersøkelser. Slike studier kan derfor ikke si noe om treffsikkerheten til personer som står bak beslutninger i den virkelige verdenen. Tetlock hevder imidlertid at han har forsøkt å få «politiske eliter» til å delta i turneringer i over 30 år, men ikke lykkes. Han definerer ikke hva han mener med politiske eliter, men viser til eksperter og sentrale personer som er involvert i politiske beslutninger. Én av forklaringene på hvorfor disse ikke

[70] Søket ble gjort 1. juni 2021 og basert på antall treff på deltagernes eksakte navn i perioden 1.1.2010–31.12.2020.
[71] Tetlock (2017), *Expert Political Judgment*, ss. xvii–xliv.

ønsker å delta er ifølge Tetlock at allerede ledende eksperter, embedsmenn og politikere har lite å tjene personlig på å bli målt grundig på treffsikkerheten.

Her kan deltagermassen i FFIs turnering representere et steg i retning av det å måle treffsikkerheten til personer hvis treffsikkerhet betyr noe i den virkelige verdenen, fordi en tredel av deltagerne arbeider eller har arbeidet med forsvars- og sikkerhetspolitiske spørsmål som en del av jobben sin. Flere av dem er også blant de mest kjente ekspertene på sine respektive fagområder i Norge. Totalt kommer 16 % av deltagerne fra FFI, som er én av Forsvarsdepartementets viktigste strategiske rådgivere i nettopp den fremtidige utviklingen av Forsvaret, og 5 % arbeider ved Forsvarets høgskole, som utgjør et annet stort, akademisk miljø innenfor forsvarssektoren.

3.2.4 Variasjon

I sammenheng med spørsmålet om deltagerne i FFIs turnering er representative for personer som arbeider med forsvars- og sikkerhetspolitikk, er det viktig å undersøke om deltagermassen varierer med minstekravene som anvendes i analysen. Det kan for eksempel tenkes at deltagerne som registrerte seg, men predikerte lite, skilte seg fra dem som deltok aktivt. Da vil også funnenes relevans avhenge av hvilket minstekrav som ble satt. Det er imidlertid lite som tyder på det.

Tabell 3.8 viser hvordan deltagermassen endrer seg ved tre ulike minstekrav til deltagelse: fra et absolutt minimumskrav om å ha svart på minst ett spørsmål i løpet av hele turneringen, via minstekravet om å ha svart på minst 20 % av spørsmålene som denne rapportens analyser tar utgangspunkt i, til det strengeste kravet om at deltagerne må ha svart på minst halvparten.

Tabellen viser at minstekravene har lite å si for hvem deltagerne i FFIs turnering var. Høyere minstekrav gir en litt større overvekt av menn, et par år høyere gjennomsnittsalder og noen få prosenter flere personer med bakgrunn fra forsvarssektoren, mens utdanningsnivået og den forsvars- og sikkerhetspolitiske erfaringen holder seg likt. Den relativt sett største forskjellen er at kvinneandelen blir enda mindre jo høyere kravene til deltagelse settes, men dette er samtidig ikke overraskende, fordi studier har vist at menn er generelt mer interessert i konkurranser enn kvinner og at forskjellene øker jo større graden av konkurranse blir.[72] Variasjonene i deltagermassene analysert her tilsier at dette funnet også kan gjelde i prediksjonsturneringer som FFIs.

[72] Niederle, M. og Vesterlund, L. (2011), 'Gender and Competition', *Annual Review of Economics*, 3:1, 601–630.

Minstekrav	FFI240
Minst ett spørsmål	86 % menn 38 år (snitt) 66 % med minst 4–5 års høyere utdanning 57 % bakgrunn fra forsvarssektoren 31 % forsvars- og sikkerhetspolitisk arbeidserfaring
Minst 20 % av alle spørsmål	90 % menn 40 år (snitt) 69 % med minst 4–5 års høyere utdanning 61 % bakgrunn fra forsvarssektoren 31 % forsvars- og sikkerhetspolitisk arbeidserfaring
Minst 50 % av alle spørsmål	92 % menn 41 år (snitt) 67 % med minst 4–5 års høyere utdanning 63 % bakgrunn fra forsvarssektoren 31 % forsvars- og sikkerhetspolitisk arbeidserfaring

Tabell 3.8 Variasjoner i deltagerne i FFIs turnering ved ulike minstekrav.

3.3 Prediksjoner

Som i EPJ og GJP måtte deltagerne i FFIs turnering oppgi et sannsynlighetsestimat for hvert svaralternativ de fikk. Det vil si at på et spørsmål med tre utfall, måtte deltagerne oppgi tre sannsynlighetsestimater, f.eks. A: 20 %, B: 30 % og C: 50 %. På binære spørsmål, der deltagerne bare bedt om å oppgi sannsynligheten for at hendelsen ville skje (f.eks. 80 %), ble sannsynligheten for at hendelsen *ikke* ville skje (20 %) beregnet automatisk. Binære spørsmål innebar derfor alltid av to sannsynlighetsestimater, mens kategoriske og ordinale av minst tre.

Det var imidlertid to forskjeller i hvordan FFIs og GJPs turneringer ble gjennomført, som gjør at det ikke er mulig med en direkte sammenligning av prediksjonene som ble samlet inn:

1) Deltagerne i FFIs turnering *kunne bare predikere én gang per spørsmål*. Det vil si at svar på spørsmål med tre svaralternativer, alltid gav tre sannsynlighetsestimater. I GJP kunne deltagerne derimot oppdatere svarene sine underveis. Dette gjorde at de *kunne predikere flere ganger per spørsmål*, noe de også ble oppfordret til å gjøre. For eksempel kunne en deltager justere estimatet for A opp fra 20 % til 50 % og for C ned fra 50 % til 20 %, mens han lot B ble stående på 50 %. For hver justering ble disse nye sannsynlighetsestimatene også registrert i GJPs datasett.

2) Deltagerne i FFIs turnering *kunne bare predikere i løpet av den første uken* etter at spørsmålet ble publisert. I løpet av denne uken kunne deltagerne endre prediksjonene sine, men det var bare det siste sannsynlighetsestimatet som ble stående. Det gjorde at

alle prediksjonene i FFIs turnering ble samlet helt i starten av spørsmålsperioden. I GJP kunne deltagerne derimot *velge selv når de predikerte*. Deltagerne kunne vente med å predikere for første gang og oppdatere tidligere prediksjoner gjennom hele spørsmålsperioden, som antagelig gjorde det lettere å treffe enn om de måtte predikere helt i starten. Dette er en viktig forskjell for sammenligninger av treffsikkerheten på tvers av studiene.

Siden disse forskjellene har betydning for sammenligninger av treffsikkerheten til deltagerne på tvers av studiene, diskuteres det her hvordan prediksjonene kan analyseres på likest mulig måte.

3.3.1 Antall

For det første avhenger antallet prediksjoner av hvilket minstekrav til deltagelse som brukes og dermed hvor mange deltagere som er med. Tabell 3.9 viser derfor det totale antallet prediksjoner som datagrunnlagene til FFIs turnering, GJP og EPJ består av, gitt de forskjellige minstekravene analysert i forrige delkapittel. Antall prediksjoner og deltagere som hver studie selv baserte seg på, er uthevet i fet skrift. Hvert sannsynlighetsestimat teller som en egen prediksjon.

Minstekrav	**FFI240**	**GJP200**	**GJP350**	**EPJ140**
Minst 15 % av alle spørsmål (GJP200)	440 675 (925)	**424 259 (801)**	989 585 (1 236)	-
Minst 20 % av spørsmål ett år (GJP350)	454 854 (1 087)	423 349 (790)	**1 070 651 (1 751)**	-
Minst 20 % av alle spørsmål (FFI)	**431 382 (857)**	421 307 (770)	941 386 (1 040)	-
Minst 50 % av alle spørsmål (EPJ)	343 482 (534)	355 199 (508)	559 353 (345)	**82 361 (284)**

Tabell 3.9 Antall prediksjoner i FFIs, GJPs og EPJs datagrunnlag ved ulike minstekrav. Antall deltagere med prediksjoner som oppfyller minstekravene i parentes.

I FFIs turnering ble det totalt samlet inn 464 342 prediksjoner fra alle 1375 deltagerne som svarte på minst ett spørsmål.[73] Samtidig viser tabell 3.9 at de forskjellige minstekravene har relativt lite å si for hvor mange prediksjoner FFIs datagrunnlaget består av, fordi den mest aktive halvparten av deltagerne i turneringen stod for det store flertallet av prediksjoner som ble samlet inn. De 857 deltagerne som oppfylte FFIs minstekrav på 20 % stod nemlig for hele 431 382 (93 %) av alle prediksjonene i hele turneringen. Den synkende svarprosenten utover i turneringen, som ble vist i figur 3.1, skjuler derfor at selv om antallet registrerte økte var det stort sett de samme personene som deltok og stod for prediksjonene gjennom hele turneringen.

Tabell 3.9 viser også at antall prediksjoner i FFI240 og GJP200 er svært like, uansett krav til deltagelse. FFI240 og GJP200 utgjør imidlertid bare halvparten av prediksjonene i GJP350.

[73] På ett av spørsmålene var det feil i svarkriteriene, som medførte endringer i spørsmålsteksten underveis i spørsmålsperioden. Her ble derfor alle prediksjonene (1 375) registrert før feilen ble rettet, slettet fra datagrunnlaget.

Antall prediksjoner i GJP350 faller heller ikke mye ved minstekravene til FFI240 og GJP200, selv om antallene deltagere blir mye mindre. Når minstekravet økes til 50 % av alle spørsmål reduseres imidlertid antallet prediksjoner i GJP350 betraktelig. Forklaringen er antagelig at en stor andel av deltagerne først ble med det tredje året, som betyr at det ikke hadde mulighet til å svare på minst halvparten av alle spørsmålene. Merk dog at tallene fra begge GJP-studiene inkluderer alle prediksjoner på samme spørsmål. FFIs tall inkluderer derimot bare én prediksjon per svaralternativ, siden deltagerne bare kunne predikere én gang.

Både FFIs og GJPs turneringer består av langt flere prediksjoner enn EPJ, som hadde det strengeste minstekravet, men der deltagerne i likhet med FFIs turnering bare predikerte én gang. Siden hver av de 284 ekspertene ble bedt å besvare rundt 140 spørsmål med tre mulige utfall hver, kunne prosjektet i teorien bestå av i overkant av 100 000 prediksjoner. Siden ikke alle ekspertene svarte på alle spørsmål, bestod det reelle datagrunnlaget til EPJ av 82 361 prediksjoner.[74]

3.3.2 Definisjoner

For det andre avhenger størrelsen på datagrunnlaget av hvordan en prediksjon defineres. Tabell 3.10 viser derfor hvor mange prediksjoner FFIs og GJPs datagrunnlag består av, dersom prediksjonene telles på forskjellige måter. Analysene her tar utgangspunktet i de respektive studienes egne minstekrav. EPJ er ikke analysert her, siden dette datasettet ikke er tilgjengelig.

Definisjon	**FFI240**	**GJP200**	**GJP350**
Antall sannsynlighetsestimater på alle svaralternativer, inkludert oppdateringer	431 382 (857)	424 259 (801)	1 070 651 (1 751)
Antall sannsynlighetsestimater på alle svaralternativer, uten oppdateringer	431 382 (857)	236 145 (801)	432 497 (1 751)
Antall ganger deltagerne predikerte, inkludert oppdateringer	124 628 (857)	185 466 (801)	452 846 (1 751)
Antall ganger deltagerne predikerte, uten oppdateringer	124 628 (857)	104 270 (801)	189 235 (1 751)
Antall sannsynlighetsestimater på alle svaralternativer, uten oppdateringer, men bare første uken	431 382 (857)	108 752 (797)	206 847 (1 723)
Antall ganger deltagerne predikerte, uten oppdateringer, men bare første uken	124 628 (857)	48 065 (797)	90 401 (1 723)

Tabell 3.10 Antall prediksjoner FFIs GJPs datagrunnlag ved ulike definisjoner. Antall deltagere med prediksjoner som oppfyller definisjonen i parentes.

[74] Tetlock (2005), *Expert Political Judgment*, s. 246.

Den første raden representerer den bredeste definisjonen. Her teller hvert sannsynlighetsestimat som en egen prediksjon. Det vil si at antallet inkluderer alle estimater på alle svaralternativer, inkludert oppdateringer. Det er denne definisjonen som er brukt i beregningen av antall prediksjoner i forrige underkapittel. Antallene prediksjoner i den første raden er derfor identisk med dem som i tabell 3.9 er uthevet med fet skrift og som studiene baserte sine egne analyser på.

Den andre raden viser også antall sannsynlighetsestimater på alle svaralternativer, men her er eventuelle oppdateringer ikke med. Det betyr at det er like mange prediksjoner per spørsmål som det er svaralternativer. Siden det i FFIs turnering ikke var mulig å oppdatere prediksjonene underveis, er antallet her det samme som i første rad. Resultatet av å utelate eventuelle oppdateringer av de første prediksjonene på et spørsmål, gjør imidlertid at antallet prediksjoner i GJP200 reduseres fra 424 259 til 236 145.[75] Det betyr at 188 114 (44 %) av prediksjonene i dette datagrunnlaget bare var oppdateringer av tidligere estimater. I GJP350 faller antallet prediksjoner enda mer når oppdaterte sannsynlighetsestimater fjernes, fra 1 070 651 til 432 497.[76] Det betyr at 638 154 (60 %) av prediksjonene i dette datagrunnlaget var oppdateringer.

Uten oppdateringer består FFI240 av langt flere prediksjoner enn GJP200 og like mange som GJP350, til tross for at GJP350 hadde dobbelt så mange deltagere og rundt 100 flere spørsmål. Forklaringen er at deltagerne i FFIs turnering svarte i gjennomsnitt på flere spørsmål enn i GJP350 (145 mot 108) og spørsmålene hadde i snitt flere svaralternativer (3,5 mot 2,4).

I GJP200-artikkelen oppgis det imidlertid at studien baserte seg på «over 150 000 prediksjoner» fra 743 deltagere på 199 spørsmål, som er et enda lavere antall enn det som er beskrevet over. Forklaringen er at definisjonen som artikkelen legger til grunn er antall *ganger* deltagerne predikerte. I GJP ble det nemlig beregnet en egen score basert på hvert sett med sannsynlighetsestimater deltagerne registrerte på et spørsmål. Hvis deltagerne predikerte flere ganger på samme spørsmål, ble treffsikkerheten på det aktuelle spørsmålet basert på snittet av alle disse scorene (se delkapittel 4.1 om hvordan scorene ble beregnet).

Den tredje raden viser derfor antall ganger deltagerne predikerte, inkludert oppdateringer. Det betyr at hvis en deltager predikerte tre ganger på samme spørsmål, teller dette som tre prediksjoner, uavhengig av antall sannsynlighetsestimater. Basert på replikasjonsdatasettet til GJP200, som inkluderer litt flere deltagere og spørsmål enn artikkelen sitert over, predikerte deltagerne 185 466 ganger til sammen, altså bare litt høyere enn antallet prediksjoner nevnt i artikkelen.

[75] Totalt finnes det 440 657 prediksjoner i det fullstendige datasettet fra deltagerne og spørsmålene som er inkludert i GJP200s replikasjonsdatasett. Dette inkluderer imidlertid hver enkelt oppdatering, selv om de ble gjort samme dag. Siden det bare var den sist registrerte prediksjonen per dag som ble brukt til å beregne treffsikkerheten, er alle prediksjoner utenom den siste fjernet. Dette reduserer det totale antallet prediksjoner med 15 086. I tillegg er det fjernet 1312 prediksjoner på spørsmål som var med i GJP200, men der deltagerne manglet en score i replikasjonsdatasettet.
[76] Totalt finnes det 1 154 864 prediksjoner i det fullstendige datasettet fra deltagerne og spørsmålene som er inkludert i GJP350s replikasjonsdatasett. Dette inkluderer imidlertid hver enkelt oppdatering, selv om de ble gjort samme dag. Siden det bare var den sist registrert prediksjonen per dag som ble brukt til å beregne treffsikkerheten, er alle prediksjoner utenom den siste fjernet. Dette reduserer det totale antallet prediksjoner med 59 765. I tillegg er det fjernet 24 448 prediksjoner på spørsmål som var med i GJP200, men der deltagerne manglet en score i replikasjonsdatasettet.

GJP350-artikkelen nevner ikke hvor mange prediksjoner studien er basert på, men ved samme fremgangsmåte er det registrert til sammen 452 846 prediksjoner.[77]

Den fjerde raden viser også antallet ganger deltagerne predikerte, men her uten oppdateringer. Denne definisjonen tilsvarer med andre ord antallet ganger deltagerne svarte på et spørsmål. Uten oppdateringene reduseres antallet ganger deltagerne i GJP200 predikerte fra 185 466 til 104 270. Det betyr at 81 196 (44 %) av gangene deltagerne predikerte var dette oppdateringer av tidligere sannsynlighetsestimater. Tilsvarende faller antallet prediksjoner i GJP350 fra 452 846 til 189 235. Det betyr at 263 611 (58 %) av gangene deltagerne i GJP350 predikerte var dette bare justeringer av eksisterende prediksjoner.

Siden deltagerne i FFIs turnering ikke kunne oppdatere sannsynlighetsestimatene sine underveis, er antallene ganger deltagerne predikerte de samme i tredje og fjerde rad (124 628). Sammenlignet med GJP200, utgjør dette et lavere antall prediksjoner i FFI240. Samtidig skyldes dette den høye andelen prediksjoner som var oppdateringer i GJP. Uten oppdateringene predikerte FFIs deltagere flere ganger enn i GJP200, antageligvis fordi FFI240 bestod av litt flere spørsmål. Deltagerne i GJP350 predikerte flere ganger enn FFIs, uavhengig av oppdateringene, som antageligvis skyldes at GJP350 bestod av både flere deltagere og spørsmål enn FFI240.

Et problem med alle definisjonene over er at de ikke tar hensyn til *når* deltagerne predikerte. Mens deltagerne i FFIs turnering bare kunne predikere i løpet av den første uken etter at spørsmålet ble publisert, kunne deltagerne i GJP velge fritt når de predikerte helt til spørsmålet ble avgjort. Det er imidlertid grunn til å anta at det blir lettere å predikere riktig mot slutten av spørsmålsperioden enn i starten, siden usikkerheten normalt reduseres jo nærmere i tid en kommer tidspunktet en hendelse skal skje (f.eks. et valg). For å kunne sammenligne treffsikkerheten til deltagerne på likest grunnlag bør det derfor tas utgangspunkt i samme prediksjonsperiode.

Den femte raden viser antall sannsynlighetsestimater på alle svaralternativer som deltagerne leverte *i løpet av den første uken* etter at spørsmålet ble publisert, slik som i FFIs turnering, mens den sjette raden viser antall ganger deltagerne predikerte innenfor samme tidsperiode. Siden alle prediksjoner i FFIs turnering ble levert den første uken, holder antallet prediksjoner seg uendret fra radene over. I GJP200 og GJP350 reduseres derimot antallet prediksjoner betydelig når de baseres på samme prediksjonstidspunkt som i FFI240.

I GJP200 faller antallet sannsynlighetsestimater fra 424 259 til 108 752 og antallet ganger deltagerne predikerte fra 185 466 til 48 065. GJP200s datagrunnlag reduseres dermed til en firedel. Antallet deltagere holder seg derimot tilnærmet uendret, som betyr at de fleste deltagerne som oppdaterte sine sannsynlighetsestimater senere, også predikerte den første uken. I GJP350 reduseres også antallet prediksjoner til en firedel av det opprinnelige. Antall sannsynlighetsestimater faller fra 1 070 651 til 206 847 og antall ganger deltagerne predikerte fra 452 846 til 90 401.

[77] At antall prediksjoner i GJP er basert på antall ganger deltagerne predikerte bekreftes av at en annen GJP-artikkel, som inkluderte 2860 deltagere som svarte på minst ett av 344 spørsmål fra de tre første årene, bygget på 494 552 prediksjoner. Dette er litt høyere enn i replikasjonsdatasettet til GJP350, der minstekravet var litt strengere ved at deltagerne måtte ha besvart minst 25 spørsmål ett av årene. Se Moore mfl. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', s. 3555.

Også i GJP350 holder antallet deltagere seg stabilt. Dette betyr at deltagerne som kan sammenlignes med FFIs basert på samme prediksjonstidspunkt, er akkurat de samme deltagerne som funnene i GJP200s og GJP350s artikler er basert på. Basert på samme prediksjonstidspunkt består FFIs datagrunnlag imidlertid av fire ganger så mange sannsynlighetsestimater som i GJP200 og dobbelt så mange som i GJP350. I tillegg blir antallet scores som treffsikkerheten kan beregnes ut fra også høyere i FFIs turnering enn begge GJP-studiene.

Således gir FFIs datagrunnlag et bedre utgangspunkt for å kunne si noe om hvor godt personer treffer når de blir stilt et spørsmål og bedt om å predikere dette, uten at treffsikkerheten deres blir påvirket av at de kan endre svaret sitt underveis. Hvis treffsikkerheten til deltagerne i GJP blir dårligere når den baseres på prediksjoner fra den første uken, tilsier dette at tidspunktet deltagerne predikerer har betydning for funnene som er beskrevet i GJPs artikler.

4 Variabler

Dette kapittelet beskriver først hvordan treffsikkerheten, som er denne rapportens avhengige variabel, har blitt målt i FFIs prediksjonsturnering. Deretter beskrives det hvordan alle de individuelle variasjonene, som utgjør rapportens uavhengige variabler, har blitt målt og hvordan deltagerne i FFIs turnering har scoret på disse sammenlignet med deltagerne i EPJ og GJP.

4.1 Treffsikkerhet

Treffsikkerheten i FFIs prediksjonsturnering ble målt ved hjelp av Brier-score. Dette er det samme scoringssystemet som ble brukt i EPJ og GJP. Brier-scoren måler evnen til å oppgi *høye* sannsynligheter til hendelser som *faktisk* skjer, og *lave* sannsynligheter til de som *ikke* gjør det, uavhengig av hvor ofte de skjer statistisk sett.[78] Dette målet på treffsikkerhet er spesielt relevant i forsvars- og sikkerhetspolitisk sammenheng, der hendelser er relativt unike og de mest nyttige prediksjonene er dem som kan fortelle oss at «dette vil skje» og «dette vil ikke skje».

4.1.1 Brier-score

Brier-systemet ble opprinnelig utviklet av Glenn W. Brier i 1950 for å måle treffsikkerheten til værmeldinger.[79] I dag er det et av de vanligste målene på treffsikkerheten til probabilistiske prediksjoner. Her vurderes ikke prediksjoner ut fra *om* de treffer, men hvor *sannsynlig* (i prosent) det riktige utfallet ble estimert å være. Skalaen går fra 0 til 2, der lavere score betyr høyere treffsikkerhet. Du får en Brier-score på 0 hvis du predikerer «helt riktig», som vil si at du hevder en hendelse er 100 % sannsynlig, og den faktisk skjer. Du får en Brier-score på 2 hvis du predikerer «helt feil», som vil si at du hevder en hendelse er 100 % sannsynlig, men den *ikke* skjer.

Den matematiske formelen som ble brukt til å beregne Brier-score i FFIs turnering var:[80]

$$Brier\ score = \sum_{i=1}^{R} (f_i - o_i)^2$$

Her er R antall svaralternativer (f.eks. to ved et binært spørsmål), i er et svaralternativ (f.eks. at hendelsen skjer og ikke skjer), f_i er en sannsynlighetsvurdering for svaralternativ i (f.eks. 60 % for at hendelsen skjer, 40 % for hendelsen ikke skjer), og o_i er utfallet av svaralternativ i (som er 1, hvis svaralternativet som predikeres er riktig, eller 0, hvis svaralternativet er feil). I beregningene under er 1-tallet uthevet med fet skrift for å vise hvilket svaralternativ som er riktig.

[78] Dette kalles *resolution*, som er én av tre komponenter som Brier-scoren kan brytes ned i. De to andre er *variation*, som måler usikkerheten ved det som predikeres, basert på hvor ofte det samme utfallet skjer (*base rate*), f.eks. om bestemt type hendelse skjer 50 % eller 100 % av gangen, og *calibration*, som sammenligner gjennomsnittlig sannsynlighetsestimat med gjennomsnittlig treffprosent (*hit rate*), f.eks. at prediksjoner på 70 % treffer 70 % av gangene.
[79] Brier, G. W. (1950), 'Verification of Forecasts Expressed in Terms of Probability', *Monthly Weather Review*, 78:1.
[80] FFIs turnering bruker den opprinnelige Brier-formelen med skala fra 0 til 2. Det finnes en nyere og mer utbredt formel, der skalaen går fra 0 til 1, men denne er bare anvendelig på binære spørsmål og kan derfor ikke brukes her.

4.1.2 Binære spørsmål

Den enkleste beregningen av Brier-score er ved binære spørsmål (ja/nei). Ta for eksempel spørsmålet: «Vil Putin stille i det russiske presidentvalget i 2024?». La oss si at du anslår at det er 60 % sannsynlig at han vil stille. Det betyr samtidig at du mener det er 40 % sannsynlig at han *ikke* stiller. Hvis Putin stiller til valg i 2024, vil du få en Brier-score på 0,32:

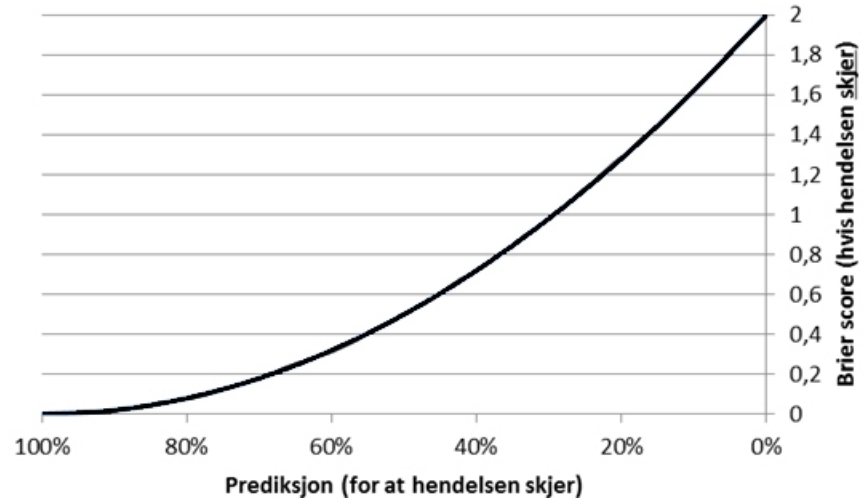
$$Brier\ score = (0{,}6 - \mathbf{1})^2 + (0{,}4 - 0)^2 = 0{,}16 + 0{,}16 = 0{,}32$$

Hvis Putin *ikke* stiller til valg, vil du derimot få en høyere Brier-score på 0,72:

$$Brier\ score = (0{,}6 - 0)^2 + (0{,}4 - \mathbf{1})^2 = 0{,}36 + 0{,}36 = 0{,}72$$

Brier-scoren er et *objektivt* mål på treffsikkerheten din, basert på avstanden mellom det du predikerte og det som virkelig skjedde. Den måler med andre ord hvor langt unna du er fra å predikere helt riktig, som forklarer hvorfor en lavere score betyr høyere treffsikkerhet.

For å si noe om hvor god en Brier-score er, er det vanlig å sammenligne scoren en fikk med scoren en ville fått ved tilnærminger en forventer å slå, som tilfeldig gjetning. Ved gjetning vil prediksjonene fordele seg likt på alle svaralternativer hvis en gjentar øvelsen mange nok ganger, f.eks. 50/50 % på mynt og kron. I EPJ ble «en pilkastende ape med bind for øynene» brukt som metafor for denne tilnærmingen. I eksempelet over ville apen ha oppgitt 50/50 % sannsynlighet for at Putin stiller til valg og ikke, som ville gitt en Brier-score på 0,5, uansett utfall.



Figur 4.1 Brier-scores ved ulike estimater, gitt at hendelsen skjedde.

Figur 4.1 illustrerer hvilken Brier-score du ville fått ved ulike sannsynlighetsestimater på et binært spørsmål, gitt at hendelsen det spørres om faktisk skjer. Som figuren viser, gjør Brier-formelen at prediksjoner som havner på feil side av 50 %, straffes hardere enn den premierer prediksjoner på riktig side.

Hvis du hadde oppjustert prediksjonen om at Putin stiller til valg til 80 %, og han gjorde det, vil du fått en mye bedre Brier-score på 0,08. Forskjellen mellom 80 % (deg) og 50 % (apen) vil da være 0,42 i Brier-score. Hvis Putin derimot ikke stilte likevel, vil det si at du bare hadde satt 20 % på det riktig svar. Dette ville gitt en mye høyere Brier-score på 1,28. Her blir avstanden ned til apens 50 %-prediksjon nesten dobbelt så stor (0,78), selv om forskjellene i prosenter er lik. Disse forskjellene blir større jo lenger ut i endene av sannsynlighetsskalaen en befinner seg.

4.1.3 Kategoriske spørsmål

Fremgangsmåten for å beregne Brier-scoren på kategoriske spørsmål, er den samme som ved binære, bare at beregningen utvides med like mange ledd som det er svaralternativer.

Ta for eksempel spørsmålet: «Fra hvilket parti vil forsvarsministeren komme fra etter stortingsvalget i 2017?». Du får fem svaralternativer og fordeler sannsynligheten slik: A: Arbeiderpartiet (20 %), B: Høyre (60 %), C: Senterpartiet (15 %), D: Fremskrittspartiet (5 %), E: Ingen av de nevnte (0 %). Hvis riktig svar ble B (60 %), ville beregningen sett slik ut:

$$Brier\ score = A{:}\,(0{,}2 - \ 0)^2 + \ B{:}\,(0{,}6 - \ \mathbf{1})^2 + \ C{:}\,(0{,}15 - \ 0)^2 + \ D{:}\,(0{,}05 - \ 0)^2 + \ E{:}\,(0{,}00 - \ 0)^2 = \ 0{,}225$$

Dette gir altså en betydelig lavere Brier-score (0,225) enn på det binære spørsmålet (0,32), selv om prediksjonen din på det riktige svaret var 60 % på begge spørsmålstyper. Forklaringen er at de resterende 40 % ble fordelt på flere svaralternativer (A: 20 %, C: 15 % og D: 5 %), som til sammen gir en mindre «straff» i form av høy Brier-score, enn på det binære spørsmålet, der alle 40 % ble satt på det gale svaret.

4.1.4 Ordinale spørsmål

Ved ordinale spørsmål, der noen svar er riktigere enn andre, er beregningen litt mer komplisert.

Ta for eksempel spørsmålet: «Hvor mange vil bli drept i islamistiske terrorangrep i Europa i 2017?». Du får fem svaralternativer og fordeler sannsynlighet på samme måte som i eksempelet over: A: 0–49 drepte (20 %), B: 50–99 drepte (60 %), C: 100–149 drepte (15 %), D: 150–199 drepte (5 %) og E: minst 200 drepte (0 %). La oss si at riktig svar ble B (60 %). Her tas det imidlertid hensyn til hvordan de andre prosentene fordelte seg i forhold til det riktige svaret.

Svaralternativene deles derfor inn i par, og det beregnes først en egen Brier-score for hvert par, basert på hvilket par som inkluderer riktige svaralternativ. Den endelige Brier-scoren baseres på snittet av scorene på alle par, i stedet for å summere Brier-scorene fra alle leddene, som ved binære og kategoriske spørsmål. I dette eksemplet ville du fått en Brier-score på bare 0,04, som er mye lavere enn i eksemplene over (0,32 på det binære og 0,225 på det kategoriske), selv om prediksjonene på riktig svar var 60 % på alle tre spørsmålstypene.

$$Brier\ score = \frac{\begin{pmatrix} A\text{: } (0{,}2-0)^2 + \ BCDE\text{: } \left((0{,}6+0{,}15+0{,}05+0)-\ \mathbf{1}\right)^2 + \\ \text{AB: } \left((0{,}2+0{,}6)-\mathbf{1}\right)^2 + CDE\text{: } \left((0{,}15+0{,}05+0)-\ 0\right)^2 + \\ ABC\text{: } \left((0{,}2+0{,}6+0{,}15)-\mathbf{1}\right)^2 + \ DE\text{: } \left((0{,}05+0)-\ 0\right)^2 + \\ ABCD\text{: } \left((0{,}2+0{,}6+0{,}15+0{,}05)-\mathbf{1}\right)^2 + \text{ E: } (0-\ 0)^2) \end{pmatrix}}{4}$$

$$= \frac{0{,}04+0{,}04+0{,}04+0{,}04+0{,}0025+0{,}0025+0+0}{4} = 0{,}04125$$

Denne måten å beregne treffsikkerheten på gjør derfor at ordinale spørsmål nærmest alltid gir en lavere Brier-score enn binære og kategoriske, selv om prediksjonene på riktig svar er de samme. Dersom det riktige svaret på de kategoriske og ordinale eksempelspørsmålene hadde blitt E i stedet for B, der prediksjonene var 0 % i begge tilfeller, hadde den ordinale beregningen fremdeles gitt en Brier-score som var litt lavere (1,29) enn den kategoriske beregningen (1,43).

Hvordan Brier-scorene beregnes, har derfor betydning for analysen av objektiv treffsikkerhet, og det er viktig når Brier-scores skal sammenlignes på tvers av forskjellige typer spørsmål.

4.1.5 Standardisert Brier-score

En annen utfordring ved Brier-systemet er at det ikke tar høyde for at noen spørsmål kan være vanskeligere å predikere enn andre. Spørsmål preget av stor usikkerhet vil generelt føre til en høyere Brier-score, fordi det blir vanskeligere å treffe. Det er for eksempel vanskeligere å forutsi morgendagens vær i Bergen enn på Mallorca. Det er dermed en fare for at noen deltagere treffer bedre enn andre fordi de velger å svare bare på de letteste spørsmålene.

For å kontrollere for variasjoner i spørsmålenes vanskelighetsgrad ble derfor alle Brier-scorene i FFIs turnering standardisert (dvs. konvertert til z-score), slik som i GJP. Denne standardiserte Brier-scoren er et *relativt* mål på treffsikkerheten, altså hvor mye bedre eller dårligere en deltager traff sammenlignet med de andre som svarte på samme spørsmål. Deltagernes sammenlagte treffsikkerhet gjennom hele turneringen er således basert på snittet av alle standardiserte Brier-scores på alle spørsmålene de svarte på. På spørsmål deltagerne ikke svarte, gis det ingen score, som gjør at de hverken premieres eller straffes ut fra antallet spørsmål de svarte på.

Dette er ikke den samme måten den sammenlagte treffsikkerheten ble beregnet på underveis i FFIs turnering. Her ble deltagernes sammenlagte score og plassering basert på en Accuracy-score, som ble beregnet ved å subtrahere medianen av alle deltagernes Brier-scores på det aktuelle spørsmålet fra hver enkelt deltagers Brier-score. Dette representerer bare en annen måte å beregne relativ treffsikkerhet på og er den samme brukt i *GJ Open*, som er en kommersiell prediksjonsturnering i regi av forskerne bak GJP.[81] De sammenlagte resultatene som ble brukt til å kåre vinnere baserte seg også på summen, ikke snittet, av deltagernes Accuracy-scores. Dette var et bevisst valg, fordi en summert score gir et større utslag fra spørsmål til spørsmål, og dermed kunne oppleves som mer motiverende for deltagerne, enn en snittscore, der forskjellene

[81] Se 'Measuring Accuracy in Prediction Markets and Opinion Poll/Pools', *Cultivate Labs*.

ville bli mindre for hvert spørsmål som ble avgjort. Ved slutten av hvert år fikk deltagerne uansett vite plassering og sammenlagtscore basert på både summen, snittet og medianen.

Som i GJP er deltagernes standardiserte Brier-scores beregnet ut fra prediksjonene til *alle* deltagere som svarte på det aktuelle spørsmålet, uavhengig av om de andre oppfylte minstekravet til deltagelse som i turneringen som helhet. Det er ikke forklart hvorfor de standardiserte Brier-scorene er beregnet slik i GJP, men en fordel er at deltagernes relative treffsikkerhet i turneringen alltid er den samme, uavhengig av hvilket utvalg som analyseres. For å kunne sammenligne resultatene på likest mulig grunnlag er de standardiserte Brier-scorene beregnet på samme måte i FFIs turnering.

I GJP200 har det å basere de standardiserte Brier-scores på alle deltagerne som svarte på det enkelte spørsmål ingen konsekvenser, siden det bare er deltagere som oppfylte minstekravet som er med i replikasjonsdatasettet. Replikasjonsdatasettet til GJP350 inneholder derimot mange flere deltagere enn dem som oppfylte minstekravet om å ha svart på 25 spørsmål minst ett av årene. Totalt består GJP350s replikasjonsdatasett av 2389 deltagere, selv om det bare er 1751 av dem som oppfylte minstekravet. Det er likevel prediksjonene fra alle 2389 deltagerne som er brukt til å beregne de standardiserte Brier-scorene til de 1751 deltagerne som er analysert i studien. Samtidig er ingen av de 2389 deltagerne i GJP350 registrert med færre enn 10 spørsmål i løpet ett turneringsår, som tydeligvis utgjør et ytterligere minstekrav for å bli inkludert i replikasjonsdatasettet. Det er det heller ingen deltagere GJP200 som har svart på mindre enn 10 spørsmål per turneringsår. Siden det i FFIs turnering ikke var et klart skille mellom turneringsårene, er det i stedet satt et minstekrav om å ha svart på minst 10 % av spørsmålene i turneringen som helhet for å bli inkludert i beregningene av standardiserte Brier-scores. FFIs minstekrav er satt til 10 % fordi dette tilsvarer omtrent den andelen som 10 spørsmål per år utgjorde i GJP350.[82]

Hvorvidt den standardiserte Brier-scoren på hvert enkelt spørsmål baseres på alle deltagere som oppfylte dette minstekravet om å ha svart på minst 10 % av spørsmålene eller på bare deltagerne som oppfylte minstekravet om å ha svart på minst 20 % har imidlertid lite å si for hvem som traff best. Basert på de foreløpige resultatene beskrevet i kapittel 5, er 99 av de 100 beste deltagerne de samme personene ved begge disse måtene å beregne standardisert Brier-score på.

[82] 10 spørsmål utgjør 10 %, 9 % og 7 % av de hhv. 102, 109 og 136 spørsmålene som ble stilt de tre årene av GJP350.

4.2 Individuelle variasjoner

Dette delkapittelet beskriver hvordan hver uavhengige variabel har blitt målt i FFIs turnering.

Ambisjonen har vært å måle alle de samme variablene som har blitt undersøkt i tidligere forskningsprosjekter for å kunne etterprøve flest mulig av de eksisterende funnene. For å kunne gjøre direkte sammenligninger av deltagerne i FFIs turnering, EPJ og GJP har ambisjonen også vært å bruke akkurat de samme psykologiske testene av individuelle evner og tenkemåter. Alle testene som har blitt brukt til å måle de uavhengige variablene i FFIs turnering, er gjengitt i vedlegg A.

De eneste variablene som ikke er målt i FFIs turnering er de situasjonelle variablene knyttet til GJPs eksperimentering med tiltak for å forbedre treffsikkerheten (opplæring og gruppearbeid), siden det ikke ble gjort tilsvarende forsøk i FFIs turnering. Fra og med det tredje året av GJP ble også noen av testene deres endret for å adressere svakheter i de opprinnelige som ble brukt. FFIs turnering har derfor bare benyttet de nyeste og mest pålitelige testversjonene.

De uavhengige variablene som har blitt målt i FFIs turnering kan grupperes i fire kategorier:

1) *Ekspertisevariabler* – personers formelle kvalifikasjoner (utdanningsnivå, relevant erfaring, spesifikk kompetanse og bruk i media).
2) *Disposisjonelle variabler* – de forskjellige forutsetningene personer har for å predikere (kognitive evner, tenkemåter, kunnskapsnivå og oppgavespesifikke ferdigheter).
3) *Innsatsvariabler* – som handler om hvor stor innsats deltagerne la ned i turneringsdeltagelsen (antall spørsmål besvart og tid brukt per spørsmål).
4) *Prediksjonsspesifikke tenkemåter* – som handler om hvilke spesifikke måter å tenke på som deltagerne benyttet seg av når de fordelte sannsynlighetene sine i turneringen.

Ekspertisevariablene har allerede blitt analysert i kapittel 3. Her analyseres derfor bare disposisjonelle variabler, deltagernes innsats og hvordan de tenkte når de predikerte i turneringen. De fleste variablene vil brukes til å etterprøve de bivariate korrelasjonene som ble analysert i GJP200-artikkelen om sammenhenger mellom individuelle variasjoner og treffsikkerheten generelt og i GJP350-artikkelen om hva som kjennetegner de aller beste (superforecasterne).

Tabell 4.1 sammenligner derfor verdiene i FFIs og GJPs turneringer på alle variablene som er relevante for de bivariate korrelasjonsanalysene. Her oppgis den gjennomsnittlige scoren, standardavviket (SD), maksimums- og minimumsverdiene og antallet deltagere (n) som verdiene er baserte på. Deltagerne som sammenlignes her er de samme som ble analysert i kapittel 3.

I denne rapporten er alle verdiene fra GJP beregnet på nytt, basert på deltagerne i hvert replikasjonsdatasett sine scores på alle individuelle variasjoner dokumentert i det fullstendige datasettet. Dette er gjort for å kunne reanalysere alle korrelasjonene med treffsikkerhet og for å kunne gå dypere inn i funnene fra GJP når resultatene sammenlignes med FFIs. Med mindre noe annet påpekes er det ingen vesentlige avvik mellom verdiene oppgitt her og i GJPs artikler. De små forskjellene som finnes skyldes antageligvis at datasettene inkluderer noen flere deltagere og

spørsmål enn artiklene. Under GJP200 rapporteres scorene på alle variabler brukt i artikkelen om systematiske sammenhenger basert på de to første årene, mens under GJP350 rapporteres scorene på relevante variabler brukt i superforecaster-artikkelen, inkludert scores på de nye testene som først ble innført det tredje året. Scorene på tester som bare ble gjennomført de to første årene, men som likevel er med i GJP350-artikkelens analyser, rapporteres under begge datasettene, men er basert på studienes hhv. 801 og 1751 deltagere.

En forutsetning for at funnene fra FFIs turnering skal være valide og reliable er at deltagerne ikke svarte helt vilkårlig. Dette gjelder både sannsynlighetsestimatene deltagerne oppgav når de ble bedt om å predikere og svarene deres på psykologiske tester og spørreundersøkelser som har blitt brukt til å måle de uavhengige variablene. Validiteten handler om i hvilken grad resultatene kan brukes til å trekke slutninger om det vi undersøker, som her er prediksjonsevne og individuelle forskjeller. Reliabiliteten handler om konsistensen i målingene av hver variabel, for eksempel mellom forskjellige påstander i en test som er ment å skulle måle den samme egenskapen.

I litteraturen er det bred enighet om hvilke kognitive prosesser som må være involvert for at spørsmål skal besvares optimalt og for at resultatene skal være valide og reliable.[83] Ideelt må respondentene: 1) forstå spørsmålet, 2) søke i hukommelsen etter relevant informasjon, 3) sammenstille informasjonen til én vurdering, og 4) oversette denne vurderingen til et svar basert på alternativene i undersøkelsen. I den grad respondenter går grundig gjennom hvert av disse stegene, driver de med *optimizing*. I virkeligheten er det imidlertid ofte slik at respondenter ikke optimaliserer, f.eks. fordi blir lei av å svare. I stedet for å gi de riktigst mulige svarene, kan de da være fornøyd med å gi tilfredsstillende svar, f.eks. ved å gå for det første «akseptable» svaret de kommer på. Dette kalles svak *satisfying*. I verste fall gjør ikke respondentene noen vurderinger i det hele tatt og bare velger det første «rimelige» svaret, f.eks. basert på holdepunkter i spørsmålet som peker mot et svaralternativ som lett kan forsvares. Dette kalles sterk *satisfying*.

Optimizing og sterk *satisfying* er endepunkter i et kontinuum av hvor grundig respondenter er i hvert steg av besvarelsesprosessen. Det er særlig tre faktorer som kan øke faren for *satisfying*:

- *Respondentenes evner.* Dette handler om hvor gode personene er til å gjennomføre komplekse mentale prosesser, tidligere erfaring med å besvare temaet spørsmålet handler om og hvorvidt personen har eksisterende vurderinger av det aktuelle spørsmålet. Dette er i stor grad dekket gjennom testene av kognitive evner og kunnskapsnivå. Som resten av kapittelet vil vise, scorer deltagerne i FFIs turnering relativt høyt på alle disse.

- *Spørsmålenes vanskelighetsgrad.* I FFIs turnering var det særlig to aspekter som kan tenkes å ha gjort prediksjonsspørsmålene vanskelige å svare på: 1) hvorvidt deltagerne forstår hva som skal predikeres, og 2) hvorvidt deltagerne klarte å omgjøre vurderingene sine til prosentvise sannsynlighetsestimater. For å undersøke dette ble deltagerne bedt om å vurdere hvor vanskelig eller lett de syntes det var å forstå spørsmålene og å oppgi svar som sannsynlighetsvurderinger, basert på en skala fra 1 til 7, der 1 var

[83] Krosnick, J. A. og Presser, S. (2010), 'Question and Questionnaire Design', i Marsden, P. V. og Wright, J. D., red., *Handbook of Survey Research*, 2. utg. (Bingley: Emerald Group Publishing Limited), ss. 263–313.

«svært vanskelig» og 7 var «svært lett». Medianscorene på disse spørsmålene var hhv. 6 og 4, som tilsier at deltagerne syntes det var ganske lett å forstå spørsmålene og hverken lett eller vanskelig å oppgi svarene sine som sannsynlighetsvurderinger.[84] De fleste deltagerne hadde altså ikke problemer med å forstå spørsmålene, selv om det å treffe naturligvis kunne oppleves som vanskelig.

- *Motivasjon.* Respondenters motivasjoner for å delta i en spørreundersøkelse kan være mangefasetterte. To viktige aspekter er hvor interessant de synes undersøkelsen er og hvor nyttig de anser sin egen deltagelse for å være. Ved registreringen ble alle deltagerne bedt om å vurdere sin egen interesse for forsvars- og sikkerhetspolitiske spørsmål, basert på en skala fra 1 til 7, der 1 var «svært liten» og 7 var «svært stor». Medianscoren på dette spørsmålet var 6, som betyr at deltagerne var ganske interessert i tematikken de ble bedt om å predikere.[85] Når deltagerne over halvveis ut i turneringen ble bedt om å oppgi hvorfor de deltok, var de viktigste motivasjonene at de syntes spørsmålene var interessante og at de ønsket å se hvor godt de klarer å predikere, mens de i svært liten grad følte seg forpliktet til å delta.[86] På spørsmål om motivasjonene deres sank eller steg i løpet av turneringen svarte de fleste at den hverken sank eller steg.[87] I tillegg vurderte deltagerne sin egen deltagelse som litt nyttig både for seg selv og for forskningen.[88]

Det er altså lite som tyder på at deltagerne ikke forstod spørsmålene eller gjorde vurderinger før de registrerte sine prediksjoner, hverken i starten eller på slutten turneringen.

Når det gjelder de psykologiske testene av deltagernes evner og tenkemåter er reliabiliteten også målt statistisk ved hjelp av Cronbachs Alpha (α), slik som i GJP.[89] Dette er et mål på hvor godt ulike deler av en test, f.eks. oppgaver, måler den samme bakenforliggende variabelen, f.eks. intelligens. Dette er også kalt internkonsistens. Alpha-verdiene på hver test er også rapportert i tabell 4.1. Ved liten eller ingen konsistens går Alpha-verdien mot 0, mens ved god konsistens mot 1. En tommelfingerregel er at verdien må være minst 0,7 for at testscoren skal ha god reliabilitet, men 0,6 omtales også som akseptabelt.[90] Antallet testledd kan imidlertid påvirke Alpha-verdiene. En test med mange påstander gir normalt høyere verdier, mens få påstander gir lavere. I tråd med GJPs praksis er det ikke beregnet Alpha-verdier for skalaer med færre enn fire testledd eller for innsatsvariablene, siden disse bare består av én verdi.

[84] Basert på svar fra 380 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen.
[85] Basert på svar fra alle 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen.
[86] Basert på svar fra 522 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen. På en skala fra 1 til 7, der 1 var «helg uenig» og 7 var «helt enig», var medianscorene 6 på begge påstandene «Jeg synes spørsmålene er interessante» og «Jeg ønsker å se hvor godt jeg klarer å predikere for min egen del», mens den bare var 2 på påstanden «Jeg føler en plikt til å delta».
[87] Basert på svar fra 380 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen. På en skala fra 1 til 7, der 1 var «sank mye» og 7 var «steg mye», var medianscoren 4.
[88] Basert på svar fra 380 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen. På en skala fra 1 til 7, der 1 var «svært unyttig» og 7 var «svært nyttig», var medianscorene 5 både på spørsmål om hvorvidt deltagelsen deres hadde vært nyttig for forskningen og på spørsmål om den hadde vært nyttig for deres egen del.
[89] Frisk, E. (2021), 'Cronbachs alfa', *Statistisk ordbok*.
[90] Taber, K. (2017), 'The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education', *Research in Science Education*, 48:1, ss. 1–24.

	FFI240						GJP200						GJP350					
	Snitt	**SD**	**Min**	**Max**	**α**	**n**	**Snitt**	**SD**	**Min**	**Max**	**α**	**n**	**Snitt**	**SD**	**Min**	**Max**	**α**	**n**
Ravens (0–12)							8,02	2,51	0	12	0,71	797	7,91	2,65	0	12	0,74	1744
Shipley-2 Abstr. (0–25)													18,7	3,35	4	25	0,74	1081
Shipley-2 Block Patterns (0–26)	17,68	4,98	1	26	0,88	423												
CRT orig. (0–3)	2,47	0,79	0	3	0,5	443	2,12	0,96	0	3		798	2,13	0,96	0	3	0,55	1411
CRT utv. (0–4)							3,42	0,97	0	4		735						
CRT utv. (0–18)	15,04	2,91	4	18	0,79	443							14,9	3,18	2	18	0,81	1086
Tallforst. (0–3)							2,69	0,55	0	3		798						
Tallforst. (0–4)	2,75	1,21	0	4	0,6	443							3,27	1,01	1	4	NA	1082
Politisk kunnskap - FFI (0–50) - GJP 1. år (0-35) - GJP 2. år (0-50) - GJP 3. år (0-55)	35,39	6,60	8	50	0,8	587	 28,85 36,72	 3,03 4,61	 18 20	 34 48	 0,52 0,64	 666 773	 28,79 36,75 31,25	 3,10 4,61 5,08	 17 19 11	 35 48 52	 0,55 0,64 0,83	 1262 935 1106
Shipley-2 Voc. (0–40)													36,91	2,56	14	40	,71	1080

	FFI240						GJP200						GJP350					
	Snitt	**SD**	**Min**	**Max**	**α**	**n**	**Snitt**	**SD**	**Min**	**Max**	**α**	**n**	**Snitt**	**SD**	**Min**	**Max**	**α**	**n**
Aktiv fordoms-fri tenkning (1–7)	6,15	0,57	3	7	0,68	522	5,92	,61	3,86	7	0,63	781	5,94	0,62	3,29	7	0,67	1615
Kognitiv lukking (1–7)	3,87	0,80	1,6	6,2	0,85	522	3,34	0,58	1,46	5,09	0,56	666						
Rev–pinnsvin – påstand (1–5)							2,37	1,03	1	5		799	2,36	1,03	1	5		1628
Rev–pinnsvin – påstand (1–7)	2,78	1,45	1	7		522												
Rev–pinnsvin – test (1–7)							3,81	0,53	1,9	5,6	0,43	782						
Motivasjon – være blant de beste (1–7)	4,95	1,54	1	7		522							4,91	1,51	1	7		1082
Kognitiv motivasjon (1–7)	5,21	0,69	2,72	6,72	0,82	365							5,82	0,64	3,5	7	0,86	1398

	FFI240						GJP200						GJP350					
	Snitt	**SD**	**Min**	**Max**	**α**	**n**	**Snitt**	**SD**	**Min**	**Max**	**α**	**n**	**Snitt**	**SD**	**Min**	**Max**	**α**	**n**
Antall unike estimater																		1751
- Alle predik.							51,16	22,26	6	101		801	45,32	24,39	2	103		
- Første uke	33,28	18,76	5	128		857	28,04	17,56	1	97		797	25,63	18,04	1	101		1723
Brier-score forståelse (0–5)	0,87	1,19	0	5	0,63	522												
Antall spørsmål besvart	145	60,31	48	240		857	130	53,40	30	211		801	108	78,30	25	347		1751
Antall prediksjo-ner per spørsmål							1,68	0,89	1	10,46		801	2,14	3,61	1	114,0		1751
Tid brukt per spørsmål (minutter)	1,39	0,58	0,33	4,98		847	3,6	-	-	-		-						

Tabell 4.1 Variabelverdier i FFIs og GJPs turneringer.

4.2.1 Ekspertise

Den første kategorien av uavhengige variabler som antas å henge sammen med evnen til å predikere internasjonal politikk er ekspertise, som er basert på formelle kvalifikasjoner som utdanningsnivå, relevant erfaring og fagspesifikk kompetanse. Det er en grunnleggende antagelse i media og i utredninger at fagfolk er bedre stilt til å vurdere fremtidige utviklinger, fordi de er høyt utdannede og har mye kunnskap om det aktuelle temaet. EPJ viste imidlertid at ekspertise ikke ser ut til å bety noe for treffsikkerheten likevel: hverken utdanning, arbeidserfaring, tilgang til gradert informasjon eller hvorvidt ekspertene kom fra akademia eller ikke hadde noe å si for treffsikkerheten. Det var små forskjeller mellom eksperter som predikerte innenfor og utenfor sine egne fagområder, og ekspertene som var mest brukt i media var de dårligste til å predikere.

I likhet med EPJ og GJP var deltagerne i FFIs turnering generelt høyt utdannede (se underkapittel 3.2.2), som gjør det mulig å etterprøve om det er forskjeller mellom treffsikkerheten til personer med ulike nivåer av høyere utdanning. Over to tredeler av deltagerne i FFIs og GJPs turneringer hadde minst 4–5 års høyere utdanning, mens nesten alle hadde det i EPJ. Til forskjell fra både EPJ og GJP hadde FFIs turnering imidlertid en gruppe deltagere uten noe høyere utdanning, som også gjør det mulig å undersøke treffsikkerheten til hele spekteret av utdanningsnivåer. Omtrent halvparten av deltagerne i FFIs turnering arbeidet også innenfor forsvarssektoren i dag, der de kan antas å ha tilgang til gradert informasjon i motsetning til dem utenfor. En betydelig andel av disse var også forskere som arbeidet akademisk. Omtrent en tredel av alle deltagerne var også det som omtales som «eksperter» på temaer som spørsmålene i FFIs turneringer handlet om, fordi de arbeidet eller hadde arbeidet med forsvars- og sikkerhetspolitikk som en del av jobben sin. Ekspertene i FFIs turnering og EPJ hadde også omtrent like lang arbeidserfaring (10–12 år), slik at det ikke er særlig forskjell mellom erfaringsnivået i de to datagrunnlagene, men det var færre eksperter i FFIs turnering som hadde blitt intervjuet i media enn i EPJ.

4.2.2 Disposisjonelle variabler

Den andre kategorien av uavhengige variabler som antas å henge sammen med treffsikkerhet er personers disposisjonelle egenskaper. Dette handler om psykologiske forutsetninger for å predikere godt, på samme måte som en kan ha anlegg for å være god på andre områder, som sjakk. I GJP ble det målt fire typer disposisjonelle variabler, som nesten alle korrelerte med treffsikkerhet: kognitive evner, kunnskapsnivå, kognitive stiler og oppgavespesifikke ferdigheter.

4.2.2.1 Kognitive evner

Den første typen disposisjonelle variabler som ble målt i GJP bestod av tre forskjellige kognitive evner, som alle ble ansett som potensielt relevante i forbindelse med politisk prediksjon:[91]

[91] I GJP-studiene omtales disse tre evnene ofte som aspekter av «intelligens». Bakgrunnen er at det i psykologien ofte skilles mellom «flytende» intelligens, som dreier seg om evnen til å løse nye oppgaver og som i liten grad beror på tidligere læring, og «krystallisert» intelligens, som i større grad handler om evnen til løse problemer basert på tidligere kunnskap. De tre kognitive evnene omtalt her er assosiert med flytende intelligens, men kognitiv refleksjonsevne er samtidig noe annet enn intelligens. For enkelthets skyld bruker denne rapporten «kognitive evner» om abstrakt resonneringsevne, kognitiv kontroll og tallforståelse, mens «kunnskapsnivå» brukes i stedet for krystallisert intelligens.

1) *Abstrakt resonneringsevne*, det vil si evnen til å trekke slutninger fra enkeltobservasjoner til mer generelle prinsipper, som er kjernen i induktiv tenkning.[92] I prediksjonssammenheng kan det for eksempel være aktuelt å se på sammenhenger mellom et dagsaktuelt spørsmål, som sannsynligheten for et kupp i Russland, og relevante historiske tilfeller. Her må en se etter regelmessigheter, utlede hypoteser og teste dem.

2) *Kognitiv kontroll* (også kalt *kognitiv refleksjonsevne*), det vil si evnen til å unngå mentale snarveier som leder til gale svar. Tenk deg at du får følgende oppgave: «Et balltre og en ball koster 1,10 dollar. Balltreet koster 1 dollar mer enn ballen. Hvor mye koster ballen?». De fleste tenker umiddelbart at ballen koster 10 cent, men det riktige svaret er 5 cent. Denne oppgaven krever at tenker oss mer om enn det som faller oss mest naturlig. Noen personer har bedre evne til å unngå slike tankefeil enn andre.

3) *Tallforståelse*, det vil si evnen til å forstå tallkonsepter som sannsynlighet. Studier har vist at selv høyt utdannede personer har vanskeligheter med relativt enkle talloppgaver, f.eks.: «Sjansen for å få en virusinfeksjon er 0,0005. Av 10 000 personer, omtrent hvor mange av dem er forventet å bli smittet?».[93] Tallforståelse antas å ha betydning for evnen til å predikere spesielt kvantitative spørsmål, som oljeprisen, siden en tallkyndig person vil lettere kunne forstå forholdet mellom dagens kurs og variasjoner over tid.

Abstrakt resonneringsevne, ofte referert til som et mål på flytende intelligens eller bare intelligens, ble i GJP målt ved hjelp av en kortversjon av *Ravens Advanced Progressive Matrices* (Ravens APM).[94] APM-versjonen av Ravens matriser er ment for topp 20 % av befolkningen. Testen består av 12 matriseoppgaver som kan brukes uavhengig av respondentenes kulturelle og lingvistiske kunnskap. Det tredje året ble det gjennomført en ytterligere test fra *Shipley Institute of Living Scale 2 (Shipley-2)*.[95] Dette er også en test av abstrakt resonneringsevne, som her måles ved en *Abstraction Test*, der deltagerne får 25 oppgaver hvor de må fullføre sekvensielle oppgaver som «*white/black, short/long, down/…*» og «*oh/ho, rat/tar, mood/…*».

GJPs deltagere klarte rundt 8 av 12 riktige på Ravens-testen, som er relativt høyt og høyere enn snittet til universitetsstudenter (7).[96] I GJP200-artikkelen oppgis snittscoren som enda høyere (8,56), mens i GJP350-artikkelen scoret deltagerne rundt 8, slik som i denne rapportens reanalyser. På *Shipley-2 Abstraction*-testen klarte deltagerne i snitt 19 av 25 riktige. Her samsvarer verdiene i GJPs datasett og artikler. I begge GJPs artikler konkluderes det med at deltagerne scoret høyere enn gjennomsnittsbefolkningen på abstrakt resonneringsevne.

[92] I GJP200-artikkelen omtales den aktuelle Ravens-testen som et mål på «induktiv resonneringsevne», mens i GJP350-artikkelen refereres det til denne som én av flere tester av «flytende intelligens».
[93] Lipkus, I. M., Samsa, G. og Rimer, B. K. (2001), 'General Performance on a Numeracy Scale among Highly Educated Samples', *Medical Decision Making*, 21:1, ss. 37–44. Det riktige svaret er 5 personer.
[94] For en norsk beskrivelse av Ravens matriser, se Helland-Riise, F, og Martinussen, M. (2017), 'Måleegenskaper ved de norske versjonene av Ravens matriser [Standard Progressive Matrices (SPM)/Coloured Progressive Matrices (CPM)]', *PsykTestBarn*, 2:2.
[95] Shipley, W. C., Gruber, C. P., Martin, T. A. og Klein, A. M. (2009), *Shipley-2 Manual* (Western Psychological Services).
[96] Bors, D. A. og Stokes, T. L. (1998), 'Raven's Advanced Progressive Matrices: Norms for First-Year University Students and the Development of a Short Form', *Educational and Psychological Measurement*, 58:3, ss. 382–398.

I FFIs turnering var det ikke mulig å bruke de samme to testene av abstrakt resonneringsevne som i GJP. Den første testen (*Ravens APM*) er bare tilgjengelig for autoriserte psykologer eller sertifiserte brukere av evnetester. APM er også en versjon av Ravens matriser som er best egnet til voksne med høyt evnenivå, som ikke nødvendigvis ville passet deltagerne i FFIs turnering, der det ikke var satt noen formelle kompetansekrav for å kunne delta. Den andre testen (*Shipley-2 Abstraction*) forutsetter at respondentene kan engelsk flytende, som gjorde den uegnet for FFIs turnering, der alt ble gjennomført på norsk og deltagernes engelskkunnskaper var ukjent.

Abstraksjonstesten er imidlertid bare én av to tester av abstrakt resonneringsevne i *Shipley-2*. Den andre er en *Block Patterns*-skala med 12 flervalgsoppgaver, som ble utviklet som et ikke-verbalt alternativ til abstraksjonstesten og derfor egnet til å måle abstrakt resonneringsevne i FFIs turnering. Det var imidlertid nødvendig å oversette testens instruksjoner til norsk. Dette ble gjort med tillatelse fra rettighetshaverne og FFIs instruksjoner har nå blitt testens offisielle oversettelse som må benyttes av andre norske brukere av samme test. Deltagernes klarte i snitt 18 av 26 riktige, som er høyere enn snittscoren til en amerikansk gjennomsnittsbefolkning (15).[97] I likhet med GJPs scoret FFIs deltagere altså relativt høyt på tester av abstrakt resonneringsevne.

Kognitiv kontroll ble i GJP først målt med tre tester. Den første var *Cognitive Reflection Test* (CRT), som er den opprinnelige og mest kjente testen av kognitiv kontroll. Denne består av tre spørsmål, der den tidligere nevnte balltre- og balloppgaven er den første.[98] Den andre var en utvidet versjon av CRT-testen som bestod av fire tilleggsoppgaver, f.eks.: «Alle blomster har kronblader. Roser har kronblader. Hvis disse to påstandene er riktige, kan vi konkludere fra dem at roser er blomster?». Det tredje året ble testen av kognitiv kontroll utvidet ytterligere, fra 4 til 18 oppgaver.[99]

FFIs turnering benyttet de samme testene av kognitiv kontroll som GJP. Dette inkluderte den opprinnelige CRT-testen med tre oppgaver (se vedlegg A-1) og den mest utvidede versjonen med 18 oppgaver (se vedlegg A-2). Spørsmålene i den opprinnelige CRT-testen ble hentet fra den norske oversettelsen av Daniel Kahnemans bok *Tenke, fort og langsomt*, der denne testen refereres.[100] Det fantes imidlertid ingen norsk oversettelse av den utvidede testen med 18 oppgaver. Denne har derfor blitt oversatt til norsk i forbindelse med FFIs turnering og kvalitetssikret av en ekstern person som kan norsk flytende, men har engelsk som morsmål.

[97] Basert på snittscoren til aldersgruppen 40–49 år, siden snittalderen i FFIs turnering var 40 år. Se Shipley mfl. (2009), *Shipley-2 Manual*, s. 49.
[98] Frederick, S. (2005), 'Cognitive Reflection and Decision Making', *Journal of Economic Perspectives*, 19:4, ss. 25–42.
[99] Testen med alle 18 oppgaver finnes i vedlegg D i Baron, J. Scott, S. Fincher, K. og Metz, S. E. (2015), 'Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)?', *Journal of Applied Research in Memory and Cognition*, 4:3, ss. 265–284. To av de fire tilleggsoppgavene introdusert det andre året finnes også i denne kilden, mens opphavet til de to andre er ukjent.
[100] Kahneman, D. (2013), *Tenke, fort og langsomt* (Oslo: Pax Forlag), ss. 74–75. I denne oversettelsen får deltagerne flere svaralternativer per oppgave (*multiple choice*). Testen er brukt både med og uten svaralternativer. For en diskusjon om bruken av svaralternativer, se Sirota, M. og Juanchich, M. (2018), 'Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the Cognitive Reflection Test', *Behavior Research Methods*, 50, 2511–2522. I den opprinnelige Frederik (2005), 'Cognitive Reflection and Decision

På den opprinnelige CRT-testen klarte GJPs deltagere 2,1 av 3 riktige, som er svært likt snittet til universitetsstudenter.[101] Deltagerne i FFIs deltagere scoret høyere med 2,5 riktige i snitt. Cronbach Alpha-verdiene er imidlertid under tilfredsstillende nivå i begge turneringer, som tilsier at scorene ikke er pålitelige. Dette skyldes antageligvis at mange hadde sett flere av oppgavene før.[102] På den mest utvidede CRT-testen var reliabiliteten mye høyere. Her fikk deltagerne i FFIs turnering 15 av 18 riktige i gjennomsnitt.[103] Her finnes det ingen referansepopulasjon å sammenligne med, men dette er akkurat like mange riktige som deltagerne i GJP klarte.

I GJP ble deltagernes tallforståelse først målt ved tre oppgaver hentet fra to forskjellige kilder.[104] Imidlertid var reliabiliteten til scorene på denne testen også lav. Med et snitt på 2,71 av 3 svarte de fleste deltagerne i GJP alt riktig. Fra det tredje året brukte GJP derfor *Berlin Numeracy Test* i stedet.[105] Dette er en relativt ny test fra 2012 som har vist seg å være en sterk predikator på andre evner, som kognitiv kontroll. Berlin-testen er også spesielt godt egnet til å skille mellom høyt og lavt utdannede personer, og passer således godt til FFIs mer heterogene deltagermasse. FFIs turnering benyttet derfor bare Berlin-testen (se vedlegg A-3), men siden GJP først brukte denne det tredje året er det bare mulig å sammenligne scorene på tallforståelse med resultatene i GJP350.

På Berlin-testen klarte FFIs deltagere 2,8 av 4 riktige, som er en lavere snittscore enn GJPs 3,3. Scorene kan imidlertid ikke sammenlignes direkte og Cronbach Alpha-verdien kan ikke beregnes i GJP, fordi GJP benyttet en adaptiv versjon av Berlin-testen der deltagerne bare fikk en ny oppgave hvis de svarte riktig på den foregående, mens alle deltagerne i FFIs turnering fikk alle fire oppgaver. I FFIs turnering var Alpha-scoren akkurat innenfor det tilfredsstillende nivået og det var svært få deltagere som hadde sett oppgavene før.[106] Dette ble ikke undersøkt i GJP.

I GJP ble det funnet signifikante korrelasjoner mellom deltagernes treffsikkerhet og scorene deres på begge tester av abstrakt resonneringsevne, på begge tester av kognitiv kontroll og på testen av tallforståelse etter at den opprinnelige testen ble erstattet av en ny med andre oppgaver.

Making', må deltagerne fylle inn svaret selv. Det står heller ingenting om bruk av svaralternativer i noen av GJPs beskrivelser av bruken av denne testen. Det er derfor heller ikke brukt alternativer i FFIs turnering.

[101] Snittscoren til studenter ved Massachusetts Institute of Technology (MIT), som i hovedsak kom fra bachelornivå, var 2,2. Se tabell 1 i Frederick (2005), 'Cognitive Reflection and Decision Making', s. 29.

[102] I FFIs turnering svarte 232 (52 %) av 443 deltagere at de hadde sett minst én av de tre oppgavene før. I GJP200-datasettet svarte 177 (22 %) av 799 deltagere at de hadde sett oppgavene før, mens 536 (49 %) av 1087 deltagere svarte det samme det tredje året i GJP350-datasettet. Hvorfor det er så stor forskjell mellom GJPs datasett, f.eks. at kontrollspørsmålet ble formulert annerledes, er uvisst, men alpha-verdiene er rundt 0,5 i begge datasett.

[103] Denne scoren samsvarer med snittene på hhv. 16,7, 15,4 og 14,6 oppgitt i GJP350-artikkelen, der de to første snittene er basert på rundt 240 deltagere, mens det siste er basert på rundt 1500.

[104] Den første oppgaven ble hentet fra Lipkus mfl. (2001), 'General Performance on a Numeracy Scale among Highly Educated Samples', mens de to siste kom fra Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K. og Dickert, S. (2006), 'Numeracy and Decision Making', *Psychological Science*, 17:5, ss. 407–413.

[105] Cokely, E. T., Galesic, M., Schulz, E. og Ghazal, S. (2012), 'Measuring Risk Literacy: The Berlin Numeracy Test', *Judgment and Decision Making*, 7:1, ss. 25–47. At testen av tallforståelse ble byttet ut med denne er forklart i fotnote 6 i vedlegg 1d. i Friedman, J. A. (2019), *War and Chance: Assessing Uncertainty in International Politics* (Oxford University Press), som er én av bøkene skrevet basert på funnene fra GJP.

[106] Bare 33 (7 %) av 443 deltagere hadde sett minst én av de fire oppgavene før.

4.2.2.2 Kunnskapsnivå

Den andre typen disposisjonelle variabler i GJP var deltagernes kunnskap om internasjonal politikk. Antagelsen var at generell kunnskap på fagområdet er relevant for vurderinger av fremtidig utvikling. Hvis du for eksempel blir bedt om å anslå sannsynligheten for at FNs sikkerhetsråd vil autorisere en militær intervensjon i Syria, kan det være nyttig å vite at rådet har fem faste medlemmer som kan legge ned veto, deriblant Russland som er alliert med det syriske regimet.

I GJP ble politisk kunnskapsnivå målt ved en riktig/galt-test hvert av de tre første årene. Her fikk deltagerne påstander som «Aserbajdsjan og Armenia har formelt avgjort sin grensekonflikt», og ble bedt om svare om de mente påstanden var riktig eller feil. Det første året av GJP bestod testen av 35 påstander, det andre året av 50 påstander og det tredje av 55 påstander.

I FFIs turnering ble det politiske kunnskapsnivået målt på tilsvarende måte. Her ble det imidlertid bare gjennomført én test med 50 påstander i løpet av det tredje året (se vedlegg A-4). Testen hadde 10 påstander hver om 5 temaer som var gjennomgående for spørsmålene i turneringen: internasjonal politikk og væpnede konflikter, økonomi, Russland, NATO/Europa og USA. Respondenter er imidlertid generelt mer tilbøyelige til å være enig enn uenig når de får presentert påstander.[107] For å unngå dette bestod FFIs test av like mange riktige og gale påstander. Det samme var tilfellet i alle de tre testene i GJP.

Formuleringene av påstander i FFIs kunnskapstest var basert på to kriterier. For det første var målet at påstandene skulle måle *generell* kunnskap, som hvorvidt USA og Russlands befolkning er omtrent like stor, ikke detaljkunnskaper eller være «lurespørsmål», som hvorvidt NATOs intervensjon i Libya skjedde i 2011 eller 2012. For det andre skulle vanskelighetsgraden variere, slik at det ble mulig å skille mellom deltagernes kunnskapsnivå, samtidig som de ikke skulle være for lette eller for vanskelige for en gjennomsnittlig forsvars- og sikkerhetspolitisk interessert deltager. Påstandene ble kvalitetssikret ut fra disse kriteriene av kollegaer på FFI fra forskjellige vitenskapsdisipliner.

I tillegg til «riktig» og «feil» kunne deltagerne i FFIs kunnskapstest svare «vet ikke» på alle påstandene. I snitt svarte deltagerne «vet ikke» tre ganger. I teorien kan dette ha gjort at deltagere som bare gjettet på påstander de ikke visste svaret på, kan ha fått en høyere score. Litteraturen er imidlertid delt i synet på hvorvidt inkluderingen av «vet ikke» påvirker resultatene i kunnskapstester.[108] På den ene siden har eksperimenter vist at det å fraråde bruken av «vet ikke» øker andelen riktige svar, som tilsier at respondentene egentlig kan mer enn de ellers ville vist. På den annen side har eksperimenter også vist at andelen riktige svar ikke økte mer enn ved tilfeldig gjetning. Effekten av å inkludere «vet ikke», varierer også med hvem respondentene er. Kvinner har f.eks. vist seg å være mer tilbøyelig enn menn til å svare «vet ikke», også når de blir oppfordret til å gjette hvis de ikke kan svaret.

[107] Krosnick og Presser (2010), 'Question and Questionnaire Design', ss. 275–278.
[108] Boudreau, C. og Lupia, A. (2011), 'Political Knowledge', i Druckman, J. N., Green, D. P., Kuklinski, J. H. og Lupia, A., red., *Cambridge Handbook of Experimental Political Science*, ss. 171–183, ss. 172–177.

Hvorvidt «vet ikke» ble inkludert i GJPs kunnskapstester er ikke beskrevet. I GJP scoret deltagerne i snitt 29 (83 %) av 35 riktige det første året, 37 (74 %) av 50 riktige det andre året og 31 (57 %) av 55 riktige det tredje året, som innebærer et fall i andelen riktige fra 83 % det første året til 57 % det tredje.[109] Reliabiliteten til GJPs test fra det første året er imidlertid under tilfredsstillende nivå og testen fra det andre året er bare akkurat innenfor grensen. Dette er antageligvis årsaken til at antallet påstander ble økt til 55 påstander det tredje året, som bidro til at testen oppnådde en tilstrekkelig høy internkonsistens. Til sammenligning var snittscoren til deltagerne i FFIs turnering 35 av 50 riktige. Dette utgjør en andel på 70 % riktige, som er omtrent lik i GJP som helhet.

Det er ikke mulig å sammenligne scorene fra FFIs og GJPs kunnskapstester direkte, siden påstandene som ble brukt var forskjellige. Det finnes heller ingen referansepopulasjon å måle scorene opp mot, men reliabiliteten til FFIs test var like høy som testen i GJPs tredje år, som betyr at testene var pålitelige mål av den samme bakenforliggende variabelen i hver sin turnering.

Det tredje året av GJP ble det også innført en ny test av en annen type tilegnet kunnskap, nemlig vokabular. Dette var en *Vocabulary Test*, som også var hentet fra *Shipley–2*. Her fikk deltagerne 40 oppgaver, der de måtte finne ut hvilke ord som lå nærmest hverandre i mening. For eksempel fikk de ordet «*large*» og måtte velge hvilket av ordene «*red – big – silent – wet*» som lignet mest. Siden også denne testen forutsatte at respondentene var flytende i engelsk, var den uegnet for bruk i FFIs turnering, og det ble ikke gjennomført andre tester av vokabular her.

I GJP ble det funnet signifikante korrelasjoner mellom treffsikkerheten og deltagernes scores på kunnskapstestene fra alle årene i turneringen. Det var også en korrelasjon med deltagernes score på vokabulartesten, men denne var relativt svakere enn med politisk kunnskapsnivå.

4.2.2.3 Tenkemåter

Den tredje disposisjonelle variabeltypen som ble målt i GJP var deltagernes tenkemåter, også kalt kognitive stiler.[110] Kognitive stiler handler om *hvordan* folk tenker, i motsetning til *hva* de tenker på eller *hvor gode* de er. De beste ekspertene i EPJ skilte seg ut nettopp ved at de var mer åpne for å ta feil, samlet mer informasjon fra ulike kilder og var mer komfortabel med usikkerhet når de predikerte. I GJP ble denne typen «fordomsfri tenkning» målt på tre måter:

1) *Aktiv fordomsfri tenkning* (*actively open-minded thinking*) handler om å behandle ulike konklusjoner likt, selv om de går imot våre foretrukne svar. Personer som scorer høyt på dette blir mindre påvirket av eksisterende oppfatninger og er mer villige til å erkjenne at de selv kan ta feil. I en tidligere studie hadde forskere involvert i GJP funnet at personer

[109] Basert på GJPs datasett og samsvarer med snittscorene oppgitt i begge GJP-artiklene.
[110] I tillegg til tenkemåtene beskrevet her ble deltagerne i GJP også målt på hvordan de reagerte på hendelser som *bare nesten* skjedde (*close calls*) og deres tro på skjebnen (*belief in fate*). Måling av den første kognitive stilen forutsetter bruk av eksperimenter, som ikke var aktuelt i FFIs turnering. Den andre tenkemåten er ikke med i GJPs datasett og dermed ikke mulig å etterprøve. For mer om disse to tenkemåtene og funnene fra GJP, se Beadle (2021), 'Hvordan forbedre treffsikkerheten til prediksjoner av internasjonal politikk?', s. 27.

som scorer høyt på aktiv fordomsfri tenkning samler mer informasjon, og at mer informasjonsinnhenting forbedrer evnen til å estimere ukjente størrelser.[111] I den grad tilgjengelig informasjon kan bidra til å forutsi utfall av politiske spørsmål, var det derfor grunn til å anta at personer med høyere grad av aktiv fordomsfri tenkning også vil være bedre til å predikere enn andre.

2) *Kognitiv lukking* (*need for closure*) handler om å trekke konklusjoner raskt, ofte før alle bevis har blitt samlet, og aversjon mot tvetydighet.[112] Fordelene ved lukkethet er større handlekraft når en beslutning skal tas, men det øker også sjansen for feilslutninger, fordi avgjørelser kan bli tatt for raskt eller man overser viktig informasjon. Kognitiv lukking bidrar også til at man holder fast ved oppfatninger selv om bevisene tilsier at de er gale. I en tidligere studie hadde Tetlock funnet at eksperter med større behov for kognitiv lukking, hadde lettere for å avvise kontrafaktiske scenarioer som beviste at teoriene deres var feil, mens de omfavner kontrafaktiske scenarioer som beviste at de var rett.[113] En antagelse i GJP var derfor at et større behov for kognitiv lukking vil være til hinder for å modellere usikkerhet ved prediksjon av hendelser i den virkelige verden.

3) *Reve- vs. pinnsvintenkning*, det vil si i hvor stor grad personer foretrekker å applisere teorier en allerede kjenner godt fra før (pinnsvinet) eller om en forsøker å trekke på forskjellige vitenskapelige retninger (reven) når politiske fenomener skal forklares. Mer pinnsvinaktige personer har ofte også et større behov for kognitiv lukking.

I FFIs turnering ble alle tre typene fordomsfri tenkning målt på samme måte som i GJP. Aktiv fordomsfri tenkning ble målt ved at deltagerne måtte oppgi, på en skala fra 1 til 7, hvor uenig eller enig de var i 7 påstander, som: «Å endre din egen oppfatning er et tegn på svakhet» (se vedlegg A-5).[114] Behovet for kognitiv lukking ble målt på samme måte, men basert på 15 påstander, som: «Jeg liker ikke situasjoner som er usikre.» (se vedlegg A-6).[115]

Deltagernes grad av reve- vs. pinnsvintenkning ble målt ved at de fikk en kort beskrivelse om hvordan rever og pinnsvin tilnærmer seg prediksjon, og deretter ble de bedt om å vurdere hvorvidt de ville beskrive seg selv som mest lik en rev eller mest lik et pinnsvin (se vedlegg A-7). Denne testen ser ut til å ha blitt utarbeidet av forskerne bak GJP.[116] I GJP ble det brukt en skala

[111] Haran, U., Ritov, I. og Mellers, B. A. (2013), 'The role of actively open-minded thinking in information acquisition, accuracy, and calibration', *Judgment and Decision Making*, 8:3, ss. 188–201.
[112] Webster, D. M. og Kruglanski, A. W. (1994), 'Individual differences in need for cognitive closure', *Journal of Personality and Social Psychology*, 67:6, ss. 1049–1062; Kruglanski, A. W. og Webster, D. M. (1996), 'Motivated closing of the mind: "Seizing" and "freezing."', *Psychological Review*, 103:2, ss. 263–283.
[113] Tetlock, P. E. (1998), 'Close-call counterfactuals and belief-system defenses: I was not almost wrong but I was almost right', *Journal of Personality and Social Psychology*, 75:3, ss. 639–652.
[114] Se vedlegget til Haran mfl. (2013), 'The role of actively open-minded thinking in information acquisition, accuracy, and calibration'. Det tredje året av GJP ble testen av aktiv fordomsfri tenkning utvidet fra syv til ni påstander. Snittscorene i GJPs datasett er derfor basert på alle ni påstandene, selv om 669 (41 %) av de 1615 deltagerne som er registrert med en score på denne testen ikke svarte på de to nyeste påstandene. I FFIs turnering er det derfor valgt å bruke de syv påstandene som den opprinnelige testen består av og som er slik den først ble brukt i GJP. Hvorvidt snittet baseres på syv eller ni påstander utgjør imidlertid svært liten forskjell for scorene til deltagerne i GJP.
[115] Roets, A. og Hiel, A. V. (2011), 'Item selection and validation of a brief, 15-item version of the Need for Closure Scale', *Personality and Individual Differences*, 50:1, ss. 90–94.
[116] Mellers (2015), 'The Psychology of Intelligence Analysis', s. 5.

fra 1 til 5, der 1 var mest lik en rev og 5 mest lik et pinnsvin. I FFIs turnering var testen lik, men det ble brukt en skala fra 1 til 7, slik som på de andre testene av fordomsfri tenkning. Resultatene fra reve- og pinnsvintestene i de to turneringene rapporteres derfor i egne rader i tabell 4.1. Siden alle testene av fordomsfri tenkning bare fantes på engelsk, måtte også disse oversettes til norsk i forbindelse med FFIs turnering.

På testen av aktiv fordomsfri tenkning scoret deltagerne i begge turneringer rundt 6 av 7, som regnes som et relativt høyt nivå.[117] På testen av kognitiv lukking scoret FFIs deltagere 3,9 av 7 sammenlignet med 3,3 i GJP, som betyr at FFI har et litt høyere behov for kognitiv lukking enn GJPs. Samtidig beskrev både FFIs og GJPs deltagere seg selv som mer reveaktige enn pinnsvinaktige. Snittscoren til FFIs deltagere var 2,8 på en skala fra 1 til 7, mens GJPs var 2,4 på en skala fra 1 til 5.[118] FFIs deltagere beskrev seg altså som litt mer reveaktig enn deltagerne i GJPs, som er overraskende, fordi deltagerne i FFIs turnering også hadde et høyere behov for kognitiv lukking, som er ett av kjennetegnene på pinnsvin-, ikke revetenkning. En mulig forklaring er at deltagerne kjenner til rever og pinnsvin siden FFI har publisert artikler om disse stereotypene før, og derfor visste at det var bedre å være en rev enn et pinnsvin i prediksjonssammenheng.[119]

I løpet av GJPs to første år ble det også gjennomført en mer omfattende test av deltagernes reve- eller pinnsvintenkning, som også er rapportert i tabell 4.1. Her ble deltagerne bedt om å vurdere, på en skala fra 1 til 7, hvor uenig eller enig de var i 10 forskjellige påstander som skulle måle graden av reve- eller pinnsvintenkning, som: «Det er utrolig hvor ofte historien gjentar seg.». Selv om testen bestod av 10 påstander, er Alpha-scoren likevel svært lav. Siden den første testen av reve- eller pinnsvintenkning bare bestod av én påstand og reliabiliteten til den andre testen er lav, er det usikkert hvor pålitelige noen av målingene av denne variabelen er. Reve- eller pinnsvintenkning synes å være en egenskap som er vanskelig å måle. Testen ble ikke videreført det tredje året, og siden reliabiliteten var så lav ble den ikke brukt i FFIs turnering heller.

Av de tre målene på fordomsfri tenkning var det bare aktiv fordomsfri tenkning som korrelerte med treffsikkerheten i GJP. Kognitiv lukking og reve- og pinnsvintenking, som var to av de egenskapene som skilte gode fra dårlige eksperter i EPJ, hang derimot ikke sammen med treffsikkerheten.

[117] Mellers (2015), 'The Psychology of Intelligence Analysis', s. 7.
[118] Snittscoren er basert på en reanalyse av GJPs datasett, men samsvarer ikke med verdiene oppgitt i tabell 1 i Mellers (2015), 'The Psychology of Intelligence Analysis'. Det er nemlig et avvik mellom artikkelens beskrivelse av hvordan reve- eller pinnsvintenkningen ble målt (basert på én påstand og en skala fra 1 til 5) og verdiene oppgitt i tabellen, der snittscoren var 3,8 og scorene går opp til 6, som antageligvis også er brukt i korrelasjonsanalysen. Derfor må verdiene i tabell 1 være basert på den andre testen av reve- eller pinnsvintenkning med ti påstander og en skala fra 1 til 7, fordi snittscoren basert resultatene i datasettet på denne testen er nettopp 3,8. I så fall tilsier deltagernes score at de anser seg selv som litt mer reveaktige enn pinnsvinaktige, ikke litt mer pinnsvinaktig slik scoren ville betydd på en skala fra 1 til 5 og slik snittet er tolket i selve artikkelen. Datasettet inneholder også en annen test basert på en enkeltpåstand og en skala fra 1 til 5, slik målingen av reve- eller pinnsvintenkningen er beskrevet i artikkelen. Det er denne som er brukt for å beregne snittet på 2,4 i denne artikkelen, og funnet herfra samsvarer med funnet på den andre testen, nemlig at deltagerne vurderte seg selv som litt mer reveaktig.
[119] Beadle, A. W. (2017), 'Er du bedre til å forutsi fremtiden enn ekspertene?', *forskning.no*, 30. sept. 2017.

I superforecaster-artikkelen er det også rapportert to ytterligere mål på deltagernes tenkemåter:

1) *Kognitiv motivasjon* (*need for cognition*), som måler folks behov for og glede av å engasjere seg i aktiviteter som krever tenkning.[120] Kognitiv motivasjon handler ikke om individers evner, men viljen til å engasjere seg i oppgaver som krever dypere tenkning og til å bruke de evnene en har.[121] Folk med høyere kognitiv motivasjon setter større pris på diskusjoner og problemløsningsoppgaver, mens folk med lavere score har lettere for å ta mentale snarveier.

2) *Ønske om å være blant de beste i turneringen*. Bakgrunnen er at prestasjoner ofte henger sammen med motivasjonen for å delta, spesielt et høyt konkurranseinstinkt.

Kognitiv motivasjon ble målt gjennom en test der deltagerne ble bedt om å oppgi, på en skala fra 1 til 7, hvor godt 18 forskjellige påstander passet dem selv, f.eks. «Jeg foretrekker komplekse fremfor enkle problemer.» (se vedlegg A-8).[122] I både FFIs og GJPs turneringer ble ønsket om å vinne målt ved at deltagerne ble spurt om hvorfor de deltok. Her ble de bedt om å oppgi, på en skala fra 1 til 7, hvor uenig eller enig de var i en rekke mulige motivasjoner. Én av disse var at de «ønsket å være blant de beste» (se vedlegg A-9). I begge turneringene ble dette spørsmålet stilt i forbindelse med det tredje året.

På testen av kognitiv motivasjon var snittscoren til FFIs deltagere 5,2 av 7 mot 5,8 i GJP. Deltagerne i FFI turnering hadde altså en litt lavere kognitiv motivasjon, men scoren må likevel anses for å være relativt høy. Deltagernes ønsker om å havne blant de beste var lik i begge turneringer, med et snitt på rundt 5 av 7. I GJP hang både deltagernes kognitive motivasjon og et sterkere ønske om å havne blant de beste sammen med treffsikkerheten, men korrelasjonene var relativt svakere enn de fleste andre variablene som ble undersøkt.

4.2.2.4 Oppgavespesifikke ferdigheter

En fjerde og siste type disposisjonell variabel som ble undersøkt i GJP var spesifikke ferdigheter knyttet til prediksjon som oppgave.[123]

Én slik oppgavespesifikk ferdighet var *forecasting granularity*, det vil si hvor «finkornede» deltagernes sannsynlighetsvurderinger var. Noen personer bryter sannsynlighetsskalaen fra 0 % til 100 % ned i flere distinksjoner enn andre, for eksempel ved å bruke 22 %, 24 % og 26 % i stedet for å runde av til nærmeste 20 % eller 30 %. I GJP viste det seg at de beste deltagerne brukte flere forskjellige sannsynlighetsestimater enn resten, og at de fikk en dårligere Brier-score hvis

[120] Cacioppo, J. T. og Petty, R. E. (1982), 'The Need for Cognition', *Journal of Personality and Social Psychology*, 42:1, ss. 116–131.
[121] Cacioppo, J. T. og Berntson, G. G. (1994), 'Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates', *Psychological Bulletin*, 115:3, ss. 401–423.
[122] For testversjonen som ble brukt, se Cacioppo, J. T. og Petty, R. E. (1984), 'The Efficient Assessment of Need for Cognition', *Journal of Personality Assessment*, 48:3, ss. 306–307.
[123] Foruten *forecasting granularity* ble deltagerne i GJP også målt på den oppgavespesifikke ferdigheten *scope sensitivity*, som omhandler hvordan vi forstår omfanget av størrelser, som tidsperspektiver. Dette er nærmere beskrevet i Beadle (2021), 'Hvordan forbedre treffsikkerheten til prediksjoner av internasjonal politikk?', ss. 27–28. *Scope sensitivity* kan imidlertid bare måles gjennom eksperimenter og ble derfor ikke gjort i FFIs turnering.

deres prediksjoner ble rundet av til nærmeste 5 %, 10 % og 33 %.[124] Sagt på en annen måte; større presisjon i sannsynlighetsestimatene var en evne som bidro til å øke treffsikkerheten.

I gjennomsnitt brukte GJPs deltagere 51 unike sannsynlighetsestimater de to første årene og 45 de tre første.[125] Til sammenligning brukte FFIs deltagere i snitt bare 33 unike sannsynlighetsestimater. Tre firedeler av prediksjonene i GJPs var imidlertid oppdateringer av tidligere sannsynlighetsestimater (se underkapittel 3.3.2). Det kan derfor tenkes at deltagerne brukte flere unike estimater nettopp fordi de justerte opp eller ned prediksjoner de allerede hadde registrert før.

Hvis vi bare tar utgangspunkt i prediksjoner fra den første uken slik som i FFIs turnering, reduseres det gjennomsnittlige antallet unike sannsynlighetsestimater i GJP fra 51 til 28 de første to årene og fra 45 til 26 de tre første. Antallene blir nå lavere enn de 33 unike estimatene som ble brukt i FFIs turnering. Deltagerne i GJP var altså mer finkornede gjennom spørsmålsperiodene som helhet, men FFIs var mer finkornede fra starten av. Deltagerne kunne i begge turneringer også bruke desimaler i estimatene sine, som gir enda mer finkornede prediksjoner. Dette ble svært sjeldent brukt, men skjedde oftere i FFIs turnering enn i GJPs. I GJP350 var det bare 6 av 1 070 651 sannsynlighetsestimater som hadde minst ett desimaltall, mens i FFIs turnering var det 1 065 av 431 382. Den høyere andelen mer finkornede prediksjoner reflekteres også i forskjellene i de maksimale antallene sannsynligheter brukt i de to turneringene.

En annen oppgavespesifikk ferdighet, som bare har blitt målt i FFIs turnering, er *forståelsen av selve scoringssystemet* som ble brukt til å beregne treffsikkerheten. Bakgrunnen for at denne variabelen er potensielt relevant for treffsikkerheten, er at det gjennom turneringen viste seg å være (overraskende) mange matematikere på topplistene. Kommentarer fra deltagerne gikk ofte på at de enten ikke forstod hvordan scoringssystemet slo ut eller at de hadde et bevisst forhold til hvordan de fordelte prosentene for å få best mulig score. En ny hypotese i FFIs turnering var derfor at bedre forståelse av Brier-score-systemet bidrar til større treffsikkerhet.

Brier-systemet straffer nemlig prediksjoner som havner på feil side av 50/50 hardt, spesielt «bombesikre» prediksjoner som slår feil (se delkapittel 4.1). Det betyr at hvis du tar sjansen på bastante prediksjoner og er heldig nok til å treffe flere ganger, men bommer av og til, vil du over tid gjøre det dårligere enn om du hadde vært mer forsiktig, men traff mindre spektakulært. En deltager som oppgir 99 % sannsynlighet på ti binære spørsmål vil få en dårligere sammenlagt Brier-score enn en deltager som oppgir 70 % sannsynlighet på akkurat de samme utfallene, hvis han bommer på mer enn ett av spørsmålene. Ved å «skru ned» relativt sikre prediksjoner kan en dermed unngå relativt harde straffer uten å tape mye treffsikkerhet om svaret blir riktig.

I GJP fikk alle deltagerne «en rask innføring» i hvordan treffsikkerheten ble beregnet.[126] I tillegg fikk en del av deltagerne opplæring i probabilistisk tenkning, inkludert konsekvensene av

[124] Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters', s. 276.
[125] Antallene virker høye sammenlignet med snittene oppgitt i GJP350-artikkelen, der superforecasterne brukte 57 i snitt, top-team individuals 29 og resten 30 basert på alle prediksjoner fra hele spørsmålsperioden. En sannsynlig forklaring er imidlertid at denne rapportens reanalyse av GJP inkluderer flere og senere prediksjoner fra det fullstendige datasettet enn dem som er brukt til å beregne treffsikkerheten i GJP350-artikkelens replikasjonsdatasett.
[126] Mellers (2015), 'The Psychology of Intelligence Analysis', s. 5.

over- og underkonfidens.[127] I FFIs turnering fikk ingen deltagere en egen innføring i hvordan Brier-scoren ble beregnet, men nettsiden beskrev hvordan treffsikkerheten ble målt og denne ble det lenket til i hver eneste mail med spørsmålsresultater. I tillegg inneholdt hver resultatmail en beskrivelse av Brier-skalaen og hvordan scorene ble brukt til å beregne sammenlagte resultater.

Deltagernes forståelse av scoringssystemet ble målt gjennom en multiple choice-test med fem oppgaver utviklet i forbindelse med FFIs turnering (se vedlegg A-10). Siden ingen av deltagerne hadde fått en innføring, stilte de i utgangspunktet likt når det kom til kunnskap om scoringssystemet. Testen ble derfor laget slik at spørsmålene både målte deltagernes helt grunnleggende kjennskap til hvordan scoringssystemet fungerer og samtidig kunne brukes til å skille mellom deltagere med bare grunnleggende forståelse og dem som hadde en dypere forståelse.

I snitt klarte FFIs deltagere bare 0,8 av 5 riktige på testen av Brier-score-forståelse. Selv om dette snittet var svært lavt og testen bestod av få testledd, var reliabiliteten likevel innenfor tilfredsstillende nivå. Hvorvidt deltagerne forstod hvor dårlig de skjønte scoringssystemet er imidlertid tvilsomt. Rett før deltagerne fikk Brier-score-testen ble nemlig de samme deltagerne bedt om å oppgi, på en skala fra 1 til 7, hvor uenig eller enig de var i følgende påstand: «Jeg forstår hvordan scoringssystemet fungerer.» Her var medianscoren 5, som betyr at de fleste mente at de forstod hvordan scoringssystemet fungerte.[128] Dette tyder på at deltagerne trodde de forstod scoringssystemet bedre enn det de faktisk gjorde.

4.2.3 Innsats

Den tredje kategorien av uavhengige variabler som kan henge sammen med treffsikkerhet er personers *innsats*. Bakgrunnen er at mengdetrening regnes som avgjørende for prestasjonsevnen på mange områder, som sport og musikk.[129] Forskning har også vist at personer med et *growth mindset* – det vil si personer som anser læring og oppnåelse som ferdigheter som kan dyrkes – har større sannsynlighet for å prestere godt enn personer med et *fixed mindset* – der evner bare anses som medfødte («Jeg er dårlig i matematikk»).[130] Personer med et *growth mindset* liker utfordringer og klarer oftere å forbedre evnen sine, mens personer med et *fixed mindset* har lettere for å gi opp når det blir vanskelig.

Betydningen av øvelse og den positive effekten av et *growth mindset* kunne derfor tenkes å gjelde også innenfor prediksjon. I GJP ble derfor deltagerne målt på hvor mye innsats de la ned i turneringen, som et indirekte mål på deres grad av *growth mindset*.[131] I FFIs turnering ble deltagernes innsats derfor vurdert ut fra to av de samme målene i GJP:

[127] Moore mfl. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', s. 3553.
[128] Basert på svar fra 522 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen.
[129] Ericsson, K. A., Krampe, R. T., og Tesch-Romer, C. (1993), 'The role of deliberate practice in the acquisition of expert performance', *Psychological Review*, 100:3, ss. 363–406.
[130] Dweck, C. (2006), *Mindset: The new psychology of success* (New York: Random House).
[131] I tillegg til variablene beskrevet her, ble deltagernes påvirkning fra og adferd i GJP også målt på hvordan de brukte nettportalen og diskusjonsforumet til turneringen, inkludert antall klikk på nyhetsartikler, kommentarer, ord, innlegg, delinger og andelene innlegg som reiste nye spørsmål eller var svar på andres kommentarer. Disse er ikke

1) *Antall spørsmål besvart*. Dette er et direkte mål på mengden øvelse deltagerne fikk i å predikere gjennom FFIs turnering, og ble målt på samme måte som i GJP.

2) *Tid brukt per spørsmål*. Dette er et annet mål på hvor mye innsats deltagerne la ned på å predikere, men denne variabelen ble målt på forskjellige måter i de to turneringene.

I snitt svarte deltagerne i FFIs turnering på 145 (60 %) av 240 spørsmål. Både dette antallet og denne andelen var svært like deltagerne i GJP200, som svarte på 130 (62 %) av 211 spørsmål.[132] Det er imidlertid et stort avvik mellom antall spørsmål besvart i replikasjonsdatasettet og artikkelen til GJP350. I GJP350-artikkelen er det ikke oppgitt hvor mange spørsmål de rundt 1750 deltagerne besvarte til sammen. I stedet oppgis bare snittene til de tre gruppene som sammenlignes – superforecasterne, top-team individuals og resten av deltagerne – som bestod av hhv. rundt 120, 120 og 1500 deltagere. Det opplyses heller ikke hvor mange spørsmål som ble stilt hvert år, men ifølge artikkelen ble det ble stilt «over 100 spørsmål» de to første årene og «rundt 150 spørsmål» det tredje året, til sammen rundt 350 spørsmål.[133] Det oppgis imidlertid at superforecasterne svarte på hhv. 76, 116 og 81 spørsmål de tre første årene av turneringen, top-team individuals på 65, 84 og 52 og resten av deltagerne på 57, 82 og 60.[134] Hvis disse snittene legges sammen svarte de tre deltagergruppene på hhv. 273 (78 %), 201 (57 %) og 199 (57 %) av 350 spørsmål, som betyr at den samlede andelen ligger litt over 57 %, altså omtrent som i GJP200.

Andelene spørsmål besvart i GJP350-artikkelen fremstår imidlertid som uriktig høye. I replikasjonsdatasettet, som også baserer seg på rundt 1750 deltagere, er deltagerne registrert med scores på bare 108 av 347 spørsmål (31 %), og det er dette antallet som brukes i denne rapporten. Det er uklart hva som skyldes den store diskrepansen mellom de to kildene, og det er uklart hvordan superforecasterne kan ha svart på 116 spørsmål i snitt det andre året hvis det bare ble stilt rundt 100 spørsmål, slik det hevdes i artikkelen. Andre GJP-studier som baserer seg på de samme tre årene som GJP350 oppgir også mye lavere antall spørsmål besvart. I én studie fra 2017, som basert seg på 2860 deltagere som hadde svart på minst ett spørsmål, oppgis et snitt på 65 av 344 (19 %) spørsmål.[135] I en annen studie fra 2020, som inkluderte 515 deltagere som hadde oppdatert sine prediksjoner på minst 10 spørsmål i løpet av alle fire årene av turneringen, og dermed må antas å være blant de mest aktive deltagerne, oppgis det at disse deltagerne svarte på 113 av 481 (24 %) spørsmål i snitt.[136] I denne rapportens analyser er det derfor verdiene fra replikasjonsdatasettet som er lagt til grunn for beregningen av antall spørsmål besvart.

I gjennomsnitt svarte deltagerne i FFIs turnering altså på omtrent en like stor andel spørsmål som i GJP200 (rundt 60 %) og på en dobbelt så høy andel spørsmål som deltagerne i GJP350

relevant her, siden FFIs turnering ikke ble organisert som en interaktiv plattform. For mer om disse variablene, se Beadle (2021), 'Hvordan forbedre treffsikkerheten til prediksjoner av internasjonal politikk?', ss. 29–30.

[132] Disse verdiene samsvarer med dem oppgitt i GJP200-artikkelen, der de 743 deltagerne som ble analysert her svarte i snitt på 121 av 199 spørsmål (61 %) og predikerte 1,58 ganger per spørsmål. Se Mellers (2015), 'The Psychology of Intelligence Analysis', s. 8.

[133] Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters', s. 268.

[134] Se tabell 2 i Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters', s. 274.

[135] Moore mfl. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', s. 3555.

[136] Atanasov, P., Witkowsi, J., Ungar, L., Mellers, B. og Tetlock, P. (2020), 'Small steps to accuracy: Incremental belief updaters are better forecasters', *Organizational Behavior and Human Decision Process*, 160, ss. 19–35, s. 23.

(30 %), gitt at den lavere svarandelen er riktig. Hvor mange spørsmål deltagerne svarte på i gjennomsnitt, er imidlertid ikke nødvendigvis avgjørende. Selv om superforecasterne svarte på flere spørsmål enn andre, var ikke antall spørsmål med blant de variablene som korrelerte med treffsikkerheten i GJP350-artikkelen. Heller ikke i GJP200-artikkelen er det rapportert noen signifikant korrelasjon med antallet spørsmål besvart basert på de to første årene.

I GJP ble tiden deltagerne brukte på hvert spørsmål bare målt det andre året. Ingen av datasettene inneholder imidlertid data på denne variabelen. Det opplyses imidlertid i GJP200-artikkelen om at den gjennomsnittlige tiden deltagerne brukte per spørsmål var 3,6 minutter.[137]

I FFIs turnering ble antallet sekunder deltagerne brukte på hver spørsmålsrunde registrert automatisk. Dette tidsrommet strekker seg fra deltagerne åpnet spørreundersøkelsen til de trykket fullfør etter siste spørsmål.[138] Tiden brukt per spørsmål er her beregnet ved å dele alle sekundene deltagerne brukte på alle spørsmålsrundene på det totale antallet spørsmål de svarte på i turneringen. Det var noen deltagere som var registrert med ekstremt lang tidsbruk (flere dager), antageligvis fordi de ikke lukket undersøkelsen i nettleseren etter at de var ferdige å svare. 87 % av rundene er imidlertid registrert med en total tidsbruk på under en halvtime. Innenfor dette intervallet er deltagernes tidsbruk i FFIs turnering beregnet til 1,4 minutter per spørsmål.

I én av ekstraunderøkselsene ble deltagerne også bedt om å oppgi hvor lang tid de brukte. Her svarte 41 % av deltagerne at de brukte mellom 0,5 og 1 minutt per spørsmål, mens 21 % svarte at de brukte mellom 1 og 1,5 minutt.[139] Det betyr at over halvparten av deltagerne mente at de brukte mellom 0,5 og 1,5 minutt per spørsmål. Selv om den registrerte tidsbruken på 1,4 minutter ligger i det øvre sjiktet av dette intervallet, er det samsvar mellom hva deltagerne trodde og hvor lang tid de faktisk brukte. Dette styrker påliteligheten til målingen av denne variabelen.

Uansett mål på tidsbruk brukte deltagerne i FFIs turnering mye mindre tid per spørsmål enn i GJP. Hva som er årsaken til dette vites ikke. Et forbehold her er at tidsbruken til deltagerne i GJP ikke er registrert i det publiserte datasettet, slik at det er mulig å etterprøve beregningene, men forskjellen er likevel så stor at det er lite sannsynlig at det ikke var en betydelig forskjell.

I tillegg til antallet spørsmål besvart og tiden deltagerne brukte på dem, ble innsatsen i GJP også målt ut fra antallet prediksjoner deltagerne registrerte per spørsmål. I GJP200 predikerte deltagerne 1,7 ganger per spørsmål i snitt, mot 2,1 i GJP350.[140] Denne variabelen er heller ikke mulig å etterprøve, siden det ikke var mulig å svare mer enn én gang per spørsmål i FFIs turnering. Deltagerne i FFIs turnering kunne riktig nok åpne spørreundersøkelsen og endre prediksjonene sine i løpet av uken spørsmålsrunden var åpen, men dette ble sjeldent gjort. På en skala fra

[137] Mellers (2015), 'The Psychology of Intelligence Analysis', s. 9.
[138] Her inkluderes altså ikke spørsmålsrunder der deltagerne ikke fullførte undersøkelsen ved å lukke vinduet før de trykket fullfør. Dette var en mulighet de hadde for å kunne endre svarene senere, men få deltagere gjorde dette.
[139] Basert på svar fra 539 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen.
[140] Det maksimale antallet prediksjoner oppgitt for GJP350 i tabell 4.1 er svært høyt (114). Årsaken er at det var én deltager som registrerte én ny prediksjon omtrent hver dag på de 30 spørsmålene han besvarte, uten at han nødvendigvis endret så mye på dem underveis. De ti neste på listen er bare registrert med 13–46 prediksjoner per spørsmål.

1 til 7, der 1 var helt uenig og 7 var helt enig, var medianscoren til deltagerne 1 på påstanden: «Jeg åpner spørsmålsrundene på nytt for å endre svarene mine før fristen går ut.».[141]

I motsetning til antallet spørsmål besvart fant GJP korrelasjoner mellom deltagernes treffsikkerhet og antall prediksjoner og tid brukt per spørsmål. Faktisk var antallet prediksjoner den variabelen som korrelerte sterkest med treffsikkerheten og tiden brukt den nest sterkeste. I GJP viste det seg også at innsatsen deltagerne la ned i turneringen kunne forklare variasjoner i treffsikkerheten utover forskjellene på de disposisjonelle og situasjonelle variablene.[142]

4.2.4 Prediksjonsspesifikke tenkemåter

Den fjerde kategorien variabler som kan tenkes å henge sammen med personers treffsikkerhet er deres *prediksjonsspesifikke tenkemåter*. Mens disposisjonelle kognitive stiler handler om hvordan en behandler informasjon generelt, er det mulig at personer tenkte på helt andre måter når de predikerer i en turnering. Deltagere som i utgangspunktet scorer lavt på kognitiv lukking kan for eksempel få et situasjonsbetinget behov for å svare raskt, fordi de ikke har mye tid å bruke på den aktuelle aktiviteten. Dette er spesielt interessant å se nærmere på, siden funnene i EPJ og GJP peker i motstridende retninger om betydningen av kognitive stiler for treffsikkerheten.

I FFIs turnering ble det derfor gjort en kvantitativ og kvalitativ kartlegging av hvordan deltagerne tenkte når de predikerte. Etter at turneringen var over fikk deltagerne en liste med 17 vanlige måter å tenke på i forbindelse med prediksjon og ble bedt om å krysse av for hvilke som var dekkende for hvordan de tenkte når de fordelte sannsynligheter i turneringen (se vedlegg A-11).

Hensikten var å undersøke hvorvidt deltagerne som traff best skilte seg fra resten basert på prediksjonsspesifikke måter å tenke på, ikke bare generelle tenkemåter. Listen over kognitive tilnærminger var basert på de kognitive stilene til reve- og pinnsvinekspertene i EPJ, vanlige fallgruver som studier fra kognitiv psykologi har vist henger sammen med dårligere treffsikkerhet og metoder for probabilistisk tenkning som deltagerne i GJP fikk opplæring i.[143]

Tabell 4.2 viser antall og andel deltagere som krysset av for hver prediksjonsspesifikke tenkemåte. De mest relevante kognitive stilene, fallgruvene og metodene er oppgitt i parentes. Resultatene er basert på svarene til 381 av 857 deltagere, som både oppfylte minstekravet i FFIs turnering og svarte på ekstraundersøkelsen som ble sendt ut etter at turneringen var avsluttet. Merk at deltagerne ble bedt om å krysse av for de tenkemåtene de mente var mest dekkende for hvordan de gikk frem. I snitt krysset deltagerne av for 5,7 tenkemåter hver, som betyr at de aller fleste anvendte forskjellige tenkemåter. At færre enn 4 % av deltagerne krysset av for «Annet» indikerer at listen med alternativer dekker de fleste tenkemåtene deltagerne brukte.

[141] Basert på svar fra 522 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen.
[142] Mellers (2015), 'The Psychology of Intelligence Analysis', ss. 8–9.
[143] Chang mfl. (2016), 'Developing expert political judgment'.

Prediksjonsspesifikke tenkemåter	Antall	Andel
Baserte meg på mageføIelsen min. (intuisjon)	254	66,7 %
Tok utgangspunkt i en teori eller generell oppfatning jeg hadde av fenomenet fra før, og brukte denne til å vurdere hva som ville skje i dette tilfellet. (deduktiv resonnering)	179	47,0 %
Tok utgangspunkt i det aktuelle spørsmålet, og tenkte gjennom hva ulike teorier ville sagt om hva som ville skje. (induktiv resonnering)	79	20,7 %
Lette etter informasjon fra flere forskjellige kilder. (aktiv fordomsfri tenkning)	51	13,4 %
Baserte meg på det første som slo meg som mest sannsynlig. (kognitiv lukking)	151	39,6 %
Baserte meg på hvordan utfallene har fordelt seg ved tilfeller av samme type hendelse. (referanseklasser)	150	39,4 %
Baserte meg på tekstene og ev. figurer som fulgte med spørsmålet, og vurderte utfallene ut fra dette. (ankring)	246	64,6 %
Estimerte hvor ofte hendelsen eller utfallet hadde skjedd før. (grunnfrekvens)	150	39,4 %
Baserte meg på snittet av flere, forskjellige estimater av utfallet. (wisdom of the crowd)	23	6,0 %
Baserte meg på et lignende, historisk tilfelle som jeg kjente utfallet av. (bruk av én historisk analogi)	89	23,4 %
Baserte meg på flere lignende, historiske tilfeller med forskjellige utfall. (bruk av flere historiske analogier)	83	21,8 %
Tok utgangspunkt i dagens situasjon/nivå, og justerte min prediksjon deretter. (ankring)	243	63,8 %
Baserte meg på den siste utviklingen som hadde skjedd i saken, da spørsmålet ble stilt. (tilgjengelighetsheuristikk)	118	31,0 %
Fordelte prosentene slik at jeg fikk best mulig score hvis jeg traff, men samtidig unngikk å få en veldig dårlig score hvis jeg bommet. (optimalisering av Brier-score)	98	25,7 %
Baserte meg på en fremskrivning av den samme utviklingen som frem til nå. (ekstrapolasjon)	125	32,8 %
Tenkte på hva som gjorde at jeg bommet/traff på tidligere spørsmål. (post-mortem analyse)	42	11,0 %
Tok hensyn til uforutsigbare, overraskende hendelser som kunne påvirke utfallet. (sorte svaner)	71	18,6 %
Annet.	14	3,7 %

Tabell 4.2 Antall og andel deltagere som brukte forskjellige prediksjonsspesifikke tenkemåter i FFIs turnering.

Et slående funn er at deltagernes vanligste måter å tenke på var kognitive stiler forbundet med dårligere treffsikkerhet. Den mest brukte tilnærmingen var å følge mageføIelsen (68 %), også kjent som *intuisjon*. Intuisjon er imidlertid én av de vanligste årsakene til at vi tenker feil. Intuisjonen vår er det som leder oss til galt svar i tester av kognitiv kontroll, som i oppgaven om balltreet og ballen. Høyere tro på intuisjon er også forbundet med mindre aktiv fordomsfri tenkning.[144] Selv ikke for eksperter anses intuisjon som relevant i forbindelse med prediksjon av politikk.[145] For at intuitiv ekspertise skal fungere, forutsettes det et miljø som er regelstyrt og forutsigbart, slik at det er mulig å lære seg spillereglene og kunne gjenkjenne situasjoner hvor vi intuitivt kan se de beste løsningen.[146] Sjakk er det klareste eksempelet på dette, mens politikk anses som det stikk motsatte, fordi spillereglene er ukjente og usikkerheten er så stor at ekspertise har begrenset betydning for evnen til å predikere fremtidige utviklinger.

Den nest vanligste tilnærmingen var å basere seg på teksten og eventuelle figurer som fulgte spørsmålene (65 %), mens den tredje vanligste var å ta utgangspunkt i dagens situasjon/nivå og justere prediksjonene deretter (64 %). Begge disse tilnærmingene er former for *ankring*, der våre vurderinger påvirkes av et bestemt referansepunkt, men de baserer seg på to forskjellige ankre som er hhv. den spesifikke spørsmålsteksten og dagens situasjon generelt. Når vi først har fått et anker (f.eks. en prisantydning) å forholde oss til, opplever vi også et mindre behov for å lete etter flere. Et dårlig anker kan imidlertid bety at vi ikke justerer nok til å treffe godt. Ankringseffekten har også vist seg å være stor, selv om ankeret vi får er åpenbart absurd.[147]

Den fjerde vanligste tilnærmingen var å ta utgangspunkt i en teori eller generell oppfatning deltagerne hadde av fenomenet fra før og bruke denne til å vurdere hva som ville skje i det aktuelle tilfellet (47 %). Dette er selve definisjonen av *deduktiv resonnering* og var det fremste kjennetegnet på tenkemåten til pinnsvinekspertene i EPJ. I deltagernes egne beskrivelser av hvordan de predikerte var det særlig to generelle oppfatninger som gikk igjen. På den ene siden beskrev flere av deltagerne at hadde et optimistisk syn på verden som var styrende for hvordan de predikerte, f.eks. at det vil bli stadig færre og mindre alvorlige kriger. På den annen side beskrev mange av deltagerne at de hadde et pessimistisk syn, preget av verstefallstenkning. Dette ble også bekreftet i andre målinger av deltagernes syn på politiske utviklingen i verden, der de fleste mente at den gikk i negativ retning.[148] Flertallet var også bekymret for den sikkerhetspolitiske

[144] Én av påstandene i testen av aktiv fordomsfri tenkning er «Intuisjon er den beste guiden i beslutningstaking», der større grad av enighet gir lavere score på aktiv fordomsfri tenkning. Se vedlegget i Haran mfl. (2013), 'The role of actively open-minded thinking in information acquisition, accuracy, and calibration', s. 201.
[145] For en diskusjon av ekspertise i politisk prediksjon, se Tetlock (2017), *Expert Political Judgment*, ss. xviii–xix.
[146] Kahneman, D. og Klein, G. (2009), 'Conditions for Intuitive Expertise: A Failure to Disagree', *American Psychologist*, 64:6, ss. 515–526.
[147] I et kjent eksperiment ble en gruppe besøkende ved et offentlig laboratorium i San Francisco bedt om å vurdere hvorvidt det høyeste redwood-treet var mer eller mindre enn 365 meter. De ble også bedt om å gjette hvor høyt de trodde det var. Her var ankeret 365 meter. En annen gruppe fikk samme spørsmål, men et lavere anker på 55 meter. De to gruppene ga meget forskjellige gjennomsnittestimater: 257 meter og 86 meter – med 171 meter i forskjell. Dette og lignende eksperimenter er beskrevet i Kahneman (2013), *Tenke, fort og langsomt*, s. 137.
[148] Basert på svar fra 526 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen. På en skala fra 1 til 5, der 1 var «svært positiv retning» og 5 var «svært negativ retning», var medianscoren 4 på spørsmål om den sikkerhetspolitiske situasjonen i verden kommer til å utvikle seg i en generelt positiv eller negativ retning de neste fem årene. 59 % svarte «litt negativ retning», mens 18 % svarte «litt positiv retning» og 15 % svarte hverken eller.

utviklingen de neste fem årene.[149] De aller fleste deltagerne beskrev likevel at de la et *status quo*-syn til grunn når de predikerte, basert på en antagelse om at «verden vil stort sett fortsette som før». Alle de tre tilnærmingene – optimistisk, pessimistisk og *status quo* – ble beskrevet av deltagerne som «strategier» de fulgte.

Den femte vanligste tilnærmingen var å basere seg på det første som slo deltagerne som mest sannsynlig (40 %). Denne måten å tenke på er et eksempel på *kognitiv lukking*, som handler om ønsket om å trekke konklusjoner raskt før mer informasjon har blitt samlet. I tillegg svarte rundt en tredel av deltagerne at de baserte seg på den siste utviklingen som hadde skjedd i saken da spørsmålet ble stilt (31 %). Her er det en fare for at prediksjonene styres av *tilgjengelighetsheuristikken*, der sannsynligheten for noe vurderes ut ifra hvor lett man kan kommer på lignende tilfeller, ikke hvor vanlig de er statistisk sett. Hendelser som nettopp har skjedd, f.eks. en nordkoreansk atomprøvesprengning, gjør lignende hendelser mer «tilgjengelige» og at det derfor fremstår som mer sannsynlig at de vil skje igjen.

Felles for de vanligste kognitive stilene i FFIs turnering var at deltagerne i liten grad søkte mer informasjon om spørsmålene de fikk før de predikerte. Bare 13 % av deltagerne krysset av for at de lette etter informasjon fra flere forskjellige kilder, som er et uttrykk for *aktiv fordomsfri tenkning*. Dette bekreftes også av deltagerne selv. På spørsmål om hvorvidt de innhentet mer informasjon enn det som stod i spørsmålsteksten, svarte 66 % nei, 30 % tja og bare 4 % ja.[150] Dette fraværet av ytterligere informasjonsinnsamling står i sterk kontrast til den relativt høye scoren deltagerne fikk på testen av aktiv fordomsfri tenkning, som handler nettopp om å samle mer informasjon og vurdere ulike sider av en sak. Dette reiser et spørsmål om i hvor stor grad deltagernes generelle tenkemåter er overførbare til deltagernes faktiske adferd i turneringen. I tillegg svarte bare 6 % av deltagerne at de baserte seg på snittet av flere, forskjellige estimater. Dette er idéen bak *wisdom of the crowd*-metoder, der flere selvstendige vurderinger kombineres slik at gale estimater blir utlignet og den aggregerte treffsikkerheten blir bedre enn de enkelte.[151]

Av tenkemåtene som forbindes med færre tankefeil og høyere prediksjonsevne var den vanligste at deltagerne så på hvordan utfallene har fordelt seg ved tilfeller av samme type hendelse (39 %), også kjent som bruk av *referanseklasser*.[152] Den nest vanligste var å estimere hvor ofte hendelsene hadde skjedd før (39 %), som er å bruke *grunnfrekvens*. Begge disse tenkemåtene er assosiert med Kahnemans «utsideperspektiv».[153] Utsideperspektivet handler om å plassere fenomener inn i en større kontekst før man begynner å studere dem mer detaljert fra innsiden. Begge teknikkene er også anbefalt for å motvirke ankringseffekten. I sine egne beskrivelser av hvordan de tenkte viste noen av deltagerne til spesifikke fallgruver de forsøkte å unngå, som å legge for stor vekt på nylige, dramatiske hendelser, ikke la seg «lede» av spørsmålstekstene og å motstå

[149] Basert på svar fra 525 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen. På en skala fra 1 til 5, der 1 var «svært bekymret» og 5 var «svært ubekymret», var medianscoren 2 på spørsmål om deltagerne var bekymret eller ubekymret for den sikkerhetspolitiske utviklingen i verden i dag. Hele 65 % svarte at de var «litt bekymret» og 13 % at de var «svært bekymret».
[150] Basert på svar fra 838 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen.
[151] Surowiecki, J. (2005), *The Wisdom of Crowds* (New York: Anchor Books).
[152] Kahneman, D. og Tversky, A. (1977), 'Intuitive prediction: Biases and corrective procedures', *Technical Report PTR- 1042-77-6* (Virginia: DARPA).
[153] For mer om utsideperspektivet, se kapittel 23 i Kahneman (2013), *Tenke, fort og langsomt*.

sin egen hang til ønske- eller verstefallstenkning. Noen få deltagere oppgav også at de aktivt søkte opp andre prediksjoner av det samme, som på *betting*-sider.

Det var enda færre som brukte andre tilnærminger forbundet med bedre treffsikkerhet. Bare 21 % av deltagerne svarte at de først tok utgangspunkt i det aktuelle spørsmålet for så å tenke gjennom hva ulike teorier ville sagt om hva som ville skje. Dette er selve definisjonen på *induktiv resonnering* og var det fremste kjennetegnet på måten reveekspertene tenkte på i EPJ.

Det var også bare én av fire deltagere som baserte seg på flere lignende, historiske tilfeller med forskjellige utfall når de predikerte (22 %), mens det var like mange som svarte at de bare baserte seg på et lignende, historisk tilfelle som de kjente utfallet av (23 %). Den første tilnærmingen anses som nødvendig for å kunne trekke lærdommer fra historien, mens den andre tilnærmingen anses som en metodisk fallgruve. Selv om historien kan være en kilde til innsikt, er historikere og statsvitere skeptiske til analogisk resonnering.[154] Historiske analogier gir nemlig en kognitiv letthet som svekker vår kritiske sans ved å skape en mer sammenhengende verden og forståelige kausalsammenhenger enn det realiteten gir dekning for. Da blir det også fristende å bruke de første og mest lettfattelige analogiene som faller oss inn. Det anbefales derfor at en skal lete like mye etter forskjeller som likheter mellom dagens situasjon og tidligere hendelser. Det forutsetter at vi leter etter flere historiske tilfeller med forskjellige utfall, ikke bare ett.

En annen tilnærming som én av tre deltagere brukte, var fremskrivning av den samme utviklingen som frem til nå (33 %), også kjent som *ekstrapolasjon*. Dette er en vanlig prediksjonsteknikk, men har ikke blitt diskutert i forbindelse treffsikkerhet i EPJ eller GJP.[155] En fare ved ekstrapolasjon er imidlertid at det ikke tas høyde for trendbrudd. Det var imidlertid bare én av fem deltagere som svarte at de tok hensyn til uforutsigbare, overraskende hendelser som kunne påvirke utfallet (19 %), også kjent som *sorte svaner*.[156] Flere av deltagerne beskrev samtidig at noen av utfallene hadde kommet overraskende på dem og gjort at de bommet grovt. Likevel var det bare 11 % som brukte å tenke over hvorfor de hadde bommet/truffet på tidligere spørsmål. Denne teknikken kalles *post-mortem*-analyse og var noe deltagerne i GJP fikk opplæring i.[157]

Samlet sett var det altså tenkemåter forbundet med dårligere treffsikkerhet som var de vanligste blant deltagerne i FFIs turnering. I deltagernes egne, kvalitative beskrivelser kommer det samtidig frem en viktig nyansering av denne observasjonen.

Mange av deltagerne forklarte nemlig at de valgte forskjellige teknikker avhengig av hvor godt grunnlag de hadde for å kunne svare på det aktuelle spørsmålet. Det var først og fremst når deltagerne følte at de manglet et godt grunnlag for å svare på spørsmålet at prediksjonene ble basert på magefølelse og gjetting. Når deltagerne hadde kjennskap til temaet fra før, tenkte de seg grundigere om og stolte mer på egne vurderinger. Når deltagerne gjorde egne vurderinger kan

[154] Neustadt, R. E. og May, E. R. (1986), *Thinking in Time: The Uses of History for Decision-Makers* (New York: The Free Press). For metodiske anbefalinger, se spesielt kapitlene 13 og 14.
[155] Ekstrapolasjon diskuteres bare i kapittel 2 i Tetlock (2005), *Expert Political Judgment*, som en tilnærming til prediksjon som ekspertene bør forventes å kunne slå, og det er ikke nevnt i Tetlock og Gardner (2015), *Superforecasting*.
[156] Taleb, N. N (2010), *The Black Swan* (New York: Random House).
[157] Chang mfl. (2016), 'Developing expert political judgment'.

disse også deles inn i to tilnærminger. Den ene gruppen deltagere beskriver en matematisk tilnærming, der de «beregnet» seg frem til sannsynlighetene, f.eks. ved å ta utgangspunkt i en normalfordelingskurve. Den andre gruppen beskrev en mer skjønnsmessig vurdering av hva de mente var en «fornuftig» fordeling av sannsynlighetene, f.eks. ved å svare på hvor sannsynlig de mente det mest sannsynlige alternativet var først, før de fordelte de resterende prosentene uten hensyn til en bestemt type fordeling.

Generelt var det bare 8 % av deltagerne som følte at de hadde et godt grunnlag for å svare på de fleste spørsmålene, mens flertallet på 45 % hadde det på de færreste spørsmålene.[158] Dette passer med funnet om at den vanligste prediksjonsspesifikke tenkemåten var å følge mageølelsen, som deltagere gjorde når de følte at de ikke hadde noe grunnlag for å svare. Det er imidlertid ikke gitt at det er en sammenheng mellom hvor godt grunnlag deltagerne følte de hadde og hvor godt de faktisk traff, hvis f.eks. betydningen av ekspertise uansett er begrenset. Alle deltagerne fikk derfor også en valgfri mulighet til å oppgi hvor sikre de følte seg på hvert enkelt spørsmål de svarte på. Her kunne de velge mellom «bare gjetter», «ganske usikker», «ganske sikker» eller «helt sikker». Svarene herfra kan således brukes til å måle om deltagerne traff bedre eller dårligere på spørsmål de bare gjettet enn på dem de følte seg relativt sikre på.

En siste tenkemåte, som én av fire deltagere oppgav, var at de fordelte prosentene sine slik at de fikk en best mulig score hvis de traff, uten å få en veldig dårlig score hvis de bommet (26 %). Denne måten å svare på reflekterer et forsøk på å optimalisere Brier-scoren, siden dette scoringssystemet straffer «bombesikre» prediksjoner som er feil hardere enn det premierer like sikre prediksjoner på riktig svar.[159] Deltagerne som brukte denne tilnærmingen kan sies å ha tatt en mer «taktisk» tilnærming.

Deltagerne fikk også et eget spørsmål om hvor taktisk de tenkte når de predikerte i turneringen. Taktisk ble her definert som hvor mye deltagerne vektla konkurranseaspektet når de predikerte, f.eks. for å få en best mulig score/plassering, i motsetning til hvor sannsynlig de «egentlig» trodde at utfallene var. Her var det flere som svarte «litt taktisk» (43 %) eller «ganske taktisk» (17 %) enn som svarte «ikke taktisk» (34 %).[160] Deltagerne fikk også spørsmål om de tenkte mer eller mindre taktisk enn før. Her var det omtrent like mange som svarte «mer taktisk enn før» (42 %) og «hverken mer eller mindre taktisk enn før» (44 %), mens et klart mindretall svarte «mindre taktisk enn før» (8 %).[161]

Det fleste deltagerne oppgav altså at de svarte litt taktisk og mange av dem mer taktisk enn før. Deltagerne gav imidlertid motstridende beskrivelser av hvilken retning taktikken deres endret seg i. På ene siden beskrev en del av deltagerne at de begynte turneringen med å være forsiktig i sine sannsynlighetsfordelinger, men gikk over til mer bastante prediksjoner for å oppnå en bedre

[158] Basert på svar fra 538 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen. 41 % av deltagerne følte de hadde et godt grunnlag for å svare på omtrent halvparten av spørsmålene.
[159] F.eks. gir prediksjonene 90 % og 100 % på galt svar Brier-scores på hhv. 1,62 og 2, som utgjør en forskjell på 0,38. Samme prediksjoner på riktig svar gir scores på hhv. 0,02 og 0, som utgjør en forskjell på bare 0,02. Det er derfor lite å tjene på å justere opp allerede høye prediksjoner sammenlignet med straffen ved å ta feil.
[160] Basert på svar fra 535 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen. I tillegg mente 7 % at de svarte «svært taktisk».
[161] Basert på svar fra 535 av 857 deltagere som svarte på minst 20 % av spørsmålene i turneringen.

plassering, selv om det betød å avvike fra hvor sannsynlig de egentlig trodde utfallene var. På den annen side var det omtrent like mange deltagere som beskrev at de startet turneringen med bastante prediksjoner, men brant seg så mye på denne taktikken at de gikk over til mer forsiktige prediksjoner. Disse motstridende endringene ser ut til å ha utlignet hverandre, fordi det er ingen stor forskjell i andelen «bastante» prediksjoner på minst 90 % i første og andre halvdel av turneringen.[162] Deltagerne var ikke bare delte i sine beskrivelsene av hvordan de endret måten de predikerte på, men også i vurderingene av hvorvidt endringene ledet til bedre eller dårligere resultater. For noen hadde endringen ledet til bedre scores, mens for andre dårligere. Det er derfor ikke gitt at mer taktiske besvarelser hang sammen med høyere treffsikkerhet.

[162] På de 150 foreløpig avgjorte spørsmålene var det 15 479 prediksjoner på minst 90 % på 98 spørsmål i rundene 1–20 (157 prediksjoner per spørsmål) mot 9726 prediksjoner på 52 spørsmål i rundene 21–40 (187 per spørsmål).

5 Foreløpige resultater

Dette kapittelet presenterer de foreløpige resultatene fra FFIs prediksjonsturnering etter at de første 150 av 240 spørsmål ble avgjort ved utgangen av 2020.

Det første delkapittelet analyserer deltagernes *generelle* treffsikkerhet. Her sammenlignes treffsikkerheten i FFIs turnering med tilfeldig gjetning og resultatene fra GJP. Dette inkluderer analyser av variasjoner i deltagernes treffsikkerhet på forskjellige typer spørsmål, temaer, tidsperspektiver og prediksjonstidspunkter, som tidligere ikke har blitt undersøkt. Funnene herfra gir konkrete svar på hvor presist vi det er mulig å forutsi forsvars- og sikkerhetspolitiske spørsmål, hvorvidt det blir vanskeligere å predikere jo lenger inn fremtiden en ser og hvordan måten turneringer gjennomføres på kan påvirke treffsikkerheten til prediksjonene en kan hente ut fra dem.

Det andre delkapittelet undersøker betydningen av *ekspertise* ved å sammenligne treffsikkerheten til fagfolk med resten av deltagerne. Her etterprøves EPJs funn om betydningen av utdanningsnivå, relevant arbeidserfaring, faglig kompetanse, tilgang til gradert informasjon, akademisk bakgrunn og bruk i media. Her diskuteres det også om vi kan stole mer på eksperters enn amatørers prediksjoner, og hvordan vi kan velge hvilke eksperter vi bør høre mer på enn andre.

Det tredje delkapittelet undersøker sammenhenger mellom treffsikkerhet og *individuelle egenskaper*. Her sammenlignes FFIs turnering med GJP200-studien, der treffsikkerheten varierte med disposisjonelle variabler, som intelligens og kunnskapsnivå, og deltagernes innsats i turneringen. I tillegg analyseres det for første gang hvilke prediksjonsspesifikke teknikker som henger sammen med bedre treffsikkerhet. Her nyanseres tidligere funn om betydningen av ulike tenkemåter. Det fjerde delkapittelet ser nærmere på *de aller beste deltagerne* i FFIs turnering. Her sammenlignes resultatene med funnene fra GJP350, der det ble identifisert en gruppe «superforecastere» som var langt bedre og som skilte seg signifikant fra alle andre deltagere. Funnene fra de to siste delkapitlene har implikasjoner for hvordan vi kan rekruttere de «riktige folkene» i prediksjonssammenheng, og hvordan vi kan forbedre treffsikkerheten generelt.

De foreløpige resultatene baseres i utgangspunktet på 274 764 sannsynlighetsestimater fra 833 deltagere som har svart på minst 20 % av de 150 avgjorte spørsmålene. Det er imidlertid ingen vesentlige forskjeller mellom spørsmålene eller deltagerne i dette foreløpige datasettet og FFIs komplette datagrunnlag analysert i kapittel 3. Unntaket er at det gjennomsnittlige tidsperspektivet til de foreløpig avgjorte spørsmålene (372 dager), som ikke overraskende er noe kortere enn alle spørsmålene som har blitt stilt (521 dager). For fullstendig analyse av det foreløpige datagrunnlaget, tilsvarende det som ble gjort av hele datasettet i kapittel 3 og 4, se vedlegg B.

FFIs foreløpige resultater bygger i utgangspunktet på litt færre spørsmål enn i GJP200 (211) og litt under halvparten så mange som i GJP350 (347). Hvis alle sannsynlighetsestimater telles med, inkludert oppdateringene som bare kunne gjøres i GJP, består FFIs foreløpige datagrunnlag av omtrent halvparten av antallet prediksjoner i GJP200 (424 259) og en firedel av i GJP350 (1 070 651). Hvis vi derimot bare teller prediksjoner registrert i løpet av den først uken etter at spørsmålene ble publisert, slik som i FFIs turnering, består datagrunnlaget som analyseres her

allerede av flere prediksjoner enn i både GJP200 (108 752) og GJP350 (206 847). Når alle de 240 spørsmålene i FFIs turnering er avgjort vil antallet prediksjoner øke til 431 382.

FFIs turnering gir således et bedre utgangspunkt enn GJPs for å kunne etablere en «baseline» for hvor godt det er mulig å treffe i enkeltstående analyser, som årlige trusselvurderinger. I motsetning til GJP var midlertid alle ekstraundersøkelsene, der de disposisjonelle variablene ble kartlagt, frivillige. Derfor består FFIs datagrunnlag av færre deltagere med verdier på alle uavhengige variabler. Dette påvirker imidlertid ikke resultatene, som vil bli forklart senere.

Denne rapportens statistiske analyser er de samme som i GJPs artikler. Det anvendes t-tester for å sjekke om snittscorene til ulike grupper deltagere er signifikant forskjellige. Pearsons *r* brukes til å måle korrelasjonene mellom deltagernes treffsikkerhet og deres individuelle egenskaper. Her omtaltes alle resultater med p-verdier på 0.05 eller lavere som signifikante, men i praksis ligger de fleste p-verdiene under 0.0001 i denne rapporten.[163] Forenklet forklart vil det si at det er en 0,01 % sannsynlighet for at korrelasjonen som er funnet bare er tilfeldig. Slike bivariate analyser kan imidlertid ikke si noe om hvilke uavhengige variabler som bidrar *mest* til forskjellene i treffsikkerhet. Effekten som finnes kan tenkes å bortfalle hvis det er en annen bakenforliggende forklaring. Det planlegges derfor å gjennomføre en multippel regresjonsanalyse på et senere tidspunkt når alle spørsmålene i FFIs turnering er avgjort. I GJP200-studien ble de bivariate analysene supplert med multippel regresjon og *Structural Equation Modeling*. GJP350-studien består derimot, i likhet med denne rapporten, kun av bivariate analyser.

De statistiske analysene som gjennomføres her kan derfor ikke si noe om årsakene til at treffsikkerheten til deltagere varierte, bare om korrelasjonene med forskjellige uavhengige variabler. Siden de fleste uavhengige variablene som undersøkes er identiske med GJPs, gir likevel FFIs turnering en unik mulighet til å etterprøve tidligere identifiserte forskjeller og sammenhenger samt et grunnlag for å diskutere årsakssammenhengene beskrevet i tidligere studier. I tillegg har hundrevis av deltagere i FFIs turnering blitt intervjuet om hvordan de tenkte når de predikerte. Denne kvalitative informasjonen brukes her til å diskutere mulige forklaringer på variasjoner i prediksjonsevnen som kan undersøkes nærmere i de kvantitative analyser senere.

Et sentralt spørsmål er om resultatene fra FFIs studie bare kan si noe om deltagerne i akkurat denne turneringen eller om deltagere i prediksjonsturneringer generelt. Hvis vi bare ønsker å si noe om treffsikkerheten til deltagerne i FFIs turnering, er det ikke nødvendig at forutsetningene til de statistiske testene er oppfylte. I stedet er snittscorene som rapporteres den *faktiske* treffsikkerheten til deltagergruppene. I tillegg kan standardavviket si noe om spredning i deltagernes snittscores. Her oppgis derfor standardavvikene (SD) til snittscorene i parentes.

Samtidig kan det å generalisere funnene fra FFIs og GJPs turneringer ha en potensielt stor verdi, fordi prediksjonsturneringer er én av få tilgjengelige metoder for innsamling av et stort antall sannsynlighetsestimater fra mange forskjellige personer. Dette anses å gi et bedre utgangspunkt for prediksjon enn å basere seg på et lite antall eksperter hvis treffsikkerhet ofte er ukjent. Vi ønsker derfor å vite hvor godt deltagere i prediksjonsturneringer treffer generelt, og hva som

[163] For lesbarhetens skyld brukes punktum, ikke komma, for å skille statistiske fra andre typer analyser i rapporten.

kjennetegner dem som treffer bedre enn andre. I så fall kan FFIs deltagere regnes som et utvalg av hele populasjonen av personer som deltar i prediksjonsturneringer.

Hvis funnene fra FFIs turnering skal gjelde for deltagere i prediksjonsturneringer generelt, er det viktigere at forutsetningene til de statistiske testene som brukes er oppfylte. Her er alle statistiske sammenhenger undersøkt ved bruk av t-tester og Pearsons *r*, slik som i GJP. Begge disse testene forutsetter imidlertid normalfordelte data, som ikke alltid er tilfellet i FFIs eller GJPs datasett. Alle resultatene er derfor etterprøvd ved bruk av Wilcoxon-tester og Spearmans r_s, som ikke forutsetter normalfordeling. Resultatene er imidlertid stort sett de samme, uansett hvilken test som brukes. For mer om de statistiske testene benyttet i denne rapporten, se boks 5.1.

Vi kan også vurdere eksisterende årsaksforklaringer ved å sammenligne funnene fra FFIs og GJPs turneringer. Sammenfallende resultater kan underbygge eksisterende hypoteser om hvordan en egenskap påvirker treffsikkerheten, mens sprikende funn kan skape tvil om disse. For å sammenligne resultatene så direkte som mulig er alle beregninger og analyser av FFIs datasett gjort på samme måte som i GJPs. Alle resultatene fra GJP er i forbindelse med denne rapporten reanalysert fra bunnen av basert på offentlig tilgjengelige og tilsendte replikasjonsdatasett.

Mens statistiske tester kan si noe om hvorvidt det er en forskjell mellom treffsikkerheten til to deltagergrupper, kan det også være interessant å vite hvor gode estimater snittscorene vi rapporterer er for populasjonen som helhet.[164] Ettersom vi ikke kjenner de reelle snittscorene til deltagere i prediksjonsturneringer generelt, bruker vi observasjonene fra FFIs turnering som et estimat. Hvis vi ønsker å generalisere, må vi imidlertid også kunne anslå snittscoren til hele populasjonen av deltagere i prediksjonsturneringer. Dette kan gjøres ved å beregne et konfidensintervall rundt snittet til deltagerne i vårt utvalg. Det vanligste er å bruke et konfidensintervall på 95 %. Dette er et intervall som vi med 95 % sannsynlighet kan si at snittet til hele populasjonen av deltagere i prediksjonsturneringer ligger innenfor.

Konfidensintervallet representerer med andre ord feilmarginene i våre beregninger, slik som feilmarginene ved partiers oppslutning i politiske meningsmålinger. Et lite konfidensintervall tilsier at estimatene er sikre, mens et stort tyder på at estimatene er mer usikre. Konfidensintervallene kan også si noe om retningen og størrelsen på forskjellene vi ser, for eksempel mellom treffsikkerheten til deltagere med ulike utdanningsnivåer. Dette kapittelet inkluderer derfor alle figurer som sammenligner treffsikkerheten til ulike deltagergrupper et 95 % konfidensintervall.[165] Når konfidensintervallene til to gruppers gjennomsnittlige scores *ikke* overlapper, er det en statistisk signifikant forskjell mellom dem. Konfidensintervaller som overlapper betyr som regel at det ikke er en signifikant forskjell mellom gruppenes gjennomsnittsscores, men utelukker det likevel ikke. I denne rapporten er det derfor både gjort beregninger av konfidensintervallene til deltagernes snittscores og gjennomført signifikanstester av alle forskjeller mellom disse.

[164] For mer om forskjellene mellom hypotesetesting og konfidensintervall, se Lysne, V. og Olsen, T. (2017), 'Konfidensintervaller – hva kan de fortelle deg?', *Norsk tidsskrift for ernæring*, 1-2017.
[165] Bruk av konfidensintervall forutsetter at dataene er relativt symmetriske om snittet, men dette kravet blir mindre viktig jo større dataene er. Mary, I., Duranczyk, J. og Stottlemyer, S. L. (2013), 'Confidence Interval: Assumptions and Conditions', *OpenStax CNX*, 21. aug. 2013.

Hvilke statistiske tester er benyttet?
Hvorvidt data fra den virkelige verdenen oppfyller statistiske testers forutsetninger er ofte et tolkningsspørsmål. Én av de viktigste forutsetningene i t-testen og Pearsons *r* er at dataene er normalfordelte. Histogrammer som viser fordelingen til alle variablene som er brukt til å måle forskjeller og korrelasjoner i denne rapporten, er samlet i et eget FFI-notat.[166]

Formen på fordelingen av den avhengige variabelen (Brier-scorene) er tilnærmet like for alle deltagergruppene som sammenlignes i FFIs og GJPs turneringer, selv om verdiene deres er forskjellige. Mange, men ikke alle, scorene ser også tilnærmet normalfordelte ut. Av de uavhengige variablene er det derimot svært få som ser normalfordelte ut, men igjen er formen på fordelingene svært like på tvers av turneringene. Når datagrunnlagene blir store reduseres imidlertid viktigheten av normalfordelte data.[167] Det kan derfor argumenteres for at antallet deltagere i FFIs og GJPs turneringer er så store at det ikke spiller så stor rolle akkurat hvilke tester som brukes.

Siden formen på fordelingene er lik på nesten alle variabler og siden det er t-tester og Pearsons *r* som har blitt benyttet i GJPs artikler, er det samme gjort i denne rapporten. For sikkerhets skyld er resultatene også etterprøvd ved bruk av Wilcoxon-tester og Spearmans r_s.

Wilcoxon-testene det her er snakk om er Wilcoxon rank-sum-test, også kjent som Mann-Whitney *U*-test, og Wilcoxon signed-rank-test for pardata. Disse testene er et alternativ til t-tester som bare forutsetter lik form på fordelingen til dataene i de to gruppene som sammenlignes.[168] Dataene må altså ikke være normalfordelte, men skjeve til samme side. Av Brier-scorene som ikke ser tilnærmet normalfordelte ut i FFIs og GJPs datagrunnlag, har alle en litt høyreskjev fordeling. Det betyr at det var litt større spredning i scorene til deltagere med dårligere treffsikkerhet enn deltagere med bedre.

Resultatene fra statistiske tester avhenger samtidig av størrelsen på deltagergruppene som analyseres. På store datasett har t-testene og Wilcoxon-testene nesten like høy teststyrke, men Wilcoxon-testene er svakere ved små datasett. Ved små, ikke-normalfordelte utvalg blir resultatene mer usikre. Spearmans korrelasjonskoeffisient, som er et alternativ til Pearsons *r*, forutsetter ingen bestemt form på fordelingen til dataene og kan derfor være bedre egnet, gitt den skjeve fordelingen til mange av de uavhengige variablene i begge turneringer.[169] Resultatene fra Wilcoxon-testene og Spearmans r_s rapporteres bare hvis de avviker fra funnene fra t-testene og Pearsons *r*. Dette er imidlertid svært sjelden tilfellet.

Boks 5.1 Statistiske tester.

[166] Beadle, A. W. (2021), 'Tilleggsdokumentasjon til foreløpige resultater fra FFIs prediksjonsturnering', *FFI-notat 21/00133* (Kjeller: FFI).
[167] For en diskusjon av nødvendigheten av normalfordeling for datagrunnlag med forskjellige størrelser, se Skovlund, E. (2017), 'Når bør man velge en ikke-parametrisk metode?', *Tidsskriftet for Den norske Legeforening*, 16. mai 2017.
[168] Wilcoxon-testene det her er snakk om er Wilcoxon rank-sum-test, også kjent som Mann-Whitney *U*-test, og Wilcoxon signed-rank-test for pardata.
[169] For en diskusjon av fordeler og ulemper ved bruk av Pearsons *r* og Spearmans r_s, se Pripp, A. H. (2018), 'Pearsons eller Spearmans korrelasjonskoeffisienter', *Tidsskriftet for Den norske legeforening*, 8. mai 2018.

5.1 Generell treffsikkerhet

Det første forskningsspørsmålet i denne rapporten var: *Hvor presist er det mulig å predikere forsvars- og sikkerhetspolitiske utviklinger?*

For å besvare dette spørsmålet måles treffsikkerheten til FFIs deltagere ved hjelp av Brier-score (se delkapittel 4.1). Her måles evnen til å oppgi *høye* sannsynligheter til hendelser som *faktisk* skjer (f.eks. 80 % sannsynlighet for at russiske militære fly vil krenke norsk luftrom det neste året) og *lave* sannsynligheter til dem som *ikke* gjør det (f.eks. 20 % sannsynlighet for at russiske fly ikke vil krenke norsk luftrom). Brier-scoren reflekterer altså ikke bare evnen til å peke på hvilke hendelser som vil skje, men også hvor sikre en er på de riktige svarene.

For å si noe om hvor gode eller dårlige FFIs deltagere var, vil Brier-scorene deres sammenlignes med tilfeldig gjetning og deltagerne i GJP. Videre måles treffsikkerheten på tvers av temaer, typer spørsmål, tidsperspektiver, prediksjonstidspunkter og eventuelle tiltak som ble gjort for å forbedre treffsikkerheten underveis, fordi det her er forskjeller mellom datagrunnlagene til og gjennomføringen av FFIs og GJPs turneringer. Dette har ikke blitt gjort tidligere i GJPs studier. Deretter undersøkes det om resultatene endrer seg hvis treffsikkerheten måles ut fra treffprosent og kalibrering i stedet for Brier-score. Treffprosenten er hvor ofte en klarer å forutsi riktig utfall, mens kalibrering handler om hvor mye vi kan stole på prediksjonene. Til slutt diskuteres det hva funnene betyr for hvor presist vi kan forutsi spørsmål av relevans for norsk sikkerhet.

For å undersøke om forskjellene i treffsikkerhet er statistisk signifikante, gjennomføres tosidige t-tester av alle gjennomsnittlige Brier-scores som sammenlignes. Alle disse analysene er etterprøvd ved hjelp av Wilcoxon-tester, som ikke forutsetter en bestemt fordeling. En statistisk signifikant forskjell betyr imidlertid ikke at forskjellen må være betydningsfull. Det kan f.eks. være en signifikant forskjell mellom to deltagergrupper, selv om scorene deres er så like at avstanden ikke betyr noe i praksis. For å illustrere hva forskjellene faktisk innebærer, informeres det om hvor høy sannsynlighet en må oppgi for det riktige utfallet på et spørsmål om hvorvidt en hendelse vil skje eller ikke, for å oppnå Brier-scorene som sammenlignes.

Dette delkapittelets sammenligninger av den generelle treffsikkerheten i FFIs og GJPs turneringer er bare basert på GJP350. Dette skyldes at GJP350 inkluderer alle spørsmålene i GJP200 og resultatene er svært like på tvers av temaer, spørsmålstype, tidsperspektiv og prediksjonstidspunkt. Formen på fordelingene av scores i FFIs og GJPs turneringer er svært like på tilnærmet alle variasjonene som sammenlignes, selv om deltagernes snitt er forskjellige.[170] Siden reanalysen av GJP er basert på en kombinasjon av de respektive studienes replikasjonsfiler og det fullstendige, offentlig tilgjengelige datasettet, har det vært nødvendig å gjøre små tilpasninger i beregningene av deltagernes Brier-scores (se boks 5.2). I denne rapporten oppgis bare Brier-scorene som følger av reanalysen av GJPs datasett. Med mindre snittscorene avviker vesentlig fra dem som er oppgitt i GJPs artikler, oppgis artiklenes verdier bare i fotnotene.

[170] For histogrammer som viser fordelingen av Brier-scores innenfor hver spørsmålskategori, se kapittel 2 i Beadle (2021), 'Tilleggsdokumentasjon til foreløpige resultater fra FFIs prediksjonsturnering'.

Hvordan er GJPs datasett analysert?
I analysen av GJP er det tatt utgangspunkt i GJP350s replikasjonsdatasett med 347 spørsmål og 1751 deltagere som oppfylte kravet om å ha svart på minst 25 spørsmål i ett av de tre årene studien baseres på. Disse antallene samsvarer med dem oppgitt i artikkelen.[171]

For å kunne kontrollere resultatene oppgitt i replikasjonsdatasettet er deltagernes Brier-scores beregnet på nytt, basert på det fullstendige datasettet som inneholder alle spørsmål, deltagere og prediksjoner som er samlet inn gjennom alle fire årene av GJP. I denne sammenheng har det vært nødvendig å behandle prediksjonene fra det komplette datasettet slik at resultatene blir likest mulig dem som er beskrevet i GJP350-artikkelen:

- Prediksjoner fra deltagere som mangler en Brier-score på det aktuelle spørsmål i replikasjonsdatasettet inkluderes ikke, selv om det kan finnes prediksjoner fra deltagerne på det samme spørsmålet i det komplette datasettet. Deltagere som har svart på spørsmål i det komplette datasettet, men som ikke er blant deltagerne i replikasjonsdatasettet, er heller ikke med, uansett om de oppfylte minstekravene eller ikke.

- Hvis en deltager er registrert med flere prediksjoner enn det Brier-scoren i replikasjonsdatasettet er beregnet ut fra, er Brier-scoren beregnet på nytt basert på alle prediksjonene som finnes i det komplette datasettet. Dette gjøres fordi det vil gi et riktigere mål på deltagernes faktisk treffsikkerhet og fordi det er en diskrepans i beregningen av de opprinnelige Brier-scorene i replikasjonsdatasettet, der de noen ganger bare er basert på prediksjoner registrert samme år som spørsmålet ble publisert, mens andre ganger også er basert på prediksjoner registrert i påfølgende år. Dette gjelder bare spørsmål som ikke ble avgjort før turneringsåret var over, og dermed fortsatt åpne for å kunne predikere etter at det neste turneringsåret hadde begynt. Brier-scorene blir typisk bedre jo senere prediksjonene som inkluderes er registrert, men dette gjelder bare et mindretall av spørsmålene. På de aller fleste spørsmålene vil scorene være identiske, men for hver enkelt deltager kan det å inkludere senere prediksjoner gi et stort utslag på Brier-scoren på noen få spørsmål. Dette har imidlertid lite å si for den gjennomsnittlige scoren til deltagerne i turneringen som helhet. Det er bare en marginal forskjell mellom den gjennomsnittlige Brier-scoren til deltagerne i GJP350 om den bare baseres på prediksjoner fra samme år som spørsmålene ble publisert (0,32) eller om den baseres på alle prediksjoner i det komplette datasettet (0,31).

- Hvis en deltager er registrert med flere prediksjoner samme dag, er det bare den siste som teller. Det samme er gjort i alle beregninger av Brier-scores i GJP-studiene.

Som følge av disse grepene reduseres antallet prediksjoner i det komplette datasettet på spørsmål og fra deltagere i GJP350s replikasjonsdatasett fra 1 154 864 til 1 070 651.

Boks 5.2 Fremgangsmåte for reanalyse av GJPs datasett.

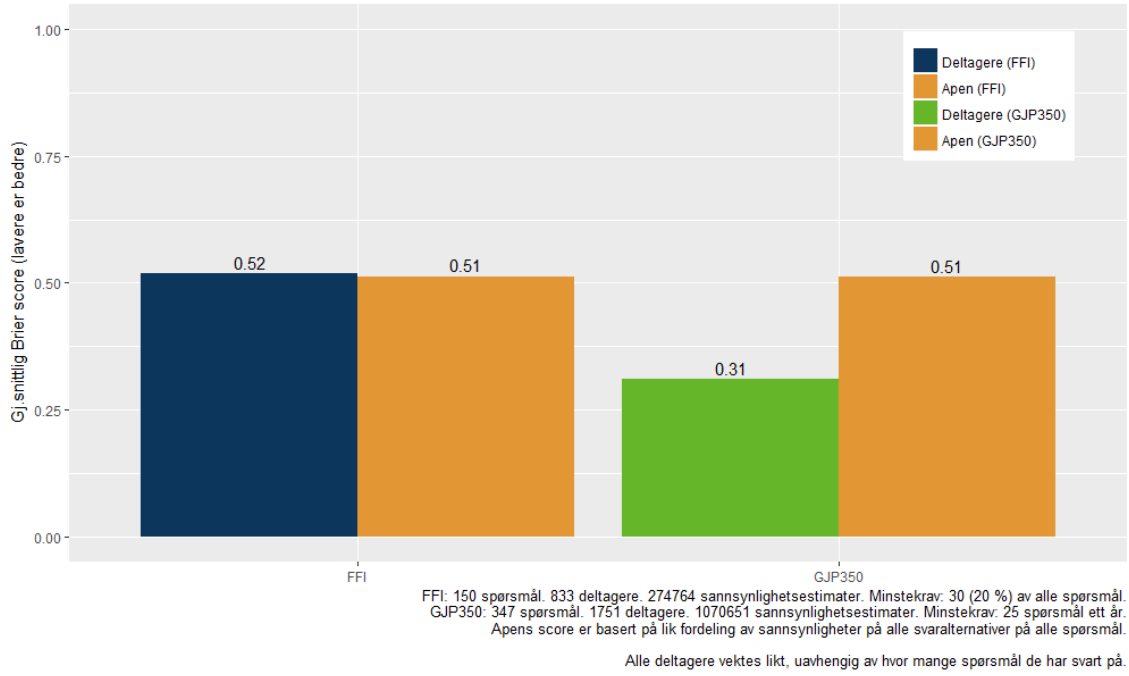
[171] Se Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters', ss. 268–270.

5.1.1 Gjennomsnittlig treffsikkerhet

Brier-scoren måler hvor presist en klarer å predikere på en skala fra 0 til 2, der *lavere* score betyr *bedre* treffsikkerhet. Du får en Brier-score på 0 hvis du oppgir en 100 % sannsynlighet for riktig utfall, mens du får en score på 2 hvis du oppgir en 100 % sannsynlig for feil svar. Brier-scoren handler ikke om den statistiske sannsynligheten for at en type hendelse skjer, som krenkelser av andre lands luftrom, men om sannsynligheten for at Russland vil gjøre det mot Norge i 2021. Dette er spesielt relevant i politisk sammenheng, der de samme typene hendelser, som krenkelser av luftrom, kan være relativt sjeldne. Da blir det enda viktigere å kunne forutsi når og hvor slike hendelser vil skje eller ikke, uavhengig av hvor vanlige de er statistisk sett.

Det er imidlertid vanskelig å si noe om hvor godt en treffer basert på en Brier-score mellom 0 og 2 alene. En vanlig tilnærming er derfor å sammenligne scoren med enkle algoritmer, som tilfeldig gjetning. Som ved mynt og kron vil prediksjoner basert på tilfeldig gjetning fordele sannsynlighetsestimatene likt på alle svaralternativer hvis øvelsen gjentas mange nok ganger, f.eks. 50/50 % på spørsmål med to mulige utfall, 33 % på hvert utfall ved tre alternativer og 25 % på hvert utfall ved fire alternativer. I EPJ ble en pilkastende ape med bind for øynene brukt som metafor for denne tilnærmingen. Her er derfor den samme algoritmen, omtalt som «apen», brukt til å sammenligne treffsikkerheten til deltagerne turneringer med tilfeldig gjetning. I tillegg kan vi diskutere den relative treffsikkerheten deltagerne i FFIs turnering ved å sammenligne den med resultatene fra lignende studier, som ekspertene i EPJ og deltagerne i GJPs turnering.

Figur 5.1 viser den gjennomsnittlige Brier-scoren til FFIs deltagere (blå), GJPs deltagere (grønn) og apene i hver turnering (oransje), basert på alle spørsmålene i hver studie.



Figur 5.1 Deltagernes gjennomsnittlig Brier-scores.

Basert på de 150 første spørsmålene er den gjennomsnittlige Brier-scoren til deltagerne i FFIs turnering 0,52 (SD: 0,11). Til sammenligning er apens Brier-score 0,51. Både deltagernes og apens treffsikkerhet tilsvarer den scoren du ville fått hvis du oppgav en 49 % sannsynlighet for det riktige svaret på et skjer/skjer ikke-spørsmål. Det er heller ingen statistisk signifikant forskjell mellom snittscorene.[172] Etter at to tredeler av spørsmålene er avgjort, har FFIs deltagere truffet omtrent akkurat like godt (eller dårlig) som tilfeldig gjetning.

I GJP traff deltagerne mye bedre. Her var deltagernes gjennomsnittsscore 0,31 (SD: 0,12).[173] Dette tilsvarer en prediksjon på 61 % sannsynlighet for det riktige av to mulige utfall. I GJP var apens score også 0,51, slik som i FFIs turnering. Det betyr at deltagerne i GJP slo FFIs deltagere og apen med like stor margin. Både forskjellen mellom GJPs og FFIs deltagere og mellom deltagerne i GJP og apen er signifikante.[174]

I utgangspunktet oppleves kanskje ikke forskjellen mellom prediksjoner på 49 % og 61 % for riktig svar som spesielt stor. Samtidig må sannsynligheter alltid fordeles. En prediksjon på 61 % for riktig svar, slik snittscoren i GJP tilsvarer, innebærer samtidig en prediksjon på 39 % for feil svar. Dette utgjør en forskjell i prediksjoner på rundt 20 prosentpoeng mellom riktig og galt utfall. I FFIs turnering utgjør forskjellen mellom 49 % for riktig svar og 51 % for galt svar bare 2 prosentpoeng. GJPs deltageres treffsikkerhet innebærer altså en relativt bedre evne til å skille mellom hendelser som skjer og ikke skjer.

I dette delkapitlet vil den relative treffsikkerheten til deltagerne i FFIs og GJPs turneringer diskuteres ved å oppgi den prosentvise prediksjonen en må oppgi for riktig utfall på et skjer/skjer ikke-spørsmål for å oppnå deres respektive Brier-scores.[175] Hensikten er å illustrere hvordan forskjellene endrer seg når det tas høyde for ulikheter i spørsmålene som ble stilt og hvordan turneringene ble gjennomført. Alternative måter å måle treffsikkerheten på, som kan si noe mer om hvor godt deltagerne traff i praksis, diskuteres mot slutten av dette delkapitlet.

[172] FFIs deltagere vs. apen: $t(832) = 1.30$, $p = 0.20$.

[173] I GJP200-artikkelen, som baserer seg på 743 deltagere og 199 spørsmål, er den gjennomsnittlige Brier-scoren oppgitt som 0,30. I replikasjonsdatasettet til samme artikkel, som baserer seg på 801 deltagere og 211 spørsmål, er snittscoren 0,32. I GJP350-artikkelen oppgis ikke snittscoren, men basert på replikasjonsdatasettet var denne 0,31. I GJP200-artikkelen var den gjennomsnittlige Brier-scoren ved tilfeldig gjetning 0,53, mens i GJP350-artikkelen oppgis ikke denne scoren. Basert på replikasjonsdatasettene blir imidlertid scoren 0,51 både i GJP200 og GJP350.

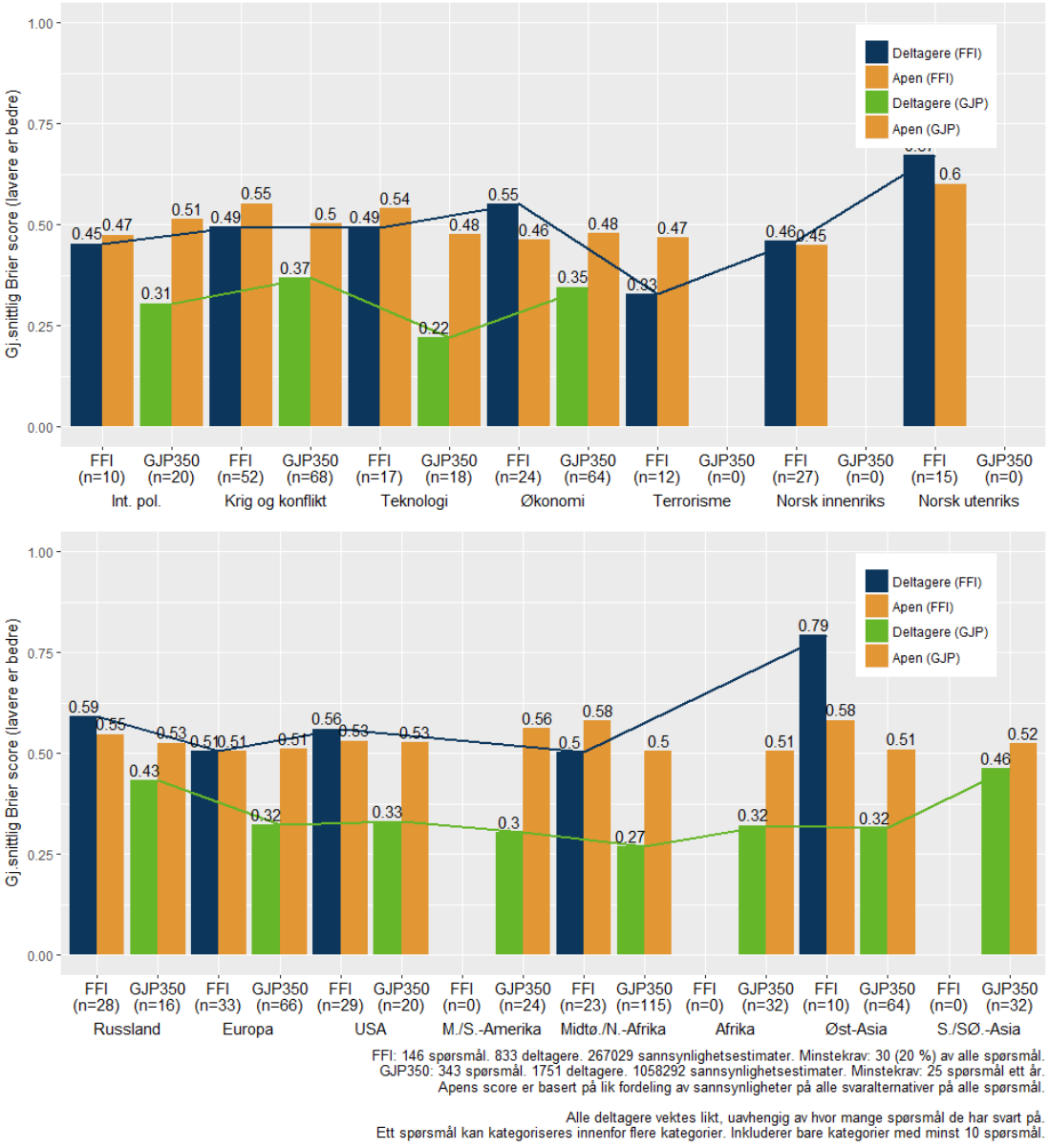
[174] GJP vs. FFIs deltagere: $t(1671) = -43.04$, $p < 0.001$. GJPs deltagere vs. apen: $t(1750) = -72.1$, $p < 0.0001$. Den statistiske forskjellen mellom snittscorene til deltagerne og tilfeldig gjetning er ikke målt i GJP350-artikkelen. Forskjellen er rapportert som signifikant i GJP200-artikkelen. Dette er også tilfellet i GJP200s replikasjonsdatasett.

[175] Å vurdere deltagernes treffsikkerhet basert på hvilket sannsynlighetsestimat Brier-scorene deres ville tilsvart på et spørsmål med bare to utfall er naturligvis en forenkling. Brier-scorene i begge turneringer er basert på forskjellige typer spørsmål der scorene beregnes på ulike måter. På kategoriske og ordinale spørsmål kreves det lavere prediksjoner for å oppnå den samme Brier-scoren som på et binært spørsmål. På et kategorisk spørsmål med fem svaralternativer kan en Brier-score på 0,5 oppnås ved å oppgi 36 % sannsynlighet for riktig svar og lik fordeling av resten av prosentene på de øvrige alternativene. På et ordinalt spørsmål med fem svaralternativer kan omtrent lik score oppnås ved bare 15 % for riktig svar, dersom svaralternativene til venstre og høyre for dette tildeles 5 % og 25 % sannsynlighet og de to siste alternativene får 25 % og 30 % sannsynlighet.

5.1.2 Tema

En første forskjell mellom datagrunnlagene i FFIs og GJPs turneringer er den tematiske fordelingen av spørsmålene som deltagerne fikk. Det kan for eksempel tenkes at spørsmålene i FFIs turnering omhandlet temaer som i seg selv var vanskeligere å predikere enn i GJP og at en høyere andel av vanskeligere temaer derfor har påvirket den gjennomsnittlige treffsikkerheten.

For å undersøke dette viser figur 5.2 den gjennomsnittlige Brier-scoren til deltagerne i FFIs og GJPs turneringer fordelt på de 15 temaene spørsmålene ble kategorisert innenfor. Linjene illustrerer mønstrene i treffsikkerhet innenfor og mellom turneringene på tvers av alle temaer.



Figur 5.2 Deltagernes gjennomsnittlige Brier-scores på hvert tema.

I utgangspunktet viser de foreløpige resultatene at deltagerne i FFIs turnering er bedre enn tilfeldig gjetning på spørsmål om internasjonal politikk generelt, krig og konflikt, teknologi, terrorisme og Midtøsten og Nord-Afrika, men dårligere enn dette på spørsmål om økonomi, norsk utenrikspolitikk, Russland, USA og Øst-Asia. Alle forskjellene mellom deltagerne og apen er statistisk signifikante, med unntak av temaene Europa og norsk innenrikspolitikk, der treffsikkerheten er helt lik. I praksis er imidlertid forskjellene små på nesten alle temaer.[176]

På spørsmål om krig og konflikt, som er det vanligste temaet i FFIs turnering, tilsvarer deltagernes og apens treffsikkerhet prediksjoner på hhv. 51 % og 48 % på riktig utfall på et skjer/skjer ikke-spørsmål. På spørsmål om Midtøsten og Nord-Afrika, som var temaet med flest spørsmål i GJP, tilsvarer scorene til deltagerne og apen prediksjoner på hhv. 63 % og 50 % på riktig utfall. GJPs deltagere var også signifikant bedre enn FFIs på alle de ni temaene med minst ti spørsmål hver i begge datasett.

Treffsikkerheten til deltagerne i de to turneringene er også svært stabil på tvers av temaer. De gjennomsnittlige Brier-scorene til deltagerne ligger rundt 0,5 i FFIs turnering og 0,3 i GJPs, som er omtrent det samme gapet mellom den generelle treffsikkerheten til deltagerne i turneringene som helhet. Det eneste temaet hvor FFIs score nærmer seg GJPs er terrorisme, men spørsmål på dette teamet kan ikke sammenlignes, fordi det ikke ble stilt et eneste spørsmål om dette i GJP.

Det er altså ikke grunnlag for å hevde at noen spørsmålstemaer i seg selv er lettere eller vanskeligere å predikere enn andre, basert på FFIs og GJPs turneringer. Det betyr samtidig at det er grunn til at å tro at er mulig å predikere spørsmål av relevans for norsk sikkerhet like presist som spørsmål om amerikansk sikkerhet.

5.1.3 Typer

En annen forskjell mellom FFIs og GJPs turneringer er typen spørsmål deltagerne fikk. I begge ble det stilt tre typer spørsmål: *binære* spørsmål med to mulige utfall, *kategoriske* spørsmål med minst tre svaralternativer, der bare ett av svarene er riktig, og *ordinale* spørsmål med minst tre svaralternativer, der noen av svarene er riktigere enn andre. Fordelingen er imidlertid ulik. Rundt halvparten av FFIs spørsmål er ordinale, mens det stort sett bare var binære i GJP.

På den ene siden burde en høyere andel ordinale spørsmål trekke den gjennomsnittlige Brier-scoren til deltagerne i FFIs turnering nedover, siden måten treffsikkerheten beregnes på ved ordinale spørsmål nærmest uansett gir en lavere Brier-score enn binære og kategoriske spørsmål (se underkapittel 4.1.4). På den annen side kan det høyere antallet svaralternativer per spørsmål ha gjort at FFIs deltagere fordelte sannsynlighetene sine jevnere og dermed oppgav en lavere sannsynlighet til det riktige svaret enn de ellers ville gjort.

[176] Det temaet hvor deltagerne traff relativt sett best er terrorisme (stort sett om antall angrep i Europa). Her tilsvarer deltagernes og apens snittscores prediksjoner på hhv. 60 % og 52 % på riktig svar. Deltagerne traff derimot mye dårligere enn apen på spørsmål om Øst-Asia (stort sett om Kina og Nord-Korea), der deltagernes score tilsvarer en prediksjon på 38 % på riktig svar mot apens 46 %.

La oss si at en deltager tror at forsvarsministeren etter stortingsvalget i 2021 vil komme fra Høyre. Hvis han får et kategorisk spørsmål om dette («Fra hvilket parti vil forsvarsministeren komme etter stortingsvalget i 2021?») med alle partier som mulige svar, kan det tenkes at han vil oppgi en lavere sannsynlighet for at svaret blir Høyre, enn om han hadde fått et binært spørsmål («Vil forsvarsministeren etter stortingsvalget i 2021 komme fra Høyre?») der han bare skal oppgi sannsynligheten for at svaret ble ja. En høyere andel kategoriske spørsmål kan dermed gjøre at sannsynligheten som tildeles det riktige utfallet, og dermed også treffsikkerheten, blir lavere enn om deltageren ble spurt om den samme hendelsen gjennom et binært spørsmål.

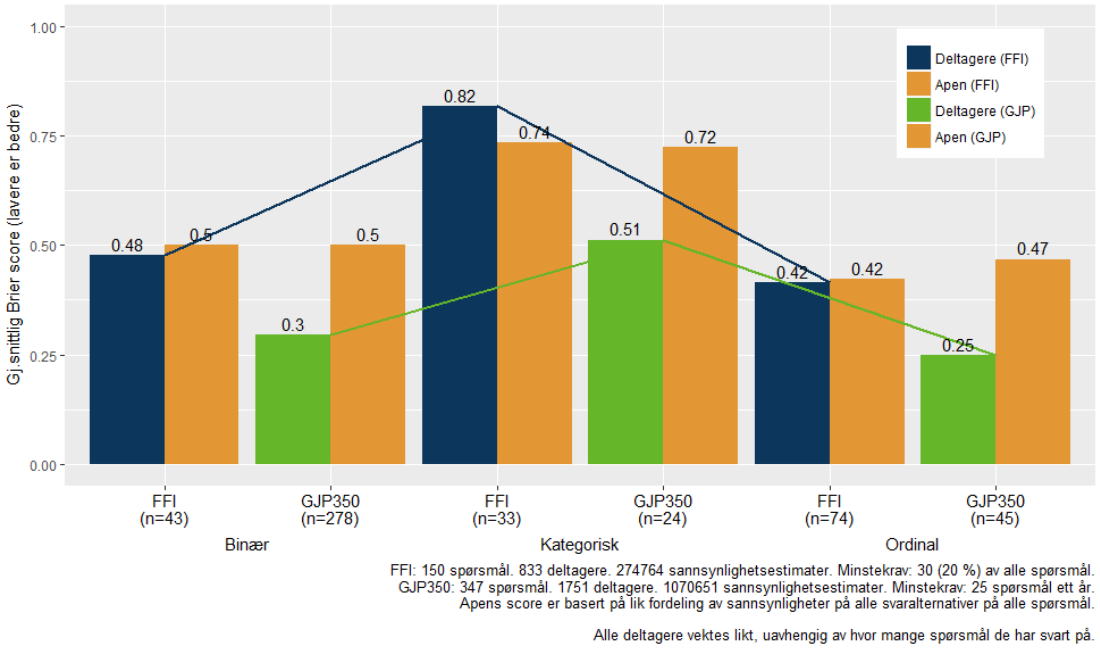
På den ene siden stemmer det at deltagerne i FFIs turnering oppgav en mye lavere sannsynlighet til de riktige svaralternativene enn deltagerne i GJP. FFIs deltageres gjennomsnittlige prediksjon på det som viste seg å være riktig svar er 47 % på tvers av alle spørsmål (binære: 64 %, kategoriske: 40 %, ordinale: 39 %). Til sammenligning var deltagerne i GJPs gjennomsnittlige prediksjon på riktig svar 72 % (binære: 74 %, kategoriske: 61 %, ordinale: 63 %).

På den annen side skyldtes ikke dette at FFIs deltagere oppgav en lavere sannsynlighet på det de *trodde* var riktig svar, slik som det ble spekulert i over. Tvert imot er det høyeste sannsynlighetsestimatet som deltagerne i FFIs turnering oppgav i snitt 72 % (binære: 83 %, kategoriske: 70 %, ordinale: 67 %).[177] Til sammenligning var de høyeste prediksjonene til GJPs deltagere i snitt 79 % (binære: 80 %, kategoriske: 71 %, ordinale: 70 %).[178] FFIs deltagere oppgav altså nesten like høye sannsynligheter som GJPs på alle spørsmålstyper. De forskjellene som finnes er så små at de har lite å si for treffsikkerheten.[179]

Problemet er at FFIs deltagere mye sjeldnere traff riktig svar. Som figur 5.3 viser, er treffsikkerheten til FFIs deltagere signifikant dårligere enn GJPs på alle typer spørsmål.[180] På de binære spørsmålene er forskjellen i treffsikkerhet omtrent like stor som ved turneringene som helhet, mens gapet er noe større på de kategoriske og noe mindre på de ordinale. Som ventet er Brier-scorene lavere på ordinale spørsmål, som det er flest av i FFIs turnering, enn på de andre typene. Dette trekker snittscoren nedover, men hjelper imidlertid lite på den relative forskjellen mellom turneringene når FFIs deltagere fortsatt treffer dårligere enn GJPs på denne typen også. I motsetning til deltagerne i GJP, sliter FFIs deltagere også med å slå tilfeldig gjetning på alle typer spørsmål. Tvert imot treffer FFIs deltagere dårligere enn apen på kategoriske spørsmål, mens de

[177] Snittene er basert på hver deltagers gjennomsnittlige sannsynlighetsestimat på alle og hver type spørsmål.
[178] Siden deltagerne i GJP kunne predikere flere ganger per spørsmål er snittet basert på den høyeste sannsynligheten på hver enkeltprediksjon, uavhengig av spørsmål, fordi å beregne snittet ut fra de høyeste prediksjonene per spørsmål vil gi et misvisende bilde av hvordan deltagerne predikerte når de skulle angi sannsynligheten for det de til enhver tid trodde var riktig svar. Det vil si at hver deltager vektes likt, som i beregningen av FFIs høyeste prediksjoner, men på bakgrunn av alle enkeltprediksjoner, ikke spørsmål, som i FFIs turnering blir det samme snittet. Forskjellen er uansett ubetydelig, ettersom snittet på alle spørsmål basert på høyeste prediksjoner per spørsmål også blir 79 %.
[179] På binære spørsmål vil en sannsynlighet på 83 % (som i FFIs turnering) gi en Brier-score på 0,06 ved riktig svar, mens en sannsynlighet på 80 % (som i GJPs turnering) vil gi en score på 0,08. Hvis sannsynlighetene er gitt på feil svar blir forskjellene bli relativt større, med Brier-scores på hhv. 1,38 og 1,28.
[180] FFIs vs. GJPs deltagere på binære spørsmål: $t(1150) = 24.83$, $p < 0.001$, på kategoriske spørsmål: $t(1804) = 30.96$, $p < 0.0001$, og på ordinale spørsmål: $t(2022) = 32.21$, $p < 0.0001$.

bare er marginalt bedre på binære og ordinale. Selv om forskjellene mellom deltagerne og apen er signifikante på alle spørsmålstyper, er forskjellene nærmest ubetydelige i praksis.[181]



Figur 5.3 Deltagernes gjennomsnittlige Brier-scores på hver spørsmålstype.

Til sammen tilsier resultatene at det ikke er forskjellen i typen spørsmål deltagerne fikk som kan forklare hvorfor FFIs treffer dårligere enn GJPs og omtrent likt med apen. Ved samme fordeling av spørsmålene som i FFIs turnering vil GJPs deltagerne fortsatt fått en Brier-score på 0,32 gitt samme snittscores per type. En annen fordeling av spørsmålstyper hadde derfor trolig ikke påvirket forskjellen i relativ treffsikkerhet.

Samtidig vil den relativt høyere andelen kategoriske spørsmål i FFIs turnering trekke den objektive Brier-scoren oppover sammenlignet med i GJPs. Det er derfor viktig å skille mellom spørsmålstypene ved sammenligninger av snittscorene til deltagerne i turneringene som helhet. I GJPs turnering hadde deltagerne også stort sett bare to svar å velge mellom, mens i FFIs hadde de som regel minst tre. Å konvertere gjennomsnittlige Brier-scores til tilsvarende prediksjoner på et binært spørsmål er derfor ikke egnet til å sammenligne treffsikkerheten på tvers av turneringene, men er fortsatt relevant for å illustrere forskjeller mellom deltagerne innad i hver.

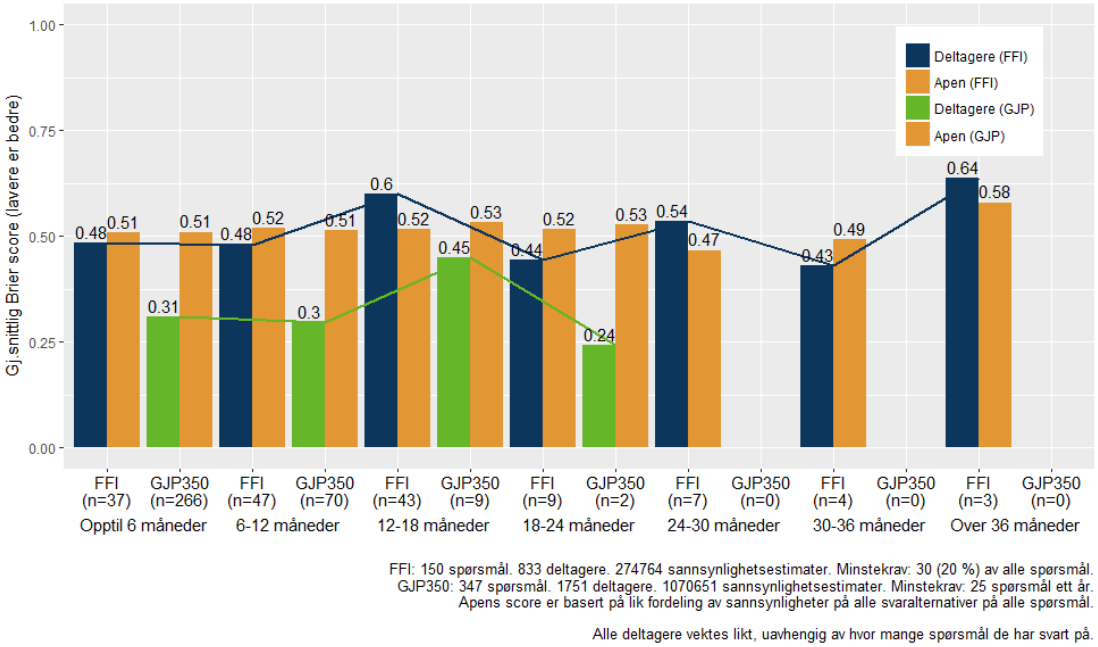
[181] FFIs deltagere vs. apen på binære spørsmål: $t(832) = -3.43$, $p < 0.001$, på kategoriske spørsmål: $t(832) = 10.47$, $p < 0.0001$, og på ordinale spørsmål: $t(832) = -2.00$, $p < 0.05$. Selv på de kategoriske spørsmålene, der gapet mellom deltagerne og apen er størst, er forskjellen ikke veldig stor. Her tilsvarer deltagernes Brier-score en prediksjon på 21,5 % på riktig svar på et kategorisk spørsmål med fire svaralternativer, gitt at de øvrige prosentene fordeles likt, mens apens score tilsvarer en litt høyere prediksjon på 25,5 %.

5.1.4 Tidsperspektiv

Sikkerhetspolitiske studier innledes ofte med et forbehold om at det blir vanskeligere å forutsi jo lenger inn i fremtiden en ser. Ifølge Tetlock var en forklaring på hvorfor ekspertene i EPJ slet med å slå tilfeldig gjetning nettopp at spørsmålene de fikk hadde et lengre tidsperspektiv enn det som faktisk var mulig å predikere.[182] Ekspertene traff bedre jo kortere frem de predikerte, men treffsikkerheten nærmet seg apens når tidsperspektivet ble 3–5 år. Til sammenligning var tidsperspektivet i GJP bare rundt 130 dager, og der traff deltagerne mye bedre enn tilfeldig gjetning.

En tredje forskjell mellom FFIs og GJPs turneringer er nettopp spørsmålenes tidsperspektiv. Det gjennomsnittlige tidsperspektivet på de 150 avgjorte spørsmålene i FFIs turnering er 372 dager. Dette er godt innenfor «grensen» på 3–5 år som EPJ fant at det var mulig å slå tilfeldig gjetning, men nesten tre ganger så langt som GJP hadde vist at det var mulig å treffe relativt godt.

Figur 5.4 viser den gjennomsnittlige Brier-scoren til deltagerne i FFIs og GJPs turneringer fordelt på de syv tidsperspektivene som spørsmålene ble kategorisert innenfor. Selv om alle forskjellene i figuren er statistisk signifikante på 0.0001-nivå, er gapene mellom deltagerne og apen i FFIs turnering relativt små, mens gapet fra GJPs deltagere og opp til både FFIs deltagere og apen er omtrent like stort som på tvers av temaene og spørsmålstypene.



Figur 5.4 Deltagernes gjennomsnittlige Brier-scores innenfor hvert tidsperspektiv.

[182] Tetlock og Gardner (2015), *Superforecasting*, s. 244.

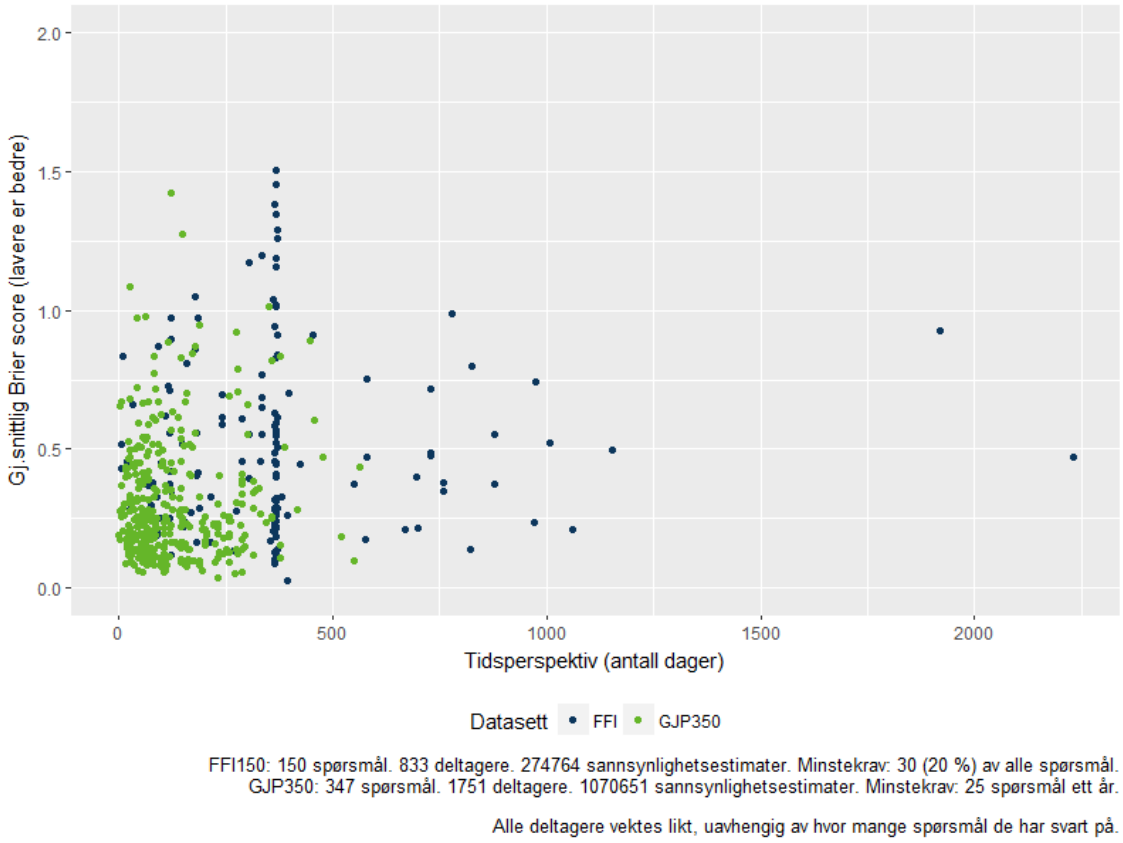
Treffsikkerheten holder seg overraskende stabil på tvers av tidsperspektivene i begge turneringer. Den forblir helt uendret selv om tidsperspektivet øker fra 0–6 til 6–12 måneder. Innenfor ett års tidsperspektiv synes det altså ikke å bli vanskeligere å forutsi jo lenger frem en ser.

Ett år er det perspektivet som vanligvis tas i trusselvurderingene til Politiets sikkerhetstjeneste (PST) og Etterretningstjenesten. De foreløpige resultatene er imidlertid tvetydige om hvor presise vi kan forvente at sannsynlighetsvurderinger i slike analyser kan være. Selv på spørsmål som bare ser et halvt år fremover, sliter FFIs deltagere med å slå tilfeldig gjetning. Mens EPJ fant at ekspertenes treffsikkerhet nærmet seg tilfeldig gjetning når spørsmålene så 3–5 år fremover, ligger deltagerne i FFIs turnering nært apens allerede fra start. GJPs deltagere treffer derimot bedre enn både FFIs deltagere og tilfeldig gjetning på spørsmål som ser 0–6 og 6–12 måneder fremover, som er de to tidsperspektivene med klart flest spørsmål i GJP.

På spørsmål med 12–18 måneders tidsperspektiv faller treffsikkerheten i FFIs turnering. Her gjør FFIs deltagere det faktisk dårligere enn apen, mens treffsikkerheten til GJPs deltagere også ser ut til å nærme seg tilfeldig gjetning etter ett år, men GJPs snitt er basert på svært få spørsmål. På spørsmål som ser over 18 måneder fremover treffer FFIs deltagere både bedre og dårligere enn ved kortere tidsperspektiver, men treffsikkerheten ser ut til å svinge i takt med apens.

Figur 5.4 viser altså en mulig knekk i treffsikkerheten når tidsperspektivet overstiger ett år i begge turneringene, men det er ingen synlig tendens til at treffsikkerheten fortsetter å falle jo lenger spørsmålene ser utover dette. Det er imidlertid få spørsmål med de lengste tidsperspektivene som er avgjort. Når alle spørsmålene i FFIs turnering er avgjort, vil det gjennomsnittlige tidsperspektivet øke fra 372 til 521 dager og antallet spørsmål som ser minst 18 måneder fremover vil øke fra 23 til 74. Dette vil derfor gi et bedre grunnlag for å undersøke mulige sammenhenger med treffsikkerheten utover ett og et halvt år frem i tid.

Treffsikkerheten innenfor de syv tidsperspektivene med seks måneders intervaller i figur 5.4 over kan imidlertid skjule en betydelig variasjon på spørsmålsnivå. Figur 5.5 viser derfor deltagernes gjennomsnittlige Brier-score på hvert spørsmål i begge turneringer. Her representerer hver prikk et eget spørsmål, rangert ut fra antallet dager frem i tid deltagerne ble bedt om å predikere.



Figur 5.5 Deltagernes gjennomsnittlige Brier-scores på hvert spørsmåls tidsperspektiv.

For det første viser figur 5.5 at de fleste spørsmålene i GJP har et kortere tidsperspektiv enn snittet på 130 dager. Medianen er bare 87 dager. 63 % av spørsmålene et tidsperspektiv på under 130 dager, mens bare 37 % så lenger frem enn dette. Det betyr at GJPs resultater er basert på et kortere tidsperspektiv enn det snittet, som er det eneste som oppgis i artiklene, gir inntrykk av.

For det andre viser figur 5.5 at svært mange av de avgjorte spørsmålene i FFIs turnering har et tidsperspektiv på ett år (rundt 365 dager). Det er altså betydelig mindre variasjon i tidsperspektivene i FFIs turnering sammenlignet med GJPs. Dette bekreftes av at medianen av tidsperspektivene til FFIs spørsmål er 367 dager, som ligger svært nært snittet på 371 dager.

For det tredje viser figur 5.5 ingen knekk i treffsikkerheten når tidsperspektivet overstiger ett år. Treffsikkerheten på FFIs spørsmål er ganske jevnt fordelt, uavhengig av antallet dager deltagerne måtte predikere. Scorene i GJP viser heller ikke et mønster som tilsier at treffsikkerheten faller jo lenger frem en ser. Det er heller ingen signifikante korrelasjoner mellom snittscoren og antall dagers tidsperspektiv per spørsmål i noen av turneringene.[183]

[183] Siden spørsmålenes tidsperspektiver er ikke normalfordelte i noen av turneringene er korrelasjonene målt ved både Pearsons r og Spearmans r_s. FFIs turnering ved Pearson: $r = 0.04$, $t(148) = 0.48$, p = 0.63, og Spearman: $r_s = 0.05$, $p = 0.55$. GJPs turnering ved Pearsons: $r = 0.08$, $t(345) = 1.49$, p = 0.14, og Spearman: $r_s = 0.006$, $p = 0.91$.

Det avdekkes dermed ingen sammenheng mellom treffsikkerheten og tidsperspektivet på spørsmålene i FFIs eller GJPs turneringer. Selv om det er riktig at treffsikkerheten til ekspertene i EPJ nærmet seg tilfeldig gjetning når tidsperspektivet nærmet seg fem år, er det bemerket at forskjellene i ekspertenes treffsikkerhet på kort sikt (1–2 år) og lang sikt (3–5 år, pluss noen spørsmål som så lenger) egentlig ikke var så store.[184] Hvordan treffsikkerheten eventuelt henger sammen med tidsperspektivet på spørsmål som ser enda lenger frem, er ennå ikke analysert i EPJ. Da EPJ ble publisert i 2005, var det bare noen av spørsmålene med et tidsperspektiv på 10 år eller mer som var avgjort. Innen 2022 vil imidlertid de fleste spørsmålene i EPJ, inkludert dem som så 25 år fremover, også være avgjort.

5.1.5 Prediksjonstidspunkt

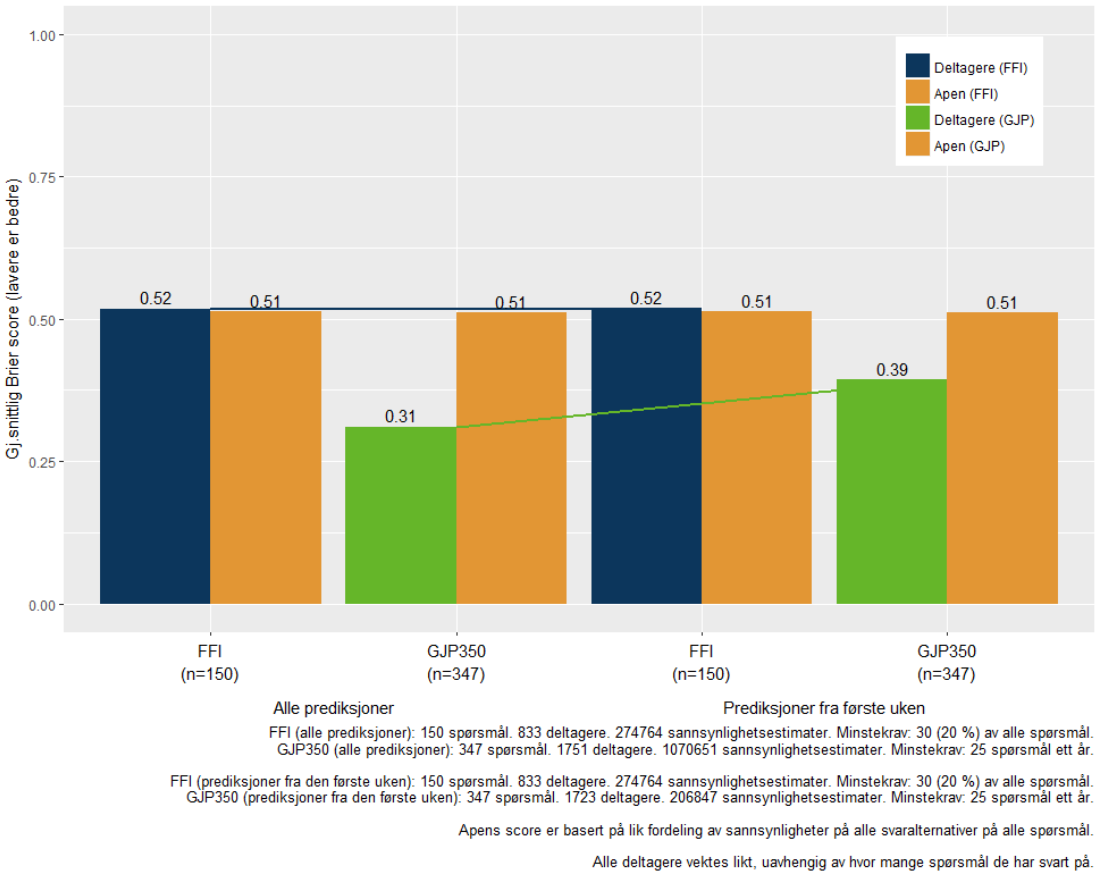
Det forrige underkapittelet viste at treffsikkerheten ikke varierte i tråd med spørsmålenes tidsperspektiver i FFIs eller GJPs turneringer – i alle fall ikke innenfor ett års sikt. Hvor langt frem en blir bedt om å se inn i fremtiden, handler imidlertid også om når en har mulighet til å registrere sine prediksjoner, ikke bare tidsperspektivet spørsmålene i utgangspunktet har.

En annen forskjell mellom FFIs og GJPs turneringer er nettopp tidspunktet deltagerne kunne svare på spørsmålene. Mens FFIs deltagere bare kunne predikere én gang i løpet av den første uken etter at spørsmålet ble stilt, kunne GJPs deltagere velge når og hvor mange ganger de ønsket å predikere helt frem til spørsmålet ble avgjort. Det er antageligvis lettere å treffe mot slutten enn i starten av et spørsmåls tidsperspektiv, f.eks. hvor mange atomprøvesprengninger Nord-Korea vil gjennomføre i 2021. Jo nærmere slutten av året, jo mer informasjon vil du ha om hva antallet kan bli. Det er derfor mulig at gapet mellom treffsikkerheten til deltagerne i FFIs og GJPs turneringer blir mindre hvis det tas høyde for forskjeller i prediksjonstidspunktene.

En første måte å måle betydningen av prediksjonstidspunktet er å sammenligne treffsikkerheten ut fra når deltagerne hadde mulighet til å predikere i de to turneringene: 1) alle prediksjoner, inkludert oppdateringer, som er registrert i løpet av hele spørsmålsperioden, slik som i GJP, og 2) bare de siste prediksjonene som er registrert i løpet av den første uken etter at spørsmålet ble publisert, slik som i FFIs turnering. Mens antallet prediksjoner er det samme i FFIs turnering ved disse to inklusjonskriteriene, reduseres antallet i GJP med tre firedeler når prediksjoner og oppdateringer etter den første uken ikke tas med.[185]

[184] Definisjonene av «kort» og «lang» sikt er basert på artikkelforfatterens personlige korrespondanse med P. Tetlock. Se Muehlhauser, L. (2019), 'How Feasible Is Long-range Forecasting?', *Open Philanthropy*, 10. okt. 2019.

[185] Siden GJPs deltagere kunne oppdatere så mange ganger de ville, er flere deltagere registrert med flere prediksjoner den første uken. Forskjellen mellom å inkludere *alle* prediksjoner eller bare den siste prediksjonen fra den første uken er imidlertid liten. Det totale antallet sannsynlighetsestimater registrert den første uken i GJP350 er 237 915. Hvis vi bare teller den sist registrerte, faller dette antallet til 206 847. Dette har likevel lite å si for den gjennomsnittlige Brier-scoren, som ved begge tellemåter blir 0,39. Selv om deltagerne i FFIs turnering hadde mulighet til å oppdatere sine prediksjoner så lenge spørsmålsuken var åpen, registrerte de aller fleste deltagerne sine prediksjoner i løpet av de første dagene etter at runden ble sendt ut. Bare 15 % av 579 deltagere som besvarte en ekstraundersøkelse mot slutten av turneringen svarte at de åpnet spørsmålsrundene senere for å oppdatere estimatene sine.



Figur 5.6 Deltagernes gjennomsnittlige Brier-scores, basert på prediksjonstidspunkt.

Til venstre i figur 5.6 vises den gjennomsnittlige Brier-scoren til deltagerne i FFIs og GJPs turneringer når alle prediksjoner er inkludert. Dette er de samme verdiene som ble vist i figur 5.1, helt i starten av dette delkapittelet. Til høyre vises snittscorene basert på bare prediksjoner fra den første uken. Figuren viser at treffsikkerheten til deltagerne i GJP faller når den beregnes med utgangspunkt i samme prediksjonstidspunkt som i FFIs turnering. Det er fremdeles en signifikant forskjell mellom treffsikkerheten til deltagerne i de to turneringene, men gapet blir nesten halvert.[186] Brier-scoren til deltagerne i GJP stiger fra 0,31 til 0,39 (SD: 0,15).[187] Omgjort til prediksjoner på et skjer/skjer ikke-spørsmål tilsvarer dette fallet en endring fra 61 % til 56 % sannsynlighet for riktig svar. GJPs deltagere er altså fortsatt mer presise i sine prediksjoner enn FFIs, men forskjellen er mindre. Gapet mellom GJPs deltagere og apen blir også nesten halvert.

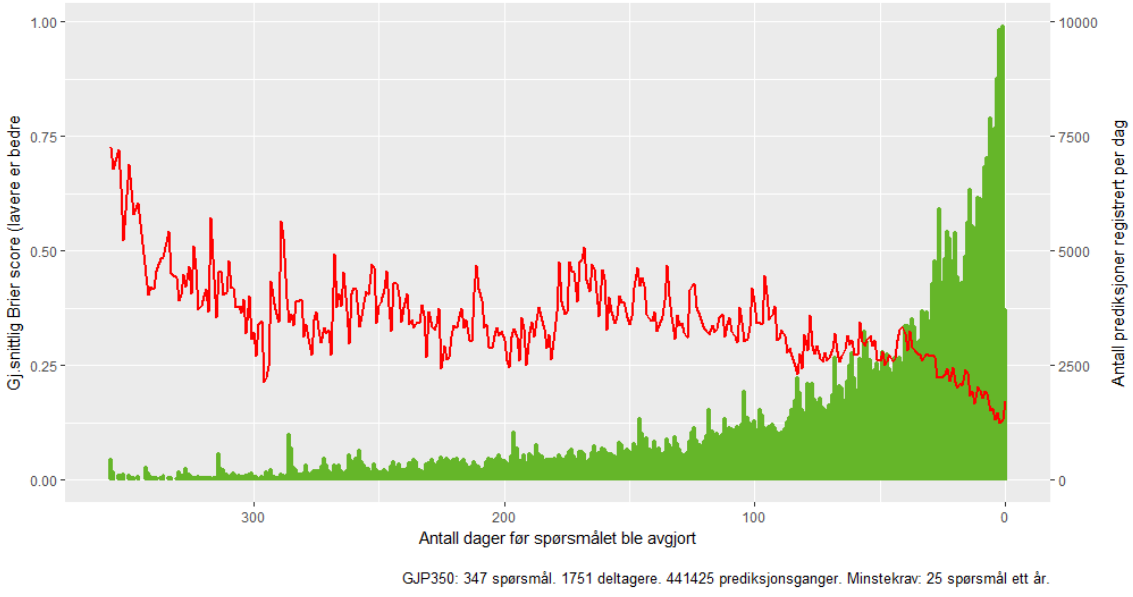
En annen måte å undersøke betydningen av prediksjonstidspunktet på er å se hvordan treffsikkerheten til deltagerne i GJP endret seg *underveis* i spørsmålsperioden. Til forskjell fra spørsmålenes tidsperspektiv defineres spørsmålsperioden her ut fra avgjørelsestidspunktet – altså fra

[186] FFIs vs. GJPs deltagere, basert på prediksjoner fra bare den første uken: $t(2146) = -22.98$, $p < 0.0001$.
[187] Forskjellen mellom disse to snittene i GJPs datagrunnlag er også signifikant. GJPs deltagere basert på alle prediksjoner vs. bare den første uken: $t(1722) = -25.11$, $p < 0.0001$. Siden deltagerne som her sammenlignes kommer fra det samme datasettet, er det gjennomført en paret t-test. Testresultatene er derfor også bare basert på de 1723 av 1751 deltagerne i GJP350 som både predikerte i løpet av og etter den første uken.

tidspunktet spørsmålene ble publisert og frem til de ble avgjort. Dette kan, men må ikke, være det samme tidspunktet som sluttdatoen, som er det seneste tidspunktet et spørsmål kan bli avgjort på og som spørsmålets tidsperspektiv er definert ut fra. For eksempel vil et spørsmål om hvorvidt det vil skje et terrorangrep i Oslo innen utgangen av året kunne avgjøres før fristen, hvis dette faktisk skjer. I FFIs turnering ble 41 (27 %) av de 150 spørsmålene avgjort før sluttdatoen. Andelen er helt lik i GJP, der 95 (27 %) av 347 spørsmål ble avgjort før fristen. Det er likevel små forskjeller mellom spørsmålenes tidsperspektiver og antallet dager det tok før de ble avgjort. I FFIs turnering var tidsperspektivet 372 dager, mens spørsmålene i snitt ble avgjort etter 301 dager. I GJP var tidsperspektivet rundt 130 dager, mens spørsmålene i snitt ble avgjort etter 115 dager.

Her tas det likevel utgangspunkt i avgjørelsestidspunktet, ikke tidsperspektivet, for å se hvordan GJPs deltagere predikerte etter hvert som det nærmet seg dagen spørsmålet ble avgjort, uavhengig av når det etter planen skulle avsluttes. Dette er spesielt interessant å se nærmere på, siden tre firedeler av prediksjonene i GJPs datagrunnlag er registrert etter den første uken og treffsikkerheten i starten på spørsmålsperioden er dårligere enn når den baseres på alle prediksjoner.

Figur 5.7 viser den gjennomsnittlige Brier-scoren (graf) til deltagerne i GJP, rangert etter antall dager det var igjen før spørsmålene faktisk ble avgjort. Det vil si at dag 100 viser snittscoren basert på alle prediksjoner på alle spørsmål registrert 100 dager før hendelsen skjedde eller tiden utløp. For å undersøke når i spørsmålsperioden deltagerne oppdaterte sannsynlighetsvurderingene underveis, viser figuren også antallet nye ganger deltagerne predikerte (søyler) hver dag frem til spørsmålene ble avgjort. Hvis en deltager predikerte 100 og 7 dager før avgjørelsestidspunktet, teller disse som én gang på dag 100 og én gang på dag 7, men ikke på dagene imellom.



Figur 5.7 Deltagernes gjennomsnittlige Brier-scores og antall prediksjoner per dag, basert på deltagernes egne prediksjoner, rangert etter hvor mange dager før spørsmålet ble avgjort de ble registrert.

For det første viser figur 5.7 at deltagerne i GJP predikerte stadig *oftere* jo nærmere de kom dagen spørsmålene ble avgjort. Av de totalt 441 425 gangene deltagerne predikerte, er 331 696 (75 %) av dem registrert i løpet av de siste 100 dagene. Dette er ikke overraskende, siden halvparten av spørsmålene ikke hadde et lengre tidsperspektiv enn 87 dager. Deltagerne fortsatte imidlertid å predikere stadig flere ganger helt frem til spørsmålene ble avgjort, med aller flest prediksjoner registrert i løpet av de tre siste dagene. Hele 215 007 (49 %) av alle gangene deltagerne predikerte skjedde i løpet av de siste 40 dagene. Rundt halvparten av prediksjonene som GJPs resultater på, er derfor prediksjoner registrert i løpet av den siste halvdelen av spørsmålsperioden. I tillegg er de aller fleste av disse bare oppdateringer av tidligere prediksjoner, siden de fleste deltagerne predikerte første gang i løpet av den første uken. Det er ikke overraskende en sterk korrelasjon mellom antall dager til avgjørelsestidspunktet og prediksjoner per dag.[188]

For det andre viser figur 5.7 at deltagerne også predikerte stadig *bedre* mot slutten av spørsmålsperioden. Det er sterke korrelasjoner mellom deltagernes Brier-scores og både antallet oppdateringer per dag (jo flere oppdateringer, jo bedre treffsikkerhet) og antallet dager inntil spørsmålet ble avgjort (jo færre dager til avgjørelsesdagen, jo bedre treffsikkerhet).[189] Mens snittscoren holder seg stabilt frem til det er 100 dager igjen, faller den betydelig etter dette, spesielt de siste 40.

Den gjennomsnittlige Brier-scoren på prediksjoner registrert over 100 dager før avgjørelsestidspunktet er 0,37 (SD: 0,07), altså litt over 0,31 som var snittet for alle prediksjoner uavhengig av prediksjonstidspunkt. Snittscoren basert på de siste 100 dagene er derimot bare 0,24 (SD: 0,06) og 0,22 (SD: 0,05) på prediksjoner fra de siste 40 dagene. Disse snittscorene tilsvarer prediksjoner på hhv. 57 % (over 100 dager igjen), 66 % (siste 100 dager) og 68 % (siste 40 dager) på riktig svar på et skjer/skjer ikke-spørsmål, sammenlignet med 61 % basert på snittet av alle prediksjonene. Forskjellen mellom turneringene blir altså også mindre når vi bare sammenligner med GJPs prediksjoner registrert over 100 dager før spørsmålene ble avgjort.

At deltagerne både traff bedre og predikerte oftere mot slutten er av stor betydning for den gjennomsnittlige Brier-scoren som er rapport i GJPs artikler. Beregningen av hver deltagers score på hvert spørsmål, baserte seg nemlig på snittet av alle prediksjoner, *uavhengig* av tidspunktet det ble predikert. Hvis en deltager predikerte én gang rett etter at spørsmålet ble publisert og oppdaterte denne tre ganger i løpet av de tre siste dagene, ble disse fire prediksjonene vektet likt. Denne måten å beregne deltagernes gjennomsnittlige Brier-score på kan gi et misvisende bilde av treffsikkerheten i GJP, siden de fleste prediksjonene er registrert helt på slutten av spørsmålsperiodene. Dette kan også være en del av forklaringen på hvorfor antall prediksjoner per spørsmål var den variabelen som korrelerte sterkest med den individuelle treffsikkerheten.[190]

[188] Siden antallet dager inntil avgjørelsestidspunktet ikke er normalfordelt, er korrelasjonene i dette underkapittelet bare målt ved Spearmans. Antall prediksjoner per dag vs. antall dager før spørsmålet ble avgjort: $r_s = -0.96$, $p < 0.0001$.

[189] GJPs deltageres Brier-scores vs. antall oppdateringer per dag: $r_s = -0.62$, $p < 0.0001$, og vs. antall dager før spørsmålet ble avgjort: $r_s = 0.65$, $p < 0.0001$.

[190] Se tabell 2 i Mellers (2015), 'The Psychology of Intelligence Analysis', s. 8.

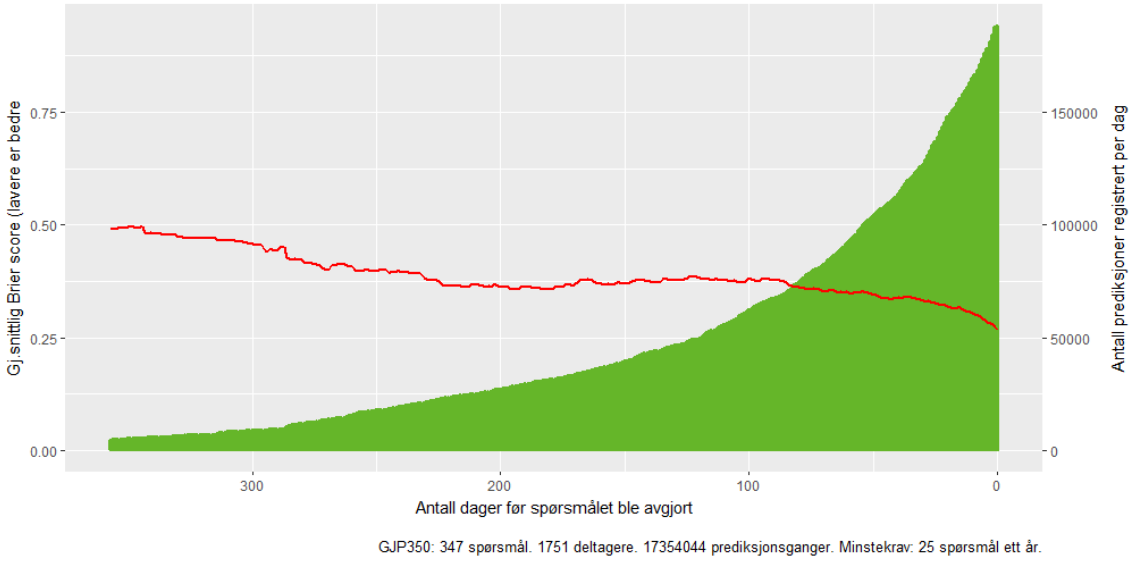
En tredje måte å ta høyde for prediksjonstidspunktet i GJP på, er derfor å beregne en egen Brier-score for hver dag gjennom hele spørsmålsperioden og basere treffsikkerheten på det aktuelle spørsmålet på snittet av alle disse. Her gjøres dette på følgende måte:

1. Etter at en deltager har predikert for første gang, blir alle påfølgende dager registrert med samme prediksjon inntil deltageren oppdaterer denne eller spørsmålet blir avsluttet.

2. På dagene før deltageren predikerte første gang, blir hans daglige score basert på snittet til alle andre deltagere, slik at han ikke premieres eller straffes for å ha predikert tidlig eller sent. Hvis en deltager predikerte 100 og 7 dager før et spørsmål ble avgjort, teller disse som én gang hver på disse dagene, pluss alle dagene før og etter.

3. Treffsikkerheten på hvert spørsmål baseres på snittet av alle daglige Brier-scores. Ved å gjøre dette for alle deltagerne på alle spørsmålene de svarte på, genereres det 17 millioner Brier-scores som kan brukes til å måle utviklingen i treffsikkerhet gjennom spørsmålsperioden.

Figur 5.8 viser den gjennomsnittlige, daglige Brier-scoren (graf) til deltagerne i GJP basert på antall dager det var igjen før spørsmålene ble avgjort. Figuren viser også det totale antallet ganger det er registrert en prediksjon per dag i spørsmålsperioden (søyler). Siden tidligere prediksjoner nå regnes med i alle påfølgende dager, viser søylene det akkumulerte antallet prediksjoner som grafen med de gjennomsnittlige, daglige Brier-scorene er basert på.



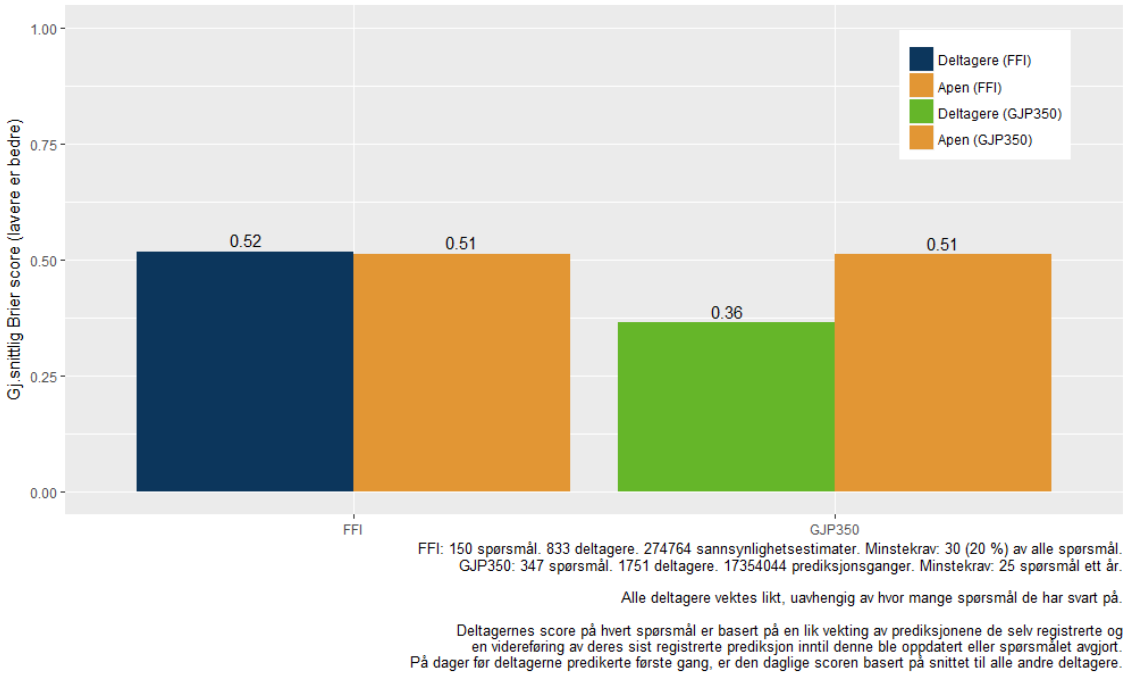
Figur 5.8 Deltagernes gjennomsnittlige Brier-scores og antall prediksjoner per dag, basert på deltagernes egne prediksjoner og en daglig videreføring av disse, rangert etter på hvor mange dager før spørsmålet ble avgjort de ble registrert.

For det første viser figur 5.8 at Brier-scoren til deltagerne i GJP ligger generelt høyere og ikke faller like brått mot slutten som i figur 5.7, nå som deltagernes tidligere prediksjoner også tas med i beregningen og treffsikkerheten baseres på prediksjoner fra hele spørsmålsperioden. Snittscoren på prediksjoner registrert frem til det er 100 dager igjen er 0,41 (SD: 0,04), altså litt høyere enn scoren på 0,37 i analysen over, som bare baserte seg på deltagernes nye prediksjoner underveis. Her blir også snittscorene fra de siste 100 og 40 dagene 0,34 (SD: 0,03) og 0,32 (SD: 0,02), som er betydelig høyere enn 0,24 og 0,22 i analysen over.

Disse nye Brier-scorene tilsvarer prediksjoner på hhv. 55 % (over 100 dager igjen), 59 % (siste 100 dager) og 60 % (siste 40 dager) på det riktige av to mulige utfall, sammenlignet med 57 %, 66 % og 68 % basert på GJPs måte å beregne treffsikkerheten på. Dette bekrefter at gapet mellom FFIs og GJPs deltagere reduseres betydelig når prediksjonstidspunktet tas høyde for.

For det andre viser figur 5.8 at antallet prediksjoner som snittscorene baseres på øker mer gradvis nå som hver dag i spørsmålsperiodene telles med, ikke bare dagene deltagerne selv predikerte. At antallet prediksjoner øker mot slutten også i denne figuren, skyldes imidlertid ikke at deltagerne predikerte oftere mot slutten, men at flere spørsmål med stadig kortere tidsperspektiv blir inkludert. Det betyr at en del av forklaringen på den raske økningen i antall nye prediksjoner mot slutten i forrige figur ikke bare handlet om at deltagerne predikerte på nytt, men også at det var stadig flere spørsmål inkludert.

Det er ikke mulig å sammenligne utviklingen i treffsikkerheten til deltagerne i GJP med resultatene fra FFIs turnering på grunn av forskjellen i når deltagerne kunne predikere. Det er likevel mulig å beregne en ny snittscore for deltagerne i GJP, basert på den ovenstående måten å måle treffsikkerheten deres på, der deltagerne ikke kan få en kunstig høy treffsikkerhet ved å predikere flere ganger mot slutten av spørsmålsperioden. Figur 5.9 den gjennomsnittlige treffsikkerheten til deltagerne i de to turneringene, basert på daglige Brier-scores i GJP.



Figur 5.9 Deltagernes gjennomsnittlige Brier-scores, basert på deltagernes egne prediksjoner og en daglig videreføring av disse gjennom hele spørsmålsperioden.

Figur 5.9 viser at når treffsikkerheten baseres på daglige prediksjoner blir den gjennomsnittlige Brier-scoren til deltagerne i GJP 0,36 (SD: 0,09). Forskjellen mellom FFIs og GJPs deltagere er fremdeles signifikant, men dette er en dårligere score enn 0,31, som er den GJP oppgir i sine artikler.[191] Samtidig er GJPs nye score litt bedre enn snittet på 0,39, basert på bare prediksjoner registrert bare i løpet av den første uken, slik som i FFIs turnering. Den nye snittscore til GJPs deltagere tilsvarer det å oppgi en 58 % sannsynlighet for det riktige utfallet på et skjer/skjer ikke-spørsmål.

Til sammen viser analysene at deltagerne i GJP blir dårligere og gapet opp til FFIs mindre når treffsikkerheten baseres på prediksjoner fra bare den første uken etter at spørsmålene ble publisert eller når scoren baseres på alle dagene i spørsmålsperiodene i stedet for bare de dagene deltagerne selv valgte å predikere. Likevel forblir GJPs deltagere bedre enn FFIs, også når det tas høyde for forskjellene i prediksjonstidspunkt. Det kan imidlertid stilles spørsmål om forskjellen mellom deltagernes treffsikkerhet fortsatt er betydningsfull.

[191] FFIs vs. GJPs deltagere, basert på daglige Brier-scores: $t(1315) = -34.57$, $p < 0.0001$.

5.1.6 Eksperimentgrupper

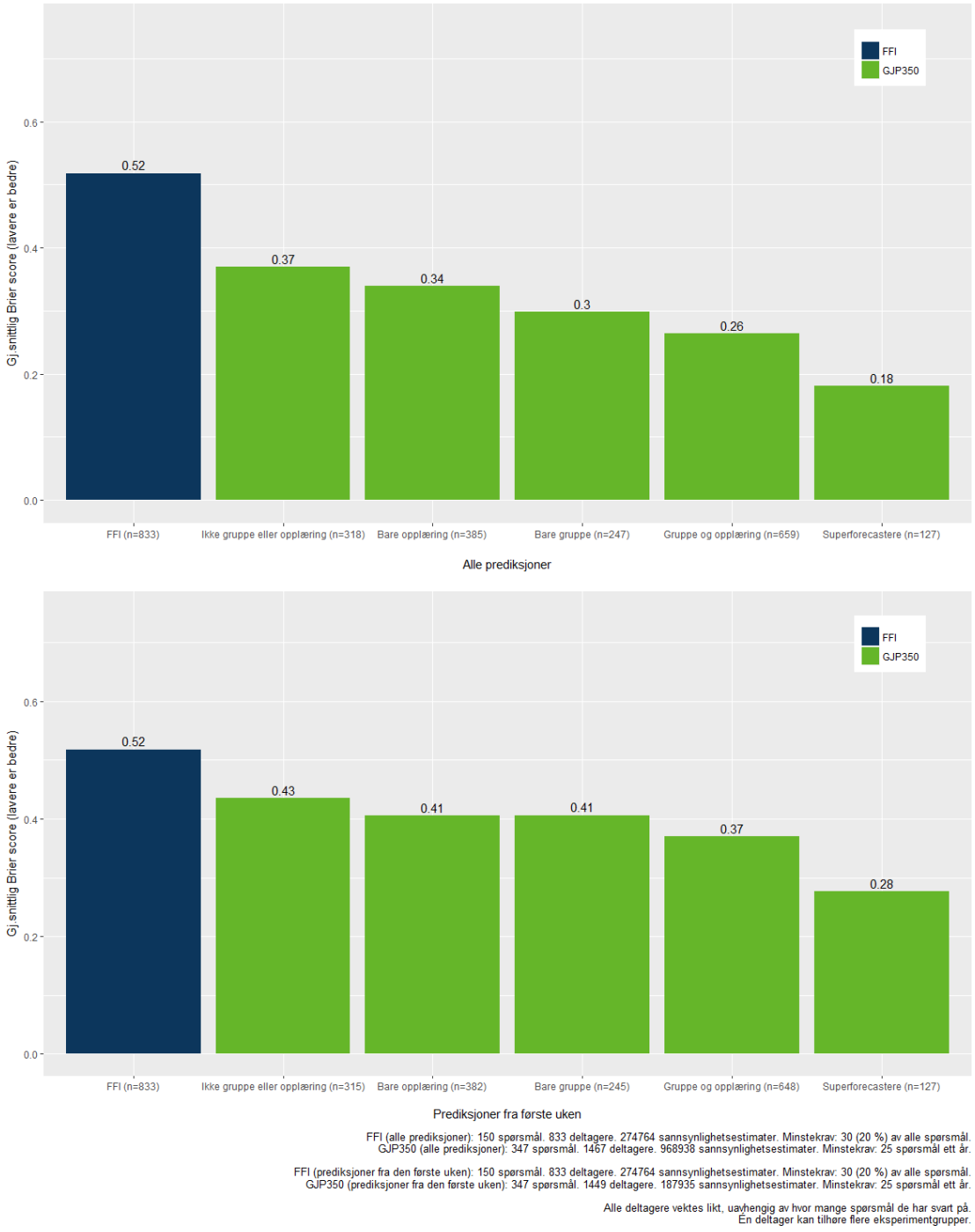
En annen viktig forskjell mellom gjennomføringene av turneringene er at det i GJPs ble eksperimentert med tiltak for å forbedre deltagernes treffsikkerhet underveis. Det var særlig tre tiltak som bidro til å øke deltagernes prediksjonsevne.[192] Det første var å gi deltagerne *opplæring* i probabilistisk tenkning. Det andre var å sette deltagerne sammen i *grupper* som kunne diskutere prediksjoner og som konkurrerte mot andre lag. De fleste deltagerne ble fordelt på eksperimentgrupper som enten ble gitt opplæring, satt i grupper, begge eller ingen av delene. Det tredje og viktigste tiltaket var identifiseringen av *superforecastere* etter hvert turneringsår, som fikk både opplæring i probabilistisk tenkning og ble satt i grupper med andre superforecastere.

I FFIs turnering er det ikke gjort noen forsøk på å forbedre treffsikkerheten underveis. Det kan derfor tenkes at fraværet av forbedringstiltak kan være en del av forklaring på hvorfor FFIs deltagere treffer dårligere. Den eneste eksperimentgruppen som er direkte sammenlignbare med FFIs deltagere er dem som hverken fikk opplæring eller ble satt i grupper.

Figur 5.10 sammenligner derfor treffsikkerheten til FFIs deltagere med hver av de fem eksperimentgruppene i GJP: 1) individuelle deltagere uten opplæring, 2) individuelle deltagere med opplæring, 3) deltagere i grupper uten opplæring, 4) deltagere i grupper med opplæring, og 5) superforecastere i grupper med opplæring.[193] Siden prediksjonstidspunktet har vist seg å redusere gapet mellom turneringene, skilles det også her mellom treffsikkerheten til eksperimentgruppene basert på alle prediksjoner (øverst) og prediksjoner fra bare den første uken (nederst).

[192] De to første tiltakene utgjør de situasjonelle variablene som ble målt i GJP200-artikkelen. Det første året inkluderte opplæringen også scenariotenkning, men denne ble ikke videreført fra det andre året, fordi det var teknikkene for probabilistisk tenkning som hadde størst effekt. Identifiseringen og bruken av superforecastere er beskrevet i GJP350-artikkelen. En siste eksperimentgruppe var deltagere som ble satt i prediksjonsmarkeder, men dette tiltaket er ikke analysert i noen av artiklene som resultatene fra FFIs turnering sammenlignes med i denne rapporten.

[193] Her inkluderes bare deltagere fra GJP som ble var registrert i én av de fem eksperimentgruppene beskrevet her i løpet av turneringens tre første år. Deltagere som ble plassert i prediksjonsmarkeder, som også viste seg å øke treffsikkerheten, er ikke inkludert her, fordi dette tiltaket ikke ble analysert i GJP200- eller GJP350-artiklene. Disse utgjør imidlertid bare en liten andel av deltagerne. I tillegg ble det gjennomført flere andre eksperimenter det fjerde året, men disse ligger utenfor perioden som studeres her. Til sammen inkluderer utvalget som analyseres her 1467 (84 %) av alle 1751 deltagerne i GJP350. 261 av disse deltagerne ble flyttet fra én eksperimentgruppe til en annen i løpet av turneringen, f.eks. fra en gruppe uten opplæring til en med. Disse er derfor inkludert i mer enn én eksperimentgruppe.



Figur 5.10 Deltagernes gjennomsnittlige Brier-scores, basert på eksperimentgruppe.

Når Brier-scorene baseres på alle prediksjoner viser resultatene at alle eksperimentgruppene i GJP traff signifikant bedre enn FFIs deltagere.[194] Konvertert til prediksjoner på et skjer/skjer

[194] Når Brier-scorene baseres på alle prediksjoner er resultatene fra t-tester av forskjellene mellom snittscoren til FFIs deltagere og hver av de fem eksperimentgruppene hhv.: 1) $t(580) = -19.978$, 2) $t(818) = -27.342$,

ikke-spørsmål tilsvarer snittscorene til GJPs fem eksperimentgrupper det å oppgi hhv. 57 %, 59 %, 61 %, 64 % og 70 % på riktig utfall. Deltagerne som predikerte helt alene oppnådde altså en snittscore som tilsvarer en prediksjon på riktig svar som var åtte prosentpoeng bedre enn FFIs deltagere (49 %), selv om disse ikke er direkte sammenlignbare gitt den ulike fordelingen av typer spørsmål i de to turneringene.

Samtidig bekreftes det nok en gang at gapet mellom FFIs og GJPs deltagere reduseres når treffsikkerheten baseres på prediksjoner fra den første uken. På den ene siden traff alle eksperimentgruppene fortsatt signifikant bedre enn FFIs deltagere, inkludert deltagerne som predikerte helt alene.[195] På den annen side faller treffsikkerheten til de fem eksperimentgruppene så mye at de nye snittscorene tilsvarer prediksjoner på hhv. 54 %, 55 %, 55 %, 57 % og 63 % på riktig svar. GJPs superforecastere treffer imidlertid fortsatt mye bedre enn alle andre.

5.1.7 Treffprosent

De probabilistiske prediksjonene som er samlet inn i FFIs og GJPs turneringer gjør det også mulig å måle treffsikkerheten til deltagerne på andre måter enn med Brier-score.

De fleste forbinder treffsikkerhet med *treffprosent*, altså hvor *ofte* vi velger det riktige svaret, uavhengig hvilke sannsynlighetsvurderinger vi la til grunn. I denne rapporten er treffprosent definert som andelen spørsmål der en deltager oppgir den høyeste sannsynligheten til det riktige svaralternativet. Hvis en deltager oppgir en sannsynlighet på minst 51 % for riktig svar, regnes dette som et helt treff (1), uavhengig av om estimatene er 60 % eller 90 %. På kategoriske og ordinale spørsmål, der prosentene kan fordeles på mer enn to alternativer, regnes det også som et helt treff (1), hvis deltagerne oppgir høyest sannsynlighet til riktig svar, selv om estimatet er under 50 % (f.eks. A: 40 %, B: 30 % og C: 30 %, der A er riktig). Hvis en deltager oppgir like høy sannsynlighet på andre svar som på det riktige, deles treffandelen på antallet svaralternativer. Helt lik fordeling av sannsynlighetene på to svaralternativer (50/50 %) regnes som et halvt treff (0,5). Lik fordeling på fire svaralternativer (f.eks. A: 25 %, B: 25 %, C: 25 % og D: 25 %) regnes som et firedels treff (0,25). Hver deltagers treffprosent er basert på snittet av alle treffandelene på alle spørsmål de svarte på.

Tabell 5.1 viser treffprosentene til FFIs (blå) og GJPs (grønn) deltagere, inkludert hver av de fem eksperimentgruppene, på alle spørsmålstyper og ved de to ulike prediksjonstidspunktene.

3) $t(503) = -31.758$, 4) $t(1466) = -45.365$, og 5) $t(409) = -59.277$. Alle disse forskjellene er signifikante på 0.0001-nivå. Det samme er alle forskjeller mellom hver av GJPs eksperimentgrupper.

[195] Når Brier-scorene baseres på prediksjoner fra den første uken er forskjellene mellom snittscoren til FFIs deltagere og de fem eksperimentgruppene fortsatt signifikante på 0.0001-nivå: 1) $t(475) = -9.365$, 2) $t(622) = -13.852$, 3) $t(324) = -10.462$, 4) $t(1196) = -21.265$, og 5) $t(239) = -32.586$. Alle forskjeller mellom eksperimentgruppene er også signifikante på minst 0.05-nivå, unntatt mellom dem som fikk opplæring eller ble satt i grupper.

Deltagere (antall per pred.tidspunkt)	**Alle prediksjoner**	**Prediksjoner første uken**
FFI (833)	-	51 % - Binære: 68 % - Kategoriske: 44 % - Ordinale: 43 %
GJP350 (hhv. 1467 og 1449)	76 % - Binære: 79 % - Kategoriske: 64 % - Ordinale: 67 %	69 % - Binære: 71 % - Kategoriske: 61 % - Ordinale: 52 %
- Individuelle deltagere uten opplæring (hhv. 318 og 315)	71 % - Binære: 73 % - Kategoriske 61 % - Ordinale: 60 %	65 % - Binære: 68 % - Kategoriske 53 % - Ordinale: 51 %
- Individuelle deltagere med opplæring (hhv. 385 og 382)	74 % - Binære: 63 % - Kategoriske 63 % - Ordinale: 65 %	68 % - Binære: 70 % - Kategoriske 60 % - Ordinale: 49 %
- Deltagere i grupper uten opplæring (hhv. 247 og 245)	77 % - Binære: 79 % - Kategoriske 70 % - Ordinale: 68 %	67 % - Binære: 69 % - Kategoriske 65 % - Ordinale: 53 %
- Deltagere i grupper med opplæring (hhv. 659 og 648)	80 % - Binære: 83 % - Kategoriske 67 % - Ordinale: 70 %	70 % - Binære: 73 % - Kategoriske 64 % - Ordinale: 54 %
- Superforecastere i grupper med opplæring (hhv. 127 og 127)	86 % - Binære: 88 % - Kategoriske 84 % - Ordinale: 78 %	78 % - Binære: 81 % - Kategoriske 80 % - Ordinale: 66 %

Tabell 5.1 Treffprosenten til FFIs og GJPs deltagere, inkludert eksperimentgrupper, basert på spørsmålstype og prediksjonstidspunkt.[196]

[196] Beregningen av treffprosentene i GJP er basert på en lik vekting av alle deltagernes gjennomsnittlige treffprosent på alle spørsmål i replikasjonsdatasettet til GJP350. Snittet på hvert spørsmål er bare basert på deltagernes egne prediksjoner og eventuelle oppdateringer, ikke en videreføring av prediksjoner på dagene imellom. Det oppgis ingen treffprosent i GJP350-artikkelen. GJP200-artikkelen oppgir derimot at treffprosenten til deltagerne var 75 % og 47 % ved tilfeldig gjetning (Mellers (2015), 'The Psychology of Intelligence Analysis', s. 6), som er tilnærmet identisk treffprosentene beregnet her basert på GJP350-datasettet. Ifølge GJP200-artikkelen ble imidlertid treffprosenten på hvert spørsmål beregnet ved å dele andelen dager de traff på antall dager de hadde predikert. Etter at en deltager

I FFIs turnering ligger den foreløpige treffprosenten på 51 %. Det vil si at deltagerne i snitt har tildelt den høyeste sannsynligheten til det riktige svaret på halvparten av spørsmålene. Til sammenligning er apens treffprosent bare 33 %. Treffprosenten i GJP var derimot 76 %, mens treffprosenten ved tilfeldig gjetning var 47 %. Mens deltagerne i både FFIs og GJPs turneringer slår apen når prediksjonsevnen blir målt i treffprosent, er det fremdeles et gap på 10 prosentpoeng mellom deltagerne i de to turneringene. Denne forskjellen er også statistisk signifikant.[197]

At treffprosenten til apen var høyere i GJP enn i FFIs turnering skyldes forskjellen i typer spørsmål. I GJP reflekterer apens treffprosent (47 %) at de fleste spørsmålene var binære, som alltid gav en treffprosent på 50 % ved lik fordeling av sannsynlighetene på to svaralternativer. I FFIs turnering var det rundt tre svaralternativer per spørsmål som forklarer apens treffprosent (33 %).

Den lavere gjennomsnittlige treffprosenten til deltagerne i FFIs turnering enn i GJP kan derfor tenkes å skyldes det samme, nemlig at deltagerne fordelte sannsynlighetene jevnere fordi de fikk flere svaralternativer å velge mellom. Dette kan undersøkes ved å sammenligne treffprosentene på binære spørsmål, der det alltid er bare to svaralternativer. Her er treffprosenten til FFIs deltagere fortsatt lavere (68 %) enn i GJP (79 %). Gapet i treffprosenten blir altså redusert fra 25 til 10 prosentpoeng når vi avgrenser analysen fra alle til binære spørsmål, men FFIs deltagere treffer fremdeles dårligere enn GJPs på samme type.

Siden tidspunktet deltagerne predikerte har vist seg å påvirke treffsikkerheten til deltagerne i GJP, er det også relevant å måle treffprosentene deres basert på prediksjoner fra den første uken. Da faller GJPs treffprosent fra 76 % til 69 %. Dette snittet er fremdeles høyere enn FFIs 51 % og forskjellen er fortsatt statistisk signifikant.[198] På binære spørsmål faller imidlertid treffprosenten i GJP fra 79 % til 71 %, som er nesten helt lik FFIs på 68 %. Denne forskjellen er også signifikant,[199] men i praksis er gapet mellom turneringenes treffsikkerhet nå nesten helt borte.

Gapet mellom treffprosentene i turneringene reduseres ikke like mye på kategoriske og ordinale spørsmål. Mens treffprosentene i FFIs turnering er 44 % på kategoriske og 43 % på ordinale, var treffprosentene i GJP 64 % på kategoriske og 67 % på ordinale. Her er forskjellene i prosentpoeng omtrent dobbelt så store som på binære spørsmål. Gapet reduseres heller ikke like mye som på de binære spørsmålene om vi bare baserer oss på prediksjoner fra den første uken. GJPs treffprosent faller riktig nok fra 64 % til 61 % på kategoriske og fra 67 % til 52 % på ordinale, men disse er fremdeles betydelig høyere enn FFIs på 44 % og 43 %.

At gapet forblir større på de kategoriske og ordinale spørsmålene kan heller ikke tilskrives at FFIs spørsmål hadde flere svaralternativer på de kategoriske og ordinale spørsmål, fordi det er

hadde predikert, ble estimatet stående alle dager inntil prediksjonen ble endret eller spørsmålet avsluttet. Treffprosenten var altså andelen dager deltagerne hadde oppgitt høyest sannsynlighet til riktig svar på alle spørsmål. Når treffprosenten beregnes basert på lik vekting av prediksjonene registrert uavhengig av dagene i spørsmålsperioden, som gjort med replikasjonsdatasettet til GJP350 her, blir treffprosenten i GJP200 også 75 %, som i artikkelen. Enten er beskrivelsen av beregningen av treffprosent i artikkelen feil eller så utgjør bruken av daglige treffprosenter ingen forskjell.

[197] FFIs vs. GJPs deltagere, basert på treffprosent: $t(2047) = -72.08$, $p < 0.0001$.

[198] FFIs vs. GJPs deltagere, basert på treffprosent og prediksjoner fra første uken: $t(2537) = -40.87$, $p < 0.0001$.

[199] FFIs vs. GJPs deltagere, basert på treffprosent på binære spørsmål og prediksjoner fra første uken: ($t(2054) = -5.51$, $p < 0.0001$.

ikke store forskjeller i antall svaralternativer per spørsmål og forskjellene i treffprosent beskrevet over holder seg uansett antall alternativer.[200] I FFIs turneringer er den gjennomsnittlige treffprosenten 54 % på spørsmål med tre svaralternativer, 43 % på fire svaralternativer og 31 % på fem svaralternativer.[201] I GJPs turnering var treffprosenten på de samme antallene svaralternativer hhv. 62 %, 71 % og 65 %, altså mellom 19 og 34 prosentpoeng høyere enn FFIs. Basert på bare prediksjoner fra den første uken faller GJPs treffprosenter med omtrent ti prosentpoeng, til hhv. 54 %, 61 % og 58 %. Fortsatt treffer FFIs deltagere dårligere enn GJPs, basert på samme prediksjonstidspunkt og de fleste antall svaralternativer.[202]

Bildet er det samme om vi sammenligner treffprosentene til deltagerne i FFIs turnering med de forskjellige eksperimentgruppene i GJP. Selv ved likt prediksjonstidspunkt oppnår den dårligste gruppen, som var individuelle deltagerne uten opplæring, en treffprosent som er 14 prosentpoeng høyere enn FFIs. På binære spørsmål forsvinner gapet igjen, men på kategoriske og ordinale spørsmål fortsetter GJPs deltagere å være systematisk mer treffsikre enn FFIs. Nok en gang skiller superforecasterne seg ut ved en betydelig høyere treffprosent på tvers av alle spørsmålstyper.

Oppsummert betyr de reduserte forskjellene når treffsikkerheten måles i treffprosent at deltagernes evner til å forutsi *hvilket* utfall som ville skje var likere og bedre enn apens i begge turneringene. Det vedvarende gapet i Brier-score betyr samtidig at FFIs deltagere var vesentlig dårligere enn GJPs til å oppgi høye *sannsynligheter* til utfallene som skjedde og lave til dem som ikke skjedde. Selv om treffsikkerheten i FFIs turnering ikke er så mye dårligere som en først kan få inntrykk av, er treffsikkerheten likevel ikke imponerende. Selv om FFIs deltagere nå blir bedre enn tilfeldig gjetning, har de likevel en treffprosent på bare 51 % på spørsmål om temaer av relevans for norsk sikkerhet med rundt tre–fire svaralternativer hver.

I likhet med Brier-scoren måler treffprosenten også evnen til å skille mellom hvilke hendelser som skjer og ikke, som er spesielt viktig i forsvars- og sikkerhetspolitiske vurderinger. Til forskjell fra Brier-scoren reflekterer imidlertid ikke treffprosenten usikkerheten ved prediksjonene. En deltager som er relativt usikker og oppgir 60 % sannsynlighet på alle riktige svar vil få den samme treffprosenten som en annen deltager som er 90 % sikker på de samme spørsmålene. Treffprosenten gjør det også enklere å slå tilfeldig gjetning, siden en i motsetning til Brier-score ikke blir straffet spesielt hardt for å oppgi høyere sannsynligheter på utfall som viser seg å være feil, selv om det kan hevdes at dette er spesielt farlig i spørsmål om nasjonal sikkerhet. Det kan derfor diskuteres hvilket av disse to målene på treffsikkerhet som er det best egnede.

[200] Det er i snitt 4,1 svaralternativer på kategoriske og 4,2 på ordinale i FFIs turnering mot 3,8 og 3,7 i GJP350.
[201] Treffprosenten ved tilfeldig gjetning vil her være 33,3 % på tre svaralternativer, 25 % på fire og 20 % på fem.
[202] Et forbehold ved sammenligningene i dette avsnittet er at det er store forskjeller i antallet spørsmål av forskjellige typer i de respektive datagrunnlagene. Mens FFIs består av 43 binære mot 33 kategoriske og 74 ordinale, består replikasjonsdatasettet til GJP350 som er analysert her, av 280 binære, 24 kategoriske og 45 ordinale. Samtidig ser det ikke ut som treffprosenten endrer seg betydelig selv om antallet spørsmålet som analyseres øker. I GJP200 var nemlig treffprosenten 77 % på binære spørsmål, 71 % på kategoriske og 65 % på ordinale, sammenlignet med 78 %, 64 % og 67 % i GJP350, selv om det sistnevnte datasettet består av rundt 150 flere spørsmål.

5.1.8 Kalibrering

Det store gapet mellom treffprosenten og sannsynlighetene deltagerne tildelte det de trodde var riktig svar, betyr at deltagerne i FFIs turnering er altfor selvsikre. Dette er i utgangspunktet ikke overraskende. En av de vanligste fallgruvene ved prediksjon er overkonfidens, der vi tror vi er bedre enn vi egentlig er.[203] Denne fallgruven forklarer hvorfor investorer til stadighet satser og taper store penger og mange byggeprosjekter overskrider tids- og kostnadsrammene sine. I forsvarssammenheng kan overkonfidens få spesielt katastrofale konsekvenser. Historikere har blant annet beregnet at den angripende part har tapt krigen i 50–75 % av tilfellene siden år 1500, og at de som vant har opplevd seieren som mer dyrekjøpt enn forventet.[204]

Et siste mål på treffsikkerheten i FFIs turnering er å se på deltagernes *kalibrering*. Når vi sammenligner sannsynlighetsestimatene deltagerne oppgav på det de *trodde* var riktig svar (konfidens) med hvor ofte de *faktisk* traff (treffprosent), kan vi måle hvor godt kalibrerte de er. Hvis en deltagers snittprediksjon på antatt riktig svar er 70 % og treffprosenten på de samme spørsmålene også er 70 %, er deltageren perfekt kalibrert. Hvis snittprediksjonen er lavere enn treffprosenten er han *under*konfident, mens hvis snittprediksjonen er høyere enn treffprosenten er han *over*konfident. Å vite hvor sikker man kan være på prediksjonene som gjøres er spesielt viktig når usikkerheten er stor, som ved forsvars- og sikkerhetspolitiske spørsmål.

I snitt oppgir deltagerne i FFIs turnering en 72 % sannsynlighet for det de trodde var riktig svar, men treffprosenten deres er bare 51 %. Dette er en overkonfidens på hele 21 %. Selv på bare binære spørsmål er overkonfidensen 15 %. Dette er en enda mer overdreven selvtillit enn ekspertene i EPJ, der overkonfidensen var 12 %.[205] I EPJ var overkonfidensen 20 % blant eksperter som predikerte langsiktige spørsmål innenfor egne ekspertiseområder, men den var 3 % blant eksperter som predikerte kortsiktige spørsmål i domener utenfor egne ekspertiseområder. I GJP var overkonfidensen derimot bare 2,7 %.[206] Selv basert på prediksjoner fra bare den uken er overkonfidensen i GJP bare 4,1 %.[207] FFIs deltagere har altså en like overdreven selvtillit som de dårligste ekspertene i EPJ og er langt mer overkonfidente enn deltagerne i GJP.

Ved godt kalibrerte personer, med en liten grad av over- eller underkonfidens, vil snittprediksjonene deres gi en god indikasjon på hvor ofte vi kan forvente at de faktisk treffer. Samtidig kan

[203] For en oppsummering av forskningen på overkonfidens, se introduksjonen i Moore mfl. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition'.
[204] Se Johnson, D. D. P. (2004), *Overconfidence and war: The havoc and glory of positive illusions* (Cambridge, MA: Harvard University Press). Se også omtalen av boken i Freedman, L. (2005), 'Overconfidence and War: The Havoc and Glory of Positive Illusions', *Foreign Affairs*, 84:2, s. 154.
[205] Se Moore mfl. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', ss. 3561–3562. Overkonfidensen var 20 % blant eksperter som predikerte langsiktige spørsmål innenfor egne ekspertiseområder, men den var 3 % blant eksperter som predikerte kortsiktige spørsmål i domener utenfor egne ekspertiseområder.
[206] Basert på replikasjonsdatasettet til GJP350, der deltagernes gjennomsnittlig prediksjon på det de trodde var riktig svar 78,5 %, mens treffprosenten var 76,4 %. Graden av overkonfidens i GJP er bare rapportert i Moore mfl. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', som i likhet med GJP350 er basert på de tre første årene av turneringen, men inkluderte flere deltagere og prediksjoner enn dem som er analysert her. Her rapporteres det at deltagerne i snitt var 65,4 % sikre på at de predikerte riktig, mens de i virkeligheten hadde rett 63,3 % av gangene, som betød en overkonfidens på 2,1 % – altså ganske lik overkonfidensen som er beregnet her.
[207] Basert på replikasjonsdatasettet til GJP350. Ved å bare se på prediksjoner fra den første uken faller deltagernes gjennomsnittlige snittprediksjon på riktig svar fra 79 % til 73 %, men treffprosenten faller fra 76 % til 69 %.

perfekt kalibrerte prediksjoner være helt unyttige, hvis de sjeldent treffer. Perfekt kalibrerte personer med snittprediksjoner på rundt 50 % på riktig utfall på binære spørsmål vil ikke kunne skille mellom hendelser som skjer og ikke. Målet er å ha høyest mulig konfidens og treffprosent samtidig. GJPs deltagere var ikke bare godt kalibrerte, men hadde en treffprosent på 69 %, selv når denne ble basert på én prediksjon helt i spørsmålsperioden, slik som i FFIs turnering.

I de påfølgende kapitlene vil treffprosent og kalibrering brukes til å diskutere hvor godt eller dårlig ulike deltagergrupper traff utover bare Brier-scorene. Kalibreringen er særlig nyttig for å si noe om hva som er feil med deltagernes treffsikkerhet, basert på hvor mye de over- eller undervurderte sin egen prediksjonsevne. En deltagergruppe med høy treffprosent, men også stor grad av overkonfidens, betyr at vi kan anta at de fleste prediksjonene vil peke på riktig utfall, men samtidig at vi bør behandle dem mer varsomt enn om gruppen hadde vært bedre kalibrert.

5.1.9 Diskusjon

De foreløpige resultatene fra FFIs turnering nyanserer de optimistiske funnene fra GJP. Etter at de 150 første spørsmålene er deltagerne i FFIs turnering akkurat like gode til å oppgi høye sannsynligheter til hendelsene som skjer og lave til de som ikke gjør det som den pilkastende apen med bind for øynene, der alle sannsynlighetene fordeles helt likt på alle de samme spørsmålene. Deltagerne i GJP var derimot langt bedre enn både FFIs og apens.

Selv om det er betydelige forskjeller i spørsmålstemaene i FFIs og GJPs turneringer, hvor mange svaralternativer deltagerne fikk og hvor langt frem de ble bedt om å predikere, ser ikke disse forskjellene ut til å forklare den relative forskjellen i deltagernes treffsikkerhet. Tvert imot sliter FFIs deltagere med å slå apen og treffer dårligere enn GJPs på tvers av alle sammenlignbare temaer, spørsmålstyper og tidsperspektiver.

Et overraskende funn er at det ikke i noen av turneringene finnes en sammenheng mellom treffsikkerheten og tidsperspektivet på spørsmålene. Selv på spørsmål som ser mindre enn et halvt år fremover, treffer FFIs deltagere like dårlig som apen. Samtidig ser det ut til at deltagerne blir hverken bedre eller dårligere på spørsmål som ser lenger frem enn dette. Selv om deltagerne i GJP generelt traff mye bedre finnes det heller ikke i denne turneringen noen sammenheng med spørsmålenes tidsperspektiv. Samlet sett er det altså lite som tilsier at det er forskjeller i selve spørsmålsgrunnlaget som kan forklare hvorfor deltagerne i FFIs turnering treffer dårligere.

Det ser derimot ut til å være sammenhenger med hvordan FFIs og GJPs turneringer ble gjennomført. Når treffsikkerheten til deltagerne i GJP bare baseres på prediksjoner fra den første uken, slik som i FFIs turnering, halveres nesten gapet mellom snittscorene i de to turneringene. Det samme gjelder alle de fem eksperimentgruppene i GJP. Alle gruppene, inkludert individer som predikerte helt alene som i FFIs turnering, er fortsatt signifikant bedre, men forskjellene mellom turneringene reduseres med mellom en tredel og halvparten når prediksjonstidspunktet blir det samme. I tillegg gjør den opprinnelige måten treffsikkerheten er beregnet på i GJP et noe misvisende bilde av hvor gode deltagerne var til å forutsi utfall fremover i tiden, siden de fleste prediksjonene er registrert helt på slutten av spørsmålsperiodene.

Forskjellen mellom deltagerne i FFIs og GJPs turneringer blir også mindre når treffsikkerheten måles ved treffprosent i stedet for Brier-score. Da slår både FFIs og GJPs deltagere apen med god margin. Gapet mellom turneringene viskes også nesten helt ut på binære spørsmål basert på prediksjoner fra den første uken alene. Forskjellen reduseres imidlertid ikke like mye på kategoriske og ordinale spørsmål, selv med samme antall svaralternativer og prediksjonstidspunkt.

I tillegg forblir FFIs deltagere langt mer overkonfidente enn GJPs deltagere, uansett prediksjonstidspunkt. En mulig forklaring kan være at spørsmålene i FFIs turnering var kvalitativt vanskeligere enn dem som ble stilt i GJP. Dette er imidlertid vanskelig å måle, både i forkant og etterkant. Hvis spørsmålene var vanskeligere har det i så fall ikke ledet FFIs deltagere til å bli mer forsiktige i prediksjoner. FFIs deltagere oppgav nemlig like høye sannsynligheter til det de *trodde* var de riktige svarene. De bommet bare rett og slett oftere enn GJPs. FFIs deltagere er til og med dårligere kalibrerte enn selv den dårligste eksperttypen i EPJ.

Selv om deltagerne i FFIs turnering ikke predikerte som om spørsmålene er vanskeligere enn i GJP, er det likevel mulig at de er det. Én måte dette kan undersøkes nærmere i videre studier, er å kategorisere spørsmålene etter type hendelser deltagerne ble bedt om å predikere, f.eks. valgresultater, kursutvikling og militær maktbruk. Hvis FFIs turnering har en større andel spørsmål om hendelsestyper der deltagerne i begge turneringer treffer dårligere kan dette være en annen forklaring på hvorfor treffsikkerheten er dårligere enn den var i GJP.

Foruten forskjellene i måten turneringene ble gjennomført på og muligheten for at spørsmålene i FFIs turnering kan være kvalitativt vanskeligere, er en annen forklaring på gapet i treffsikkerheten at det handler om selve deltagerne. Det er derfor variasjonene i treffsikkerheten til deltagernes ekspertise og individuelle egenskaper som vil undersøkes nærmere i de neste delkapitlene.

5.2 Ekspertise

Det andre forskningsspørsmålet i denne rapporten var: *Er eksperter bedre til å predikere forsvars- og sikkerhetspolitiske utviklinger enn andre?*

For å besvare dette spørsmålet sammenlignes treffsikkerheten til deltagerne i FFIs turnering, basert på vanlige mål på ekspertise. Ekspertise er definert som ferdigheter og kunnskaper som skiller en spesielt dyktig og sakkyndig person fra andre.[208]

Først sammenlignes treffsikkerheten til deltagere med forskjellige utdanningsnivåer, som ofte brukes som et mål på personers generelle ekspertise. Dess høyere utdanning, jo dyktigere anses gjerne en person for å være. Ett av funnene fra EPJ var imidlertid at utdanningsnivå ikke hadde noe å si for ekspertenes prediksjonsevne. Her vil funnene fra EPJ etterprøves basert på deltagere med en større variasjon i utdanningsnivåer enn det som tidligere har blitt undersøkt.

Personer med ekspertise innenfor et spesifikt fagområde omtales ofte som eksperter. En forventer gjerne at eksperter er bedre enn andre til å forutsi utviklinger innenfor sitt eget fagfelt. Det var derfor overraskende når EPJ fant at de mest selvsikre ekspertene var dårligere enn andre på sine egne ekspertiseområder. I GJP kom også en overraskende stor andel av superforecasterne fra disipliner som fysikk, biologi og programvareutvikling, ikke fra statsvitenskap. Dette delkapittelet vil derfor sammenligne treffsikkerheten til deltagere med ulik kompetanse innenfor forsvars- og sikkerhetspolitikk, som var det overordnede emnet til spørsmålene i FFIs turnering.

Foruten utdanning og relevant kompetanse kan det også tenkes at sektoren en person arbeider innenfor kan påvirke prediksjonsevnen gjennom ulik tilgang til informasjon. I EPJ ble det undersøkt om tilgang til gradert informasjon og hvorvidt ekspertene kom fra akademia hang sammen med treffsikkerheten, men heller ikke her ble det funnet noen sammenheng. Her etterprøves EPJs funn ved å sammenligne deltagere som i dag arbeider innenfor den norske forsvarssektoren, og som ofte har tilgang til gradert informasjon, med deltagere fra andre sektorer. Innad i forsvarssektoren sammenlignes også treffsikkerheten til forskere, offiserer og spesialister, som gjør det mulig å sammenligne «akademikere» med «praktikere» i det norske forsvarsmiljøet.

Til slutt undersøkes det hvordan eksperter som brukes i media treffer sammenlignet med andre fagfolk og alle andre deltagere. Bakgrunnen er at EPJs analogi om den pilkastende apen har blitt tolket som at «hvem som helst» kan slå eksperter, men de ble aldri sammenlignet med amatører. Et siste overraskende funn i EPJ var at de dårligste ekspertene var de mest siterte. Her undersøkes det derfor om de norske ekspertene også treffer dårligere jo oftere de har blitt brukt i media.

Til forskjell fra forrige delkapittel måles treffsikkerheten til deltagerne ved hjelp av *standardiserte* Brier-scores (se underkapittel 4.1.5). Det vil si at hver deltagers opprinnelige Brier-score er konvertert til en z-score på hvert spørsmål. Hensikten er å ta høyde for at deltagerne kan ha valgt spørsmål med ulike vanskelighetsgrader. Den standardiserte scoren måler dermed delta-

[208] Smeby, J.-C. (2021), 'ekspertise', *Store norske leksikon*.

gernes treffsikkerheten i forhold til hverandre. Siden lavere Brier-score betyr høyere treffsikkerhet, betyr en *negativ* standardisert Brier-score at deltageren var *bedre* enn snittet til alle andre, mens en *positiv* betyr at deltagerne var *dårligere*. Deltagernes treffsikkerhet er basert på snittet av alle standardiserte Brier-scorene de fikk på alle spørsmålene de svarte på. Deltagergruppenes treffsikkerhet er basert på snittet av alle individuelle scores. Det vil si at alle deltagere vektes likt innad i hver gruppe, uavhengig av hvor mange spørsmål de svarte på, så lenge de oppfylte minstekravet.

Som i delkapittel 5.1 benyttes t-tester for å måle om det er statistisk signifikante forskjeller mellom deltagernes snittscores.[209] Siden utdanningsnivå er den eneste ekspertisevariabelen i GJPs datasett, består de øvrige analysene kun av sammenligninger av deltagergrupper innad i FFIs turnering. Alle resultatene diskuteres opp mot funnene i EPJ, der de fleste sammenhengene som undersøkes her ble analysert. Deltagernes treffsikkerhet sammenlignes også med apens.[210]

De standardiserte Brier-scorene kan bare si noe forskjellene mellom deltagergruppene relativt sett, ikke noe om hvor godt de traff objektivt sett. En signifikant forskjell kan være liten og ubetydelig i praksis. Ved alle tilfeller der det er en signifikant forskjell mellom to deltagergruppers snittscores, oppgis det derfor hvor presist disse traff ved hjelp av de tre ulike måtene å måle prediksjoners treffsikkerhet på som ble diskutert i forrige kapittel:

1) *Brier-score*, som måler hvor godt deltagerne traff på en skala fra 0 til 2, der lavere betyr bedre treffsikkerhet. Brier-scoren reflekterer evnen til å oppgi høye sannsynligheter til riktige svar og lave til gale. I FFIs turnering er deltagernes Brier-score 0,52, mens apens er 0,51. Det betyr at deltagerne sliter med å slå tilfeldig gjetning, der sannsynlighetene fordeles helt likt på alle svaralternativer på alle spørsmål.

2) *Treffprosent*, som måler hvor gode deltagerne var til å peke på riktig utfall, uavhengig av sannsynlighetene de oppgav. I FFIs turnering er treffprosenten til deltagerne 51 %, mens den er 33 % ved tilfeldig gjetning. I motsetning til Brier-scoren sier ikke treffprosenten noe om evnen til å vurdere usikkerhetene rundt hendelsene som predikere, som kan være spesielt viktig i forsvars- og sikkerhetspolitisk sammenheng. Deltagere som er helt sikre og deltagere som er veldig usikre vil få den samme treffsikkerheten.

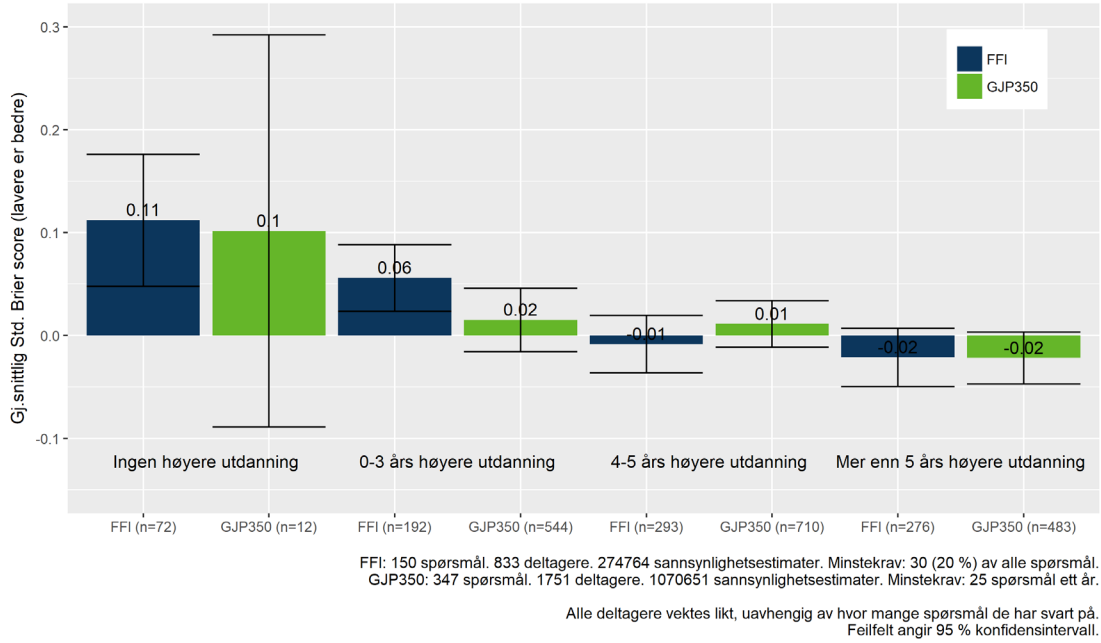
3) *Kalibrering*, som måler avstanden mellom hvor sikre deltagerne var og hvor godt de faktisk traff. Dette måles ved å sammenligne snittprediksjon på det deltagerne *trodde* var riktig svar (konfidens) med treffprosenten deres. En positiv verdi betyr at deltagerne trodde de var bedre enn de var (*over*konfidente), mens en negativ verdi betyr at de egentlig var bedre enn de trodde (*under*konfidente). I FFIs turnering er snittprediksjonen på antatt riktig svar 72 %, mens treffprosenten 51 %. Dette gir en overkonfidens på hele 21 %, som betyr at deltagerne har en tendens til å overpredikere hendelser.

[209] For histogrammer som viser fordelingen av de standardiserte Brier-scorene innenfor hver ekspertkategori, se kapittel 3 i Beadle (2021), 'Tilleggsdokumentasjon til foreløpige resultater fra FFIs prediksjonsturnering'.
[210] Apens standardiserte Brier-score er beregnet som om han var én av deltagerne i turneringen, men apen er ikke med i beregningen av deltagernes score og påvirker dermed ikke snittene deres.

5.2.1 Utdanningsnivå

I EPJ viste det seg at utdanningsnivået til ekspertene ikke hadde noe å si for treffsikkerheten. Det var imidlertid hele 96 % av de 284 ekspertene i EPJ som hadde utdanning på minst mastergradsnivå, slik at studien egentlig bare sammenlignet eksperter med master- eller doktorgrad. I GJP var ikke deltagerne like høyt utdannede, men 99 % hadde utdanning på minst bachelornivå, som var det eneste formelle kravet for å delta. I FFIs turnering hadde derimot 9 % ingen høyere utdanning, som gjør det mulig å sammenligne deltagere på tvers av alle utdanningsnivåer. Selv om GJP ikke har publisert noe om sammenhengene med deltagernes utdanningsnivå, er denne variabelen registrert i det offentlige datasettet. Det er derfor mulig å sammenligne treffsikkerheten til FFIs og GJPs deltagere på hvert utdanningsnivå. Siden EPJs datasett ikke er tilgjengelig er funnene som diskuteres her basert på resultatene fra studiens korrelasjonsanalyse.[211]

Figur 5.11 viser treffsikkerheten til deltagerne i FFIs og GJPs turneringer på hvert utdanningsnivå. Andelene deltagere per utdanningsnivå er relativt like i begge turneringer, men det var flere deltagere uten høyere utdanning i FFIs turnering. I tråd med funnene fra EPJ er det ingen signifikant forskjell mellom de aller mest utdannede deltagerne, hverken i FFIs eller GJPs turneringer. Når vi sammenligner alle utdanningsnivåene finner vi derimot en signifikant forskjell i FFIs turnering mellom dem med minst 4–5 års høyere utdanning og dem med mindre.[212]



Figur 5.11 Relativ treffsikkerhet, basert på utdanningsnivå.

I motsetning til EPJ viser altså resultatene fra FFIs turnering at det *er* en forskjell mellom treffsikkerheten til deltagere med ulike utdanningsnivåer, men at skillet går mellom dem som har

[211] Se tabell 3.1. i Tetlock (2005), *Expert Political Judgment*, s. 69.
[212] «Ingen høyere utdanning» vs. hhv. «4–5 års høyere utdanning»: $t(100) = 3.43$, $p < 0.001$, og «Over 5 års høyere utdanning»: $t(101) = 3.79$, $p < 0.001$. «0–3 års høyere utdanning» vs. hhv. «4–5 års høyere utdanning»: $t(425) = 2.96$, $p < 0.01$, og «Over enn 5 års høyere utdanning»: $t(421) = 3.54$, $p < 0.001$.

minst fire års høyere utdanning, som for mange vil tilsvare mastergradsnivå eller høyere, og dem som har opptil tre års høyere utdanning, som normalt tilsvarer bachelorgradsnivå eller lavere. I GJP er det ingen signifikante forskjeller, men det er så få deltagere uten høyere utdanning at det ikke er mulig å si noe sikkert om forskjellen mellom disse og resten. Snittet til GJPs deltagere med lavest utdanningsnivå er likevel betydelig dårligere enn resten, slik som i FFIs.

Hvor store er forskjellene i treffsikkerhet i praksis? Tabell 5.2 viser at de høyest utdannede i FFIs turnering treffer best, uansett treffsikkerhetsmål, og det største skillet går fremdeles mellom bachelor- og mastergradsnivå. De høyest utdannede har rundt 4 prosentpoeng lavere overkonfidens enn de lavest utdannede. Dette skyldes ikke at de høyest utdannede var mer forsiktige i sine prediksjoner, men at treffprosenten deres var bedre. Treffsikkerheten er likevel ikke imponerende. Mens deltagerne uten høyere utdanning har en treffprosent på 47 %, har de høyest utdannedes treffprosent bare 52 %. Dette er riktig nok bedre enn apens treffprosent (33 %), men Brier-scorene viser at alle deltagergruppene sliter med å slå apen (0,51). Det er altså ikke slik at deltagerne med høyest utdanning er så mye bedre enn tilfeldig gjetning; det er heller det at de med lavest som er enda dårligere. Faktisk er deltagerne på alle utdanningsnivåer langt dårligere kalibrerte enn ekspertene i EPJ, som hadde en overkonfidens på 12 %.[213]

Utdanningsnivå (antall deltagere)	**Brier-score**	**Treffprosent**	**Kalibrering**
Ingen høyere utdanning (72)	0,56	47,1 %	24,1 %
1–3 års høyere utdanning (192)	0,54	49,4 %	23,4 %
4–5 års høyere utdanning (293)	0,51	51,3 %	20,4 %
Over 5 års høyere utdanning (276)	0,50	52,0 %	19,8 %

Tabell 5.2 Objektiv treffsikkerhet, basert på utdanningsnivå i FFIs turnering.

Til sammenligning bekrefter tabell 5.3 hvordan deltagerne i GJP traff mye bedre enn FFIs, også i praksis. Treffprosenten ligger rundt 76 % på alle utdanningsnivåer og kalibreringen god, med en overkonfidens på bare 2 %. Alle Brier-scorene er også betydelig bedre enn apens (0,51).

Utdanningsnivå (antall deltagere)	**Brier-score**	**Treffprosent**	**Kalibrering**
Ingen høyere utdanning (12)	0,30	76,1 %	0,8 %
1–3 års høyere utdanning (544)	0,31	76,3 %	2,3 %
4–5 års høyere utdanning (710)	0,31	76,1 %	2,3 %
Over 5 års høyere utdanning (483)	0,30	76,9 %	1,9 %

Tabell 5.3 Objektiv treffsikkerhet, basert på utdanningsnivå i GJP350.

[213] Se Moore mfl. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', ss. 3561–3562. Overkonfidensen var 20 % blant eksperter som predikerte langsiktige spørsmål innenfor egne ekspertiseområder, men den var 3 % blant eksperter som predikerte kortsiktige spørsmål i domener utenfor egne ekspertiseområder.

5.2.2 Forsvars- og sikkerhetspolitisk kompetanse

En utbredt antagelse er at eksperter med kompetanse på et spesifikt tema har bedre forutsetninger enn andre for å vurdere fremtidig utvikling innenfor sine egne fagområder. En Russland-ekspert antas å være bedre til å forutsi hva som vil skje i Russland enn en Kina-ekspert. I EPJ viste det seg imidlertid at ekspertene som predikerte innenfor sine egne regionale ekspertiseområder slet med å slå dem som predikerte utenfor. De mest selvsikre ekspertene («pinnsvinene») var til og med dårligere enn de andre til å predikere spørsmål om sine egne regioner.

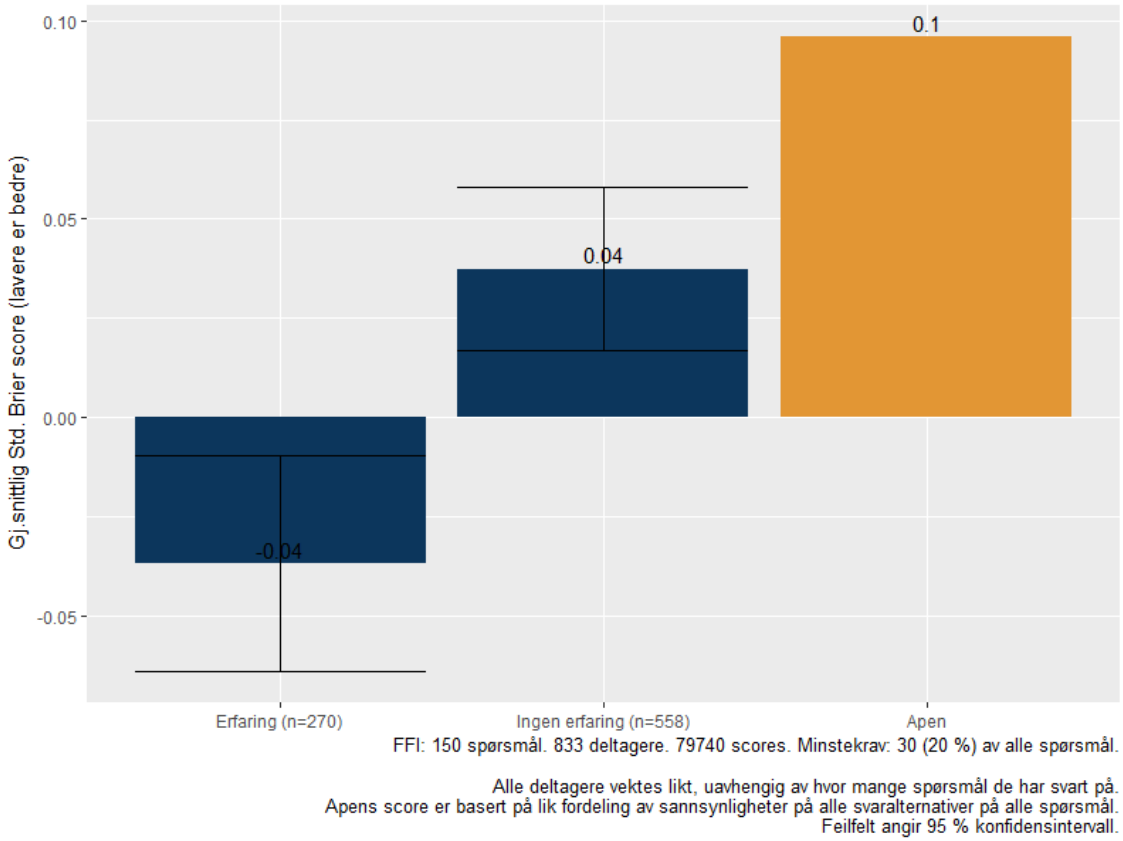
Én forklaring på dette er at politiske spørsmål, som dem ekspertene i EPJ ble bedt om å predikere, er preget av så mye usikkerhet at betydningen av ekspertise avtar. Områdeekspertene i EPJ kunne sikkert mye mer om regionene sine enn andre, men et høyere kunnskapsnivå ledet ikke til en bedre prediksjonsevne. Det kan tenkes at spørsmålene i FFIs turnering er beheftet med enda større usikkerhet, siden det vanligste temaet var krig og konflikt, som er sjeldne og spesielt uforutsigbare sosiale fenomener. Forrige delkapittel viste imidlertid at det er lite som tilsier at det er noen temaer som i seg selv er mye vanskeligere å predikere enn andre. Tvert imot treffer FFIs deltagere gjennomgående dårligere enn GJPs på akkurat de samme temaene.

Kompetanse består både av utdanning og erfaring. I de fleste stillingsutlysninger er det krav om relevant kompetanse, basert på en antagelse om at personer med utdanning eller erfaring på det aktuelle området har ferdigheter og kunnskaper som gjør dem bedre egnet enn personer uten. I denne rapporten defineres relevant kompetanse som utdanning i eller arbeidserfaring med forsvars- og sikkerhetspolitiske temaer. I FFIs turnering er det 270 (32 %) av 833 deltagere som arbeider eller har arbeidet med forsvars- og sikkerhetspolitiske spørsmål som en del av jobben sin. Det er altså en tredel av deltagerne som har profesjonell erfaring og relevant kompetanse på forsvar- og sikkerhetspolitikk. Nesten alle disse hadde også spisskompetanse på flere relevante temaer, som internasjonal politikk, økonomi, teknologi, NATO, Russland eller USA (se underkapittel 3.2.3). Det er disse deltagerne som i utgangspunktet utgjør ekspertgruppen som her analyseres.

De øvrige to tredeler av deltagerne i FFIs turnering har aldri arbeidet med forsvars- og sikkerhetspolitiske spørsmål som en del av jobben sin. Flere av dem er antageligvis fagfolk, men uten relevant kompetanse i denne sammenhengen.

For å undersøke betydningen av relevant kompetanse kan vi først sammenligne treffsikkerheten til deltagere med og uten noe forsvars- og sikkerhetspolitisk arbeidserfaring i det hele tatt. Figur 5.12 viser at begge disse to gruppene var i snitt bedre enn tilfeldig gjetning, men deltagerne med forsvars- og sikkerhetspolitisk arbeidserfaring var signifikant mer treffsikre enn dem uten.[214] I EPJ ble denne forskjellen aldri undersøkt, siden alle deltagerne i studien var politiske eksperter.

[214] «Erfaring» vs. «ingen erfaring»: $t(578) = -4.29$, $p < 0.0001$. «Erfaring» vs. apen: $t(269) = -9.64$, $p < 0.0001$. «Ingen erfaring» vs. apen: $t(557) = -5.57$, $p < 0.0001$.



Figur 5.12 Relativ treffsikkerhet, basert på relevant arbeidserfaring.

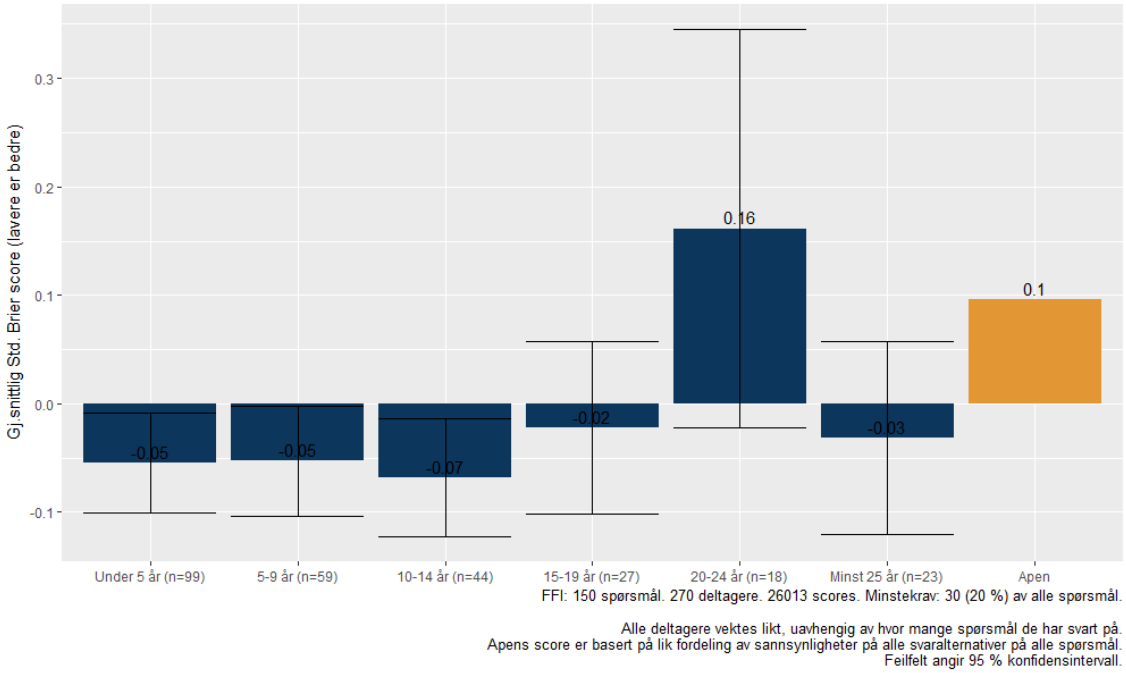
Samtidig viser tabell 5.4 at den gjennomsnittlige Brier-scoren til ekspertene i FFIs turnering bare er marginalt bedre enn snittet til alle deltagerne i turneringen (0,52) og omtrent akkurat like god som apen (0,51). Disse fagfolkene med relevant arbeidserfaring hadde også bare 2 prosentpoeng høyere treffprosent og tilsvarende lavere grad av overkonfidens enn deltagerne uten. Forskjellen mellom deltagere med og uten noe relevant kompetanse er derfor relativt liten i praksis, og enda mindre enn forskjellen mellom deltagerne med og uten minst 4 års høyere utdanning.

Relevant arbeidserfaring (antall deltagere)	**Brier-score**	**Treffprosent**	**Kalibrering**
Erfaring (270)	0,5	52,2 %	19,7 %
Ingen erfaring (558)	0,53	50,0 %	22,0 %

Tabell 5.4 Objektiv treffsikkerhet, basert på relevant arbeidserfaring.

Det kan likevel være store variasjoner i *hvor mye* arbeidserfaring fagfolk har. I oppnevnelser av ekspertutvalg foretrekkes ofte «senior» fagfolk med relativt lang arbeidsfaring på fagfeltet. I snitt hadde FFIs fagfolk ti års arbeidserfaring med forsvars- og sikkerhetspolitiske spørsmål, som er omtrent like lang relevant erfaring som ekspertene i EPJ sine tolv år.

Figur 5.13 viser treffsikkerheten til de forsvars- og sikkerhetspolitiske fagfolkene i FFIs turnering, basert på hvor mange års arbeidserfaring de selv oppgav at de hadde. I likhet med EPJ avdekkes det ingen sammenheng mellom mengden arbeidserfaring og eksperters treffsikkerhet.[215]



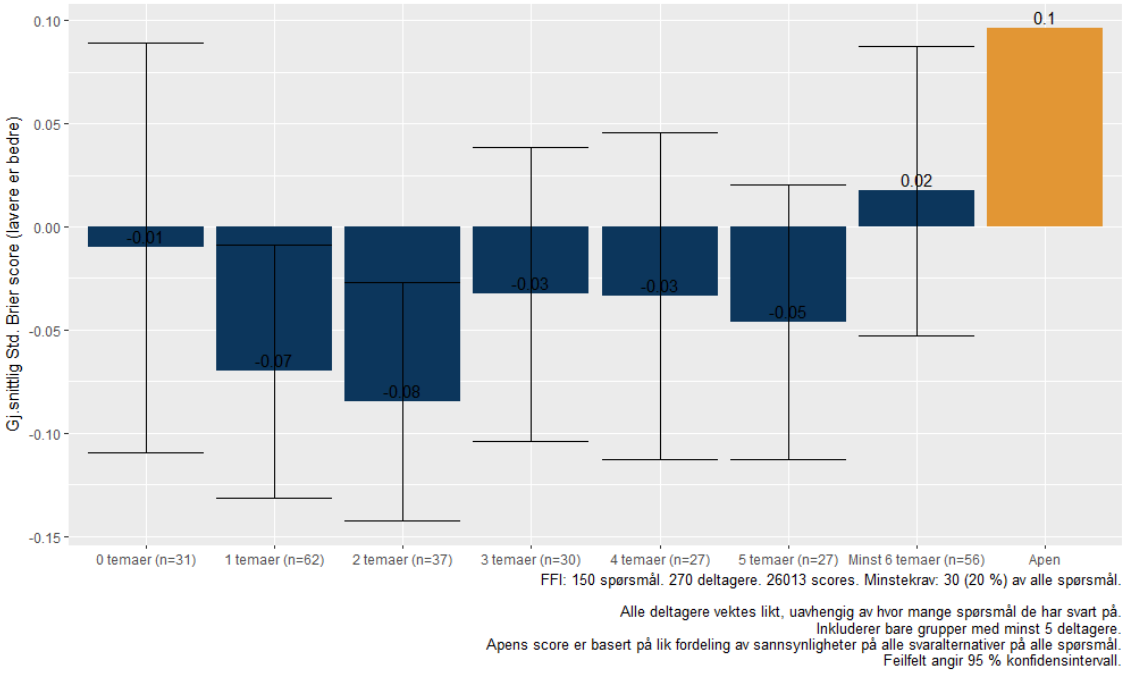
Figur 5.13 Relativ treffsikkerhet, basert antall års relevant arbeidserfaring.

Et annet vanlig mål på ekspertise er breddekompetanse. Kompetanse på flere forskjellige områder anses ofte som en fordel, fordi det gir en bedre helhetsforståelse enn ren spesialisering. Her er fagfolkenes breddekompetanse basert på hvor mange av spørsmålstemaene i FFIs turnering som deltagerne oppgav at de hadde kompetanse på. Hvis en bredere forståelse av forsvars- og sikkerhetspolitiske temaer henger sammen med høyere treffsikkerheten, burde fagfolk med kompetanse på relativt mange temaer treffe bedre enn deltagerne med kompetanse på relativt få.

Av alle 270 fagfolk i FFIs turnering oppgav 246 (91 %) at de hadde kompetanse på minst 1 av de 18 temaene som spørsmålene ble kategorisert innenfor. Figur 5.14 viser treffsikkerheten deres, basert på deltagernes totale antallet kompetanseområder. I snitt oppgav fagfolkene at de hadde kompetanse på 3,4 temaer hver. Figuren viser ingen sammenheng mellom kompetanse på

[215] Det er i utgangspunktet signifikante forskjeller på 0.05-nivå mellom snittscorene til deltagerne med 20–24 års erfaring og flere av de andre erfaringsnivåene. Med unntak av de to første gruppene er imidlertid utvalgene små og fordelingene av scores usymmetriske, som gjør både t-testen og Wilcoxon-testen lite egnet. Større utvalg og mer symmetrisk fordeling oppnås ved å samle alle deltagerne med minst 10 års erfaring. Etter dette grepet avdekkes det ingen signifikante forskjeller mellom treffsikkerheten til deltagerne med under 5 års erfaring, 5–9 år eller minst 10 års erfaring.

flere temaer og bedre treffsikkerhet. De små forskjellene som vises er ikke statistisk signifikante.[216] Det er heller ingen signifikant korrelasjon mellom ekspertenes individuelle snittscores og antall kompetansetemaer.[217]



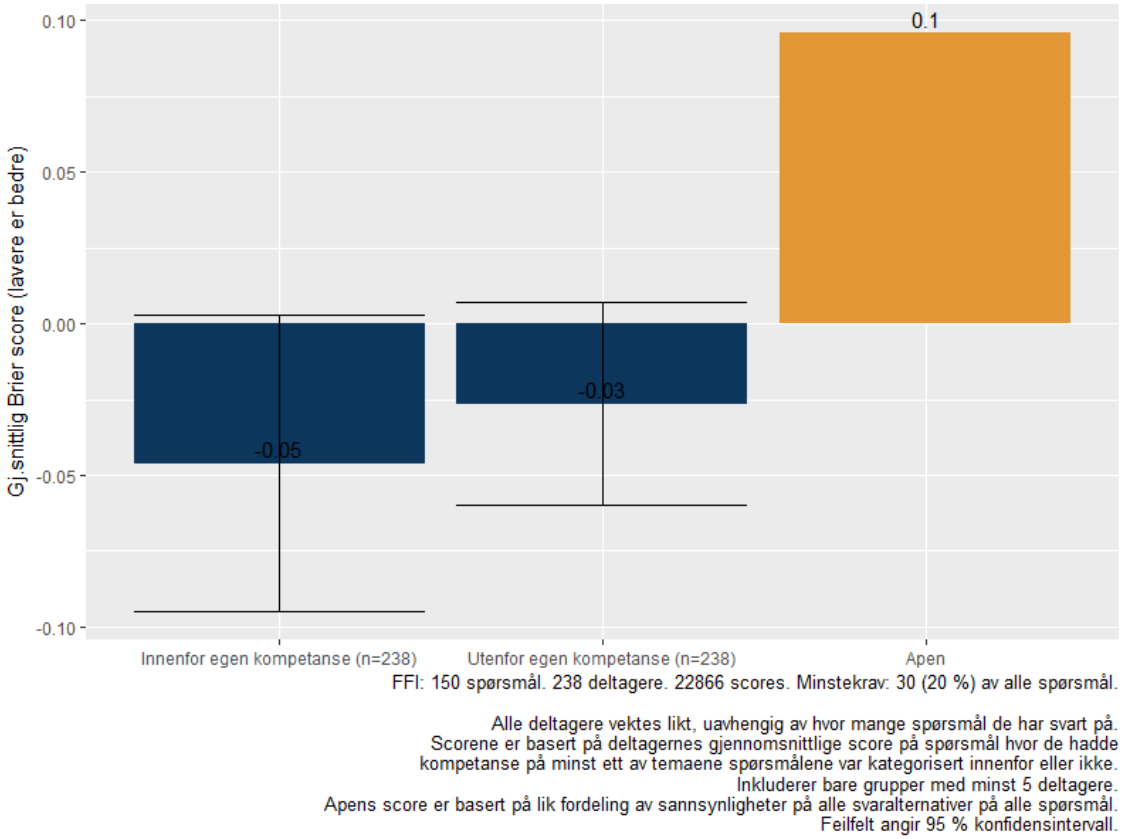
Figur 5.14 Relativt treffsikkerhet, basert antall kompetansetemaer totalt.

Siden fagfolkene i FFIs turnering oppgav at de hadde kompetanse på rundt tre temaer hver, betyr dette samtidig at de fikk spørsmål om mange andre temaer enn det de selv var eksperter på. Fagfolk forventes først og fremst å treffe bedre på spørsmål de selv har spesialisert seg på. Disse omtales ofte som *Subject Matter Experts* (SME) innenfor sine områder. For å undersøke betydningen av relevant spisskompetanse kan vi sammenligne treffsikkerheten til de samme fagfolkene på spørsmål innenfor og utenfor deres egne ekspertisetemaer.

Figur 5.15 viser fagfolkenes treffsikkerhet på spørsmål der de hadde kompetanse på minst ett av de samme temaene som spørsmålene omhandlet, sammenlignet med treffsikkerheten deres på spørsmål der de ikke hadde kompetanse på noen relevante temaer. Her inkluderes bare deltagere som predikerte minst én gang både innenfor og utenfor egne kompetanseområder. Det er imid-

[216] Selv om det er signifikante forskjeller på 0.05-nivå mellom snittscorene til deltagerne med kompetanse på 1 eller 2 temaer og deltagerne med minst 6 på enten t- eller Wilcoxon-testene, er det relativt få deltagere i hver gruppe og scorene deres er ikke symmetrisk fordelte. Ved å samle deltagerne i tre omtrent like store grupper med kompetanse på hhv. 0–1 temaer (93 deltagere), 2–4 temaer (94 deltagere) og minst 5 temaer (83 deltagere) blir utvalgene større og fordelingene av scores mer symmetriske. Da er det ikke lenger statistisk signifikante forskjeller mellom snittscorene.
[217] Siden antall kompetanseområder ikke er symmetrisk fordelt er korrelasjonene målt ved både Pearsons r og Spearmans r_s. Pearson: $r = 0.05$, $t(268) = 0.89$, p = 0.35. Spearman: $r_s = 0.11$, $p = 0.09$.

lertid ingen signifikant forskjell mellom ekspertenes treffsikkerhet på temaer de hadde spisskompetanse på og ikke.[218] Også dette funnet er i tråd med EPJ, der ekspertene som predikerte innenfor sine egne regioner traff omtrent like godt som ekspertene som predikerte utenfor.

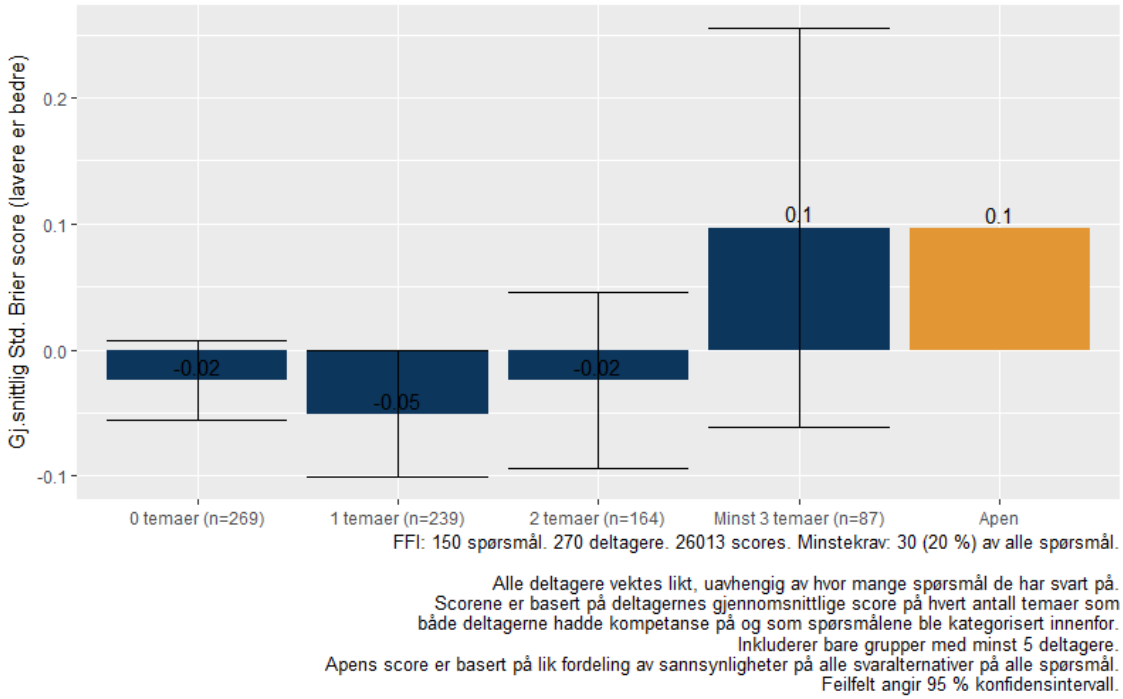


Figur 5.15 Relativ treffsikkerhet, basert spørsmål innenfor og utenfor fagfolkenes egne kompetanseområder.

Til slutt kan det tenkes at det kreves både bred og spesialisert kompetanse på spørsmålenes temaer for at ekspertise skal henge sammen med treffsikkerheten. Det er naturlig å anta at en fagperson med kompetanse på NATO, USA og økonomi er bedre til å predikere spørsmål om alle disse tre temaene enn en annen med kompetanse på bare økonomi. For å undersøke dette kan vi måle forskjellene i treffsikkerhet ut fra antallet felles temaer som fagfolkene hadde kompetanse på og som spørsmålene også er kategorisert innenfor. I snitt hadde fagfolkene ingen kompetanse på noen relevante temaer på 61 % av spørsmålene, mens de hadde kompetanse på ett relevant tema på 33 % av spørsmålene, kompetanse på to relevante temaer på 13 % av dem og kompetanse på minst tre relevante temaer på bare 5 %.

[218] «Innenfor egen kompetanse» vs. «utenfor egen kompetanse»: $t(237) = –0.85$, $p = 0.40$. Siden snittene som sammenlignes er basert på de samme deltagerne er det her benyttet en paret t-test, basert på de 238 deltagerne som svarte på spørsmål både innenfor og utenfor sine egne kompetanseområder. Totalt var det 269 deltagere som predikerte spørsmål innenfor egen kompetanse og 239 deltagere utenfor.

Figur 5.16 viser fagfolkenes treffsikkerhet basert på antall temaer deltagernes kompetanse og spørsmålene hadde til felles. Heller ikke her er det noen signifikante forskjeller mellom deltagernes snittscores.[219] Det spilte altså ingen noen rolle om deltagerne hadde kompetanse på ingen, noen få eller mange av temaene som spørsmålene omhandlet.



Figur 5.16 Relativ treffsikkerhet, basert antall spørsmålsrelevante temaer fagfolkene hadde kompetanse på.

5.2.3 Ansatt i forsvarssektoren

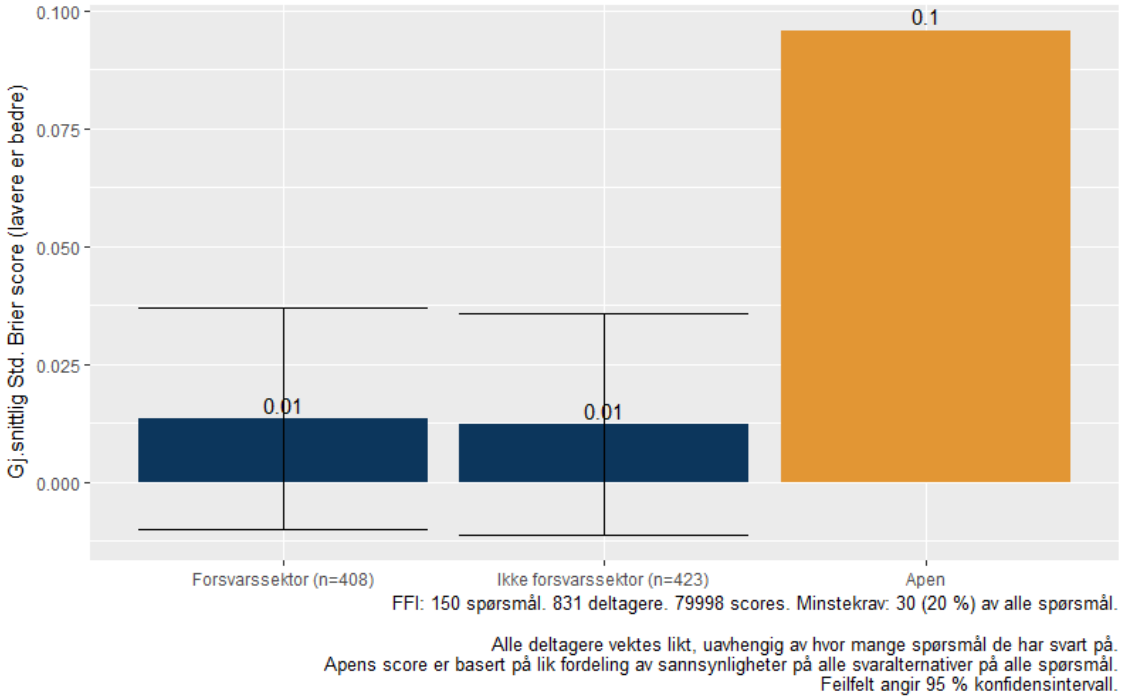
Foruten utdanningsnivå og kompetanse kan det også tenkes at prediksjonsevnen kan henge sammen med hvilken sektor en arbeider innenfor. To andre variabler som ble målt i EPJ var nemlig hvorvidt ekspertene hadde tilgang til gradert informasjon og om de kom fra akademia eller ikke, men heller ikke her ble det funnet noen sammenhenger med treffsikkerhet.[220] I tråd med dette funnet traff de beste deltagerne i GJP, som ikke hadde tilgang til gradert informasjon, 30 % bedre enn et prediksjonsmarked bestående av amerikanske etterretningsanalytikere.[221]

I FFIs turnering var 408 (49 %) av alle 833 deltagere ansatt i forsvarssektoren på tidspunktet de registrerte seg. Av disse 408 deltagerne kom 132 (32 %) fra FFI, hvorav de alle fleste var fors-

[219] Mange deltagere er inkludert i flere, men ikke i alle snittene som sammenlignes her. Her er det derfor gjennomført parede t-tester av snittene til hver kombinasjon av antall temaer, basert på deltagere med scores på begge. Det vil si 238 deltagere med snittscores på 0 og 1 temaer, 163 deltagere på 0 og 2 temaer, 86 deltagere på 0 og minst 3 temaer, 164 deltagere på 1 og 2 temaer, 87 deltagere på 1 og minst 3 temaer og 87 deltagere på 2 og minst 3 temaer.
[220] Tabell 3.1 i Tetlock (2005), *Expert Political Judgment*, s. 69.
[221] Tetlock mfl. (2017), 'Bringing probability judgments into policy debates via forecasting tournaments'.

kere, 51 (13 %) fra Hæren, hvorav mange tilhørte Etterretningsbataljonen, og 41 (10 %) fra Forsvarets høgskole, der de fleste arbeidet på Institutt for forsvarsstudier, Stabsskolen eller krigsskolene. Det ble ikke kartlagt hvorvidt hver deltager hadde tilgang til gradert informasjon, men det er grunn til å anta at mange av dem som arbeidet innenfor forsvarssektoren hadde dette i motsetning til deltagerne som arbeidet i andre sektorer. Av de 423 (51 %) deltagerne som *ikke* var ansatt i forsvarssektoren kom 100 (24 %) fra faglig, vitenskapelig og teknisk tjenesteyting, 63 (15 %) fra offentlig administrasjon, 51 (12 %) fra informasjon og kommunikasjon og 44 (10 %) var ikke yrkesaktive/pensjonerte.[222]

Figur 5.17 viser imidlertid at det ikke er noen signifikant forskjell mellom treffsikkerheten til ansatte i forsvarssektoren og ikke.[223] Gitt at sektortilhørighet er en gyldig proxy for informasjonstilgang, støtter dette EPJs funn om at det ikke er noen sammenheng mellom treffsikkerhet og tilgang til gradert informasjon – heller ikke på forsvars- og sikkerhetspolitiske spørsmål.



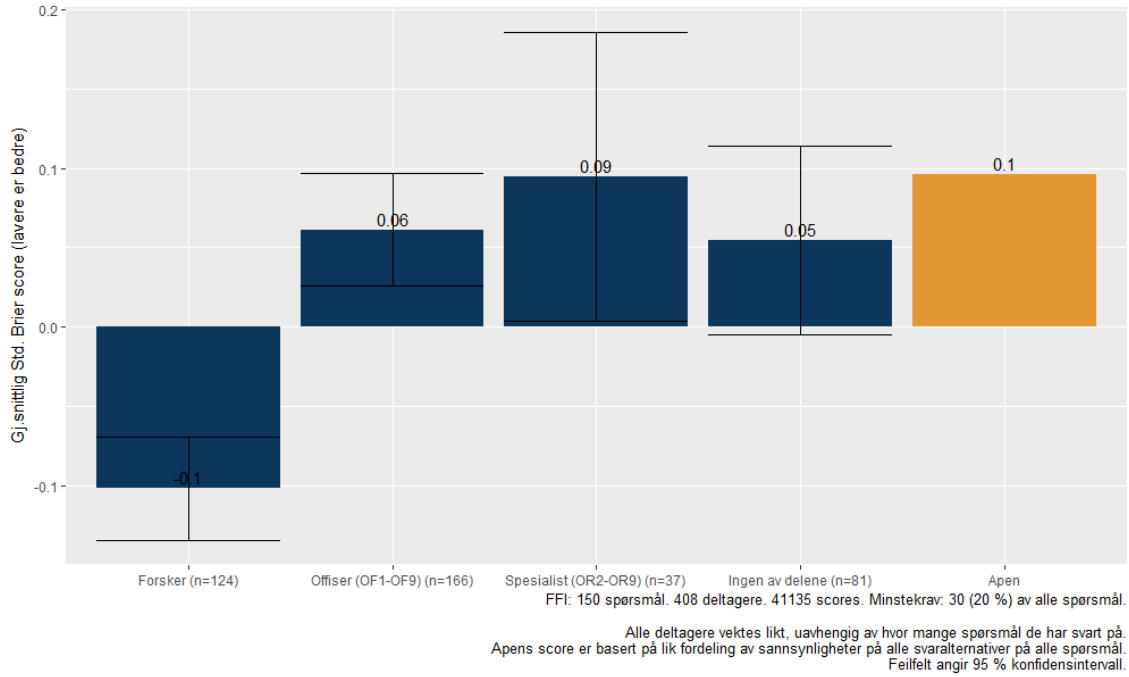
Figur 5.17 Relativ treffsikkerhet, basert på hvorvidt deltagerne er ansatt innenfor eller utenfor forsvarssektoren i dag.

Deltagerne i FFIs turnering ble heller ikke spurt spesifikt om de kom fra akademia eller ikke, men alle ansatte i forsvarssektoren ble bedt om å oppgi hva slags stilling de hadde. Av alle 408

[222] Av disse 423 deltagerne var det 105 (25 %) som tidligere hadde vært ansatt i forsvarssektoren, men siden de ikke var det nå antas det at de ikke hadde tilgang til gradert informasjon når de deltok i turneringen.
[223] «Forsvarssektor» vs. «ikke forsvarssektor»: $t(829) = 0.07$, $p = 0.94$.

forsvarssektoransatte var 124 (30 %) forskere, 166 (41 %) offiserer (OF1–9), 37 (9 %) spesialister (OR2–9) og 81 (20 %) ingen av delene. Her brukes skillet mellom forskere og øvrige stillingstyper som en proxy for å måle treffsikkerheten til deltagere fra akademia og ikke.[224]

Figur 5.18 viser at forskerne treffer signifikant bedre enn alle de andre stillingskategoriene.[225] Det er derimot ingen signifikante forskjeller mellom offiserer og spesialister eller mellom de militære og resten. I utgangspunktet kan dette utfordre resultatene fra EPJ, der det ikke ble funnet noen forskjell mellom eksperter som kom fra akademia og ikke. En mulig forklaring er imidlertid at EPJs definisjon av akademia ser ut til å være avgrenset til forsknings- og universitetsmiljøer, siden ekspertene som *ikke* kom fra akademia er kategorisert som at de kom fra staten, tenketanker og stiftelser, internasjonale organisasjoner eller privat sektor, som også kan inkludere akademikere. Funnene fra EPJ og FFIs turnering er derfor ikke direkte sammenlignbare.



Figur 5.18 Relativ treffsikkerhet, basert på stillingskategori i forsvarssektoren i dag.

Gapet mellom forskerne og de militære er den største relative forskjellen som er funnet så langt. Tabell 5.5 viser at forskernes Brier-score er bedre enn apens (0,51), mens alle de andre stillingskategorienes scores er dårligere. I praksis tilsvarer forskernes score en prediksjon på 52 % for det riktige svaret på et spørsmål med to mulige utfall, mens offiserenes en prediksjon på 48 %. Treffprosenten til forskerne er også 4–5 prosentpoeng høyere enn offiserenes og spesialistenes. Den største forskjellen er at forskerne har 7–8 prosentpoeng lavere overkonfidens enn begge

[224] Det ble ikke undersøkt hva disse deltagerne fra forsvarssektoren arbeidet med, men av alle 408 ansatte i forsvarssektoren var det 183 (45 %) som arbeidet med forsvars- og sikkerhetspolitikk. Siden det finnes offiserer som driver med forskning, f.eks. på Forsvarets høgskole, er alle offiserer som også krysset av for at de var forskere regnet som forskere i denne sammenhengen. Dette gjaldt imidlertid bare tre deltagere.

[225] «Forsker» vs. «offiser»: $t(287) = -6.7$, $p < 0.0001$; «spesialist»: $t(46) = -4.1$, $p < 0.001$; «ingen av delene»: $t(128) = -4.6$, $p < 0.0001$.

militære stillingstyper. Forskerne er altså betydelig bedre kalibrerte, men likevel dårligere enn både ekspertene i EPJ og deltagerne i GJP, med overkonfidens på hhv. 12 % og 3 %.

Stilling i forsvarssektoren (antall deltagere)	**Brier-score**	**Treffprosent**	**Kalibrering**
Forsker (124)	0,46	53,7 %	17,5 %
Offiser (OF1–9) (166)	0,54	49,8 %	24,9 %
Spesialist (OR2–9) (37)	0,55	48,9 %	25,7 %
Ingen av delene (81)	0,53	49,2 %	21,7 %

Tabell 5.5 Objektiv treffsikkerhet, basert på stillingskategori i forsvarssektoren i dag.

5.2.4 Bruk i media

Den eneste bakgrunnsvariabelen som i EPJ hang sammen med treffsikkerheten til ekspertene var hvor ofte de ble brukt i media: Dess *mer kjent* ekspertene var, jo *dårligere* var treffsikkerheten deres. Dette funnet var basert på en omvendt korrelasjon mellom treffsikkerheten og ekspertenes selvrapporterte kontakt med media og antallet Google-treff på navnene deres.[226]

I EPJ var definisjonen av en ekspert at personen arbeidet med trender av betydning for stater, regioner eller verden generelt.[227] Det var ingen krav om å ha vært i media, men 173 (61 %) av 284 eksperter hadde blitt intervjuet av minst ett stort medium og 60 (21 %) intervjuet minst 10 ganger. Det betyr at hele 233 (82 %) av ekspertene i EPJ hadde blitt brukt i media.

Alle de 270 deltagerne i FFIs turnering som arbeidet med forsvars- og sikkerhetspolitikk oppfyller EPJs ekspertdefinisjon, siden de også kan sies å arbeide med trender av betydning for stater, regioner eller verden generelt. Samtidig er FFIs definisjon snevrere enn EPJs, fordi den bare inkluderer eksperter på forsvars- og sikkerhetspolitikk, ikke politikk generelt.

I FFIs turnering var det 68 (25 %) av de 270 ekspertene som oppgav at de hadde blitt intervjuet i media om forsvars- og sikkerhetspolitiske spørsmål, hvorav 49 av dem hadde blitt sitert eller omtalt minst 10 ganger. De øvrige 202 ekspertene hadde derimot ikke blitt intervjuet. FFIs eksperter inkluderer altså i utgangspunktet omtrent like mange eksperter som i EPJ totalt, men det var langt færre av dem som hadde blitt brukt i media.[228]

[226] Tetlock (2005), *Expert Political Judgment*, ss. 62–63. Det er den selvrapporterte kontakten med media (på en skala fra 1 til 7) som er rapportert i tabell 3.1 (s. 69), mens berømthet («*fame*») var basert på antall Google-treff.
[227] For definisjonen av hvem som kvalifiserte som «ekspert», se Tetlock (2005), *Expert Political Judgment*, s. 239ff.
[228] Forskjellen i andelene eksperter som hadde blitt brukt i media i EPJ og FFIs turnering kan tyde på seleksjonsbias. I en studie av norske forskeres deltagelse i media svarte 66 % at de hadde blitt intervjuet av journalist. Carlsen, B., Müftüoglu, I. B. og Riese, H. (2014), 'Forskning i media: Forskere om motivasjon og erfaringer fra medieintervjuer', *Norsk medietidsskrift*, 21:3, ss. 188–208. Andelen var høyere blant forskere fra samfunnsvitenskapelige fag (78 %) enn fra naturvitenskapelige og tekniske fag (55 %). Disse andelene ligger mye nærmere andelen i EPJ (82 %) enn i FFIs turnering (25 %). To mulige forklaringer kan være at en betydelig andel av forskerne i FFIs turnering kommer fra naturvitenskapelige fag, selv om de jobber med forsvars- og sikkerhetspolitikk, og mange av dem jobber gradert, som gjør forskerne mindre synlige og redusere sannsynligheten for at de ville latt seg intervjue. Det kan derfor tenkes

De resterende 563 deltagerne i FFIs turnering hadde ingen forsvars- og sikkerhetspolitisk arbeidserfaring og regnes ikke som eksperter i denne sammenhengen. En femtedel av disse kom fra faglig, vitenskapelig og teknisk tjenesteyting og inkluderer derfor flere fagfolk som antageligvis også har blitt brukt som eksperter i media, men på andre fagområder enn forsvars- og sikkerhetspolitikk. Disse kategoriseres her sammen med alle andre deltagere som «amatører», fordi de er personer som gjennom turneringsdeltagelsen har vurdert forsvars- og sikkerhetspolitiske spørsmål uten å ha eller å ha hatt dette som levebrød.[229] Mange av deltagerne i denne gruppen kom fra offentlig administrasjon, informasjon og kommunikasjon eller var pensjonister.

Amatørene er her inkludert som en kontrollgruppe, fordi analogien fra EPJ – om at eksperter treffer omtrent like godt som en pilkastende ape – kan gi et inntrykk av at «hvem som helst» kan slå profesjonelle eksperter. Dette ble imidlertid aldri undersøkt i EPJ, siden det bare var eksperter som ble sammenlignet med hverandre. Amatørene i FFIs turnering er likevel ikke hvem som helst, men personer med generelt stor interesse for forsvars- og sikkerhetspolitiske spørsmål.[230]

Figur 5.19 viser treffsikkerheten til fagfolkene som har blitt brukt som eksperter i media, fagfolkene som ikke har det og amatørene uten noe arbeidserfaring fra forsvars- og sikkerhetspolitikk. Resultatene viser at ekspertene brukt i media treffer bedre enn andre eksperter, men forskjellen er ikke signifikante. Det er derimot en signifikant forskjell både mellom ekspertene i media og amatørene og mellom ekspertene utenfor media og amatørene.[231]

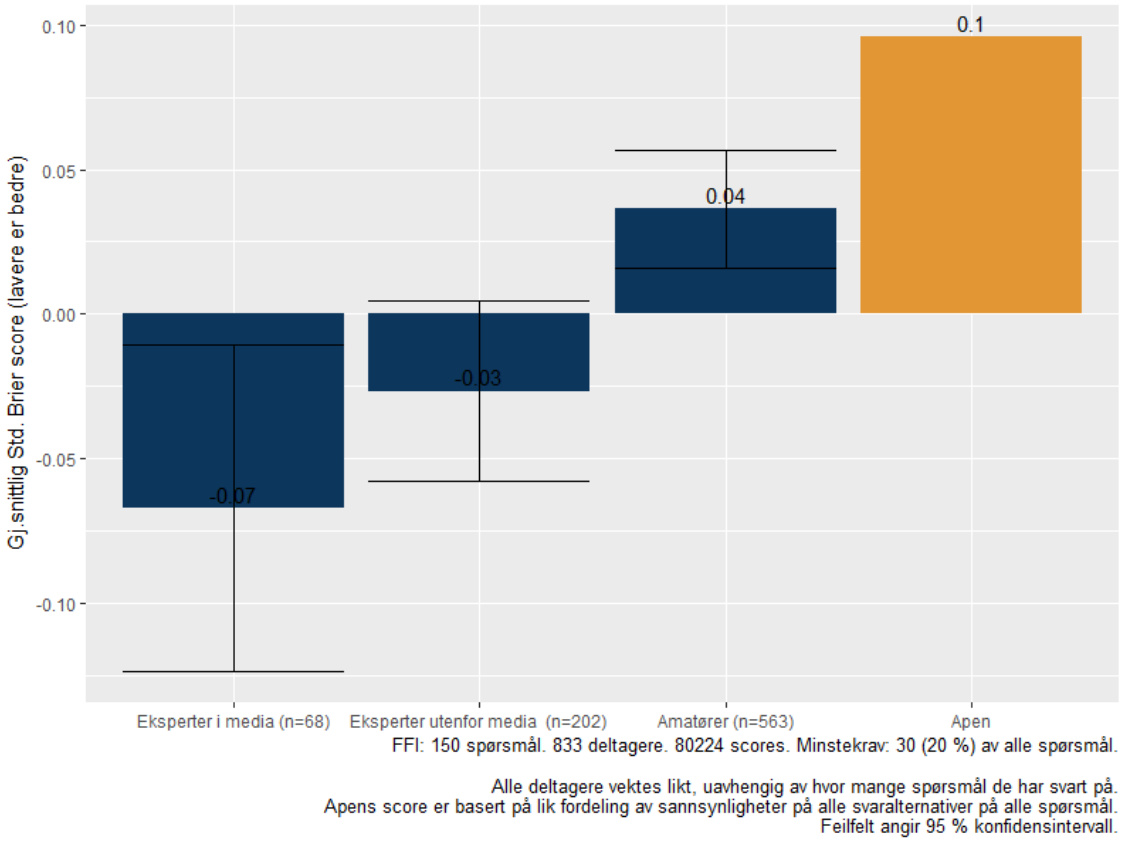
Resultatene viser at det ikke er grunnlag for å hevde at «hvem som helst» kan slå forsvars- og sikkerhetspolitiske eksperter, hverken dem som brukes i media eller ikke. I motsetning til funnene fra EPJ tyder resultatene heller ikke på at eksperter i media er dårligere enn dem utenfor.

at den lavere andelen i FFIs turnering er mer representativ for hvor mye eksperter innenfor forsvars- og sikkerhetspolitikk lar seg intervjue i media sammenlignet med eksperter som arbeider med politiske spørsmål generelt.

[229] Nilstun, C. (2021), 'amatør', *Store norske leksikon*.

[230] Under registreringen ble deltagerne bedt om å vurdere sin egen interesse for forsvars- og sikkerhetspolitiske spørsmål. På en skala fra 1 til 7, der 1 var «svært liten» og 7 var «svært stor», var medianscoren 6 («ganske stor») blant 833 deltagere som har svart på minst 20 % av de avgjorte spørsmålene.

[231] Eksperter i media vs. amatører: $t(87) = -3.44$, $p < 0.0001$. Eksperter utenfor media vs. amatører: $t(390) = -3.33$, $p < 0.0001$.



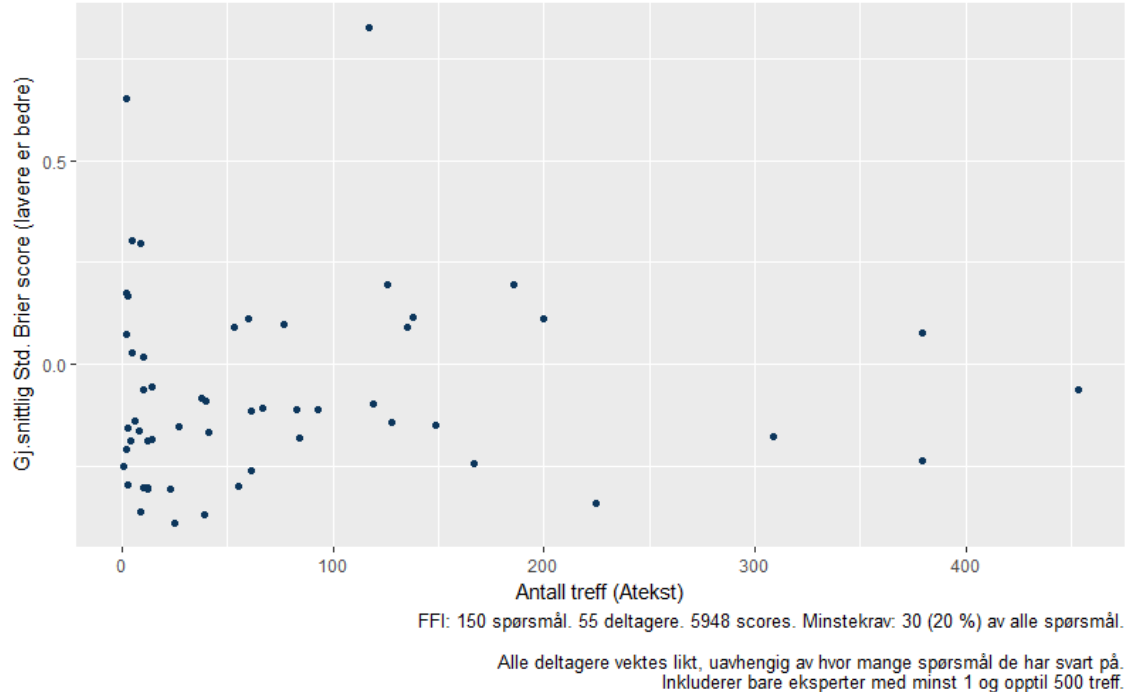
Figur 5.19 Relativ treffsikkerhet, basert på eksperters bruk i media.

For å etterprøve den omvendte korrelasjonen mellom berømmelse og treffsikkerhet i EPJ har antallet treff på de 68 ekspertene i FFIs turnering som har blitt intervjuet i media blitt kartlagt, basert på Google og den digitale arkivtjenesten Atekst, som bare inneholder internettkilder og artikler fra de største norske papiravisene, fagbladene og magasinene.[232]

Figur 5.20 viser spredningen av snittscorene til 55 eksperter med minst 1 relevant treff i Atekst i løpet av de 10 siste årene frem til turneringen ble avsluttet (2010–2020). 4 eksperter gav ingen treff, mens 9 eksperter med over 500 treff er ikke med i figuren, fordi disse var registrert med veldig mange flere treff enn resten (1000–4000), men de er med i analysen uten at det påvirker resultatene. Plottet viser ingen sammenheng mellom treffsikkerhet og antall treff. Siden dette

[232] To forskjellige kilder ble brukt for å redusere sjansen for at resultatene blir påvirket av hvordan bruken av ekspertene i media er målt. Søkene er gjort ved å telle treff på ekspertenes eksakte navn. Hvis ekspertene er siterte med forskjellige versjoner av navnet, f.eks. med og uten mellomnavn eller initialer, er antallet treff basert på summen av treff på kombinasjonene. En mulig feilkilde er at slike søk også inkluderer treff på andre personer med identiske navn, men det samme ville ha vært tilfellet i EPJ. Dette ble tatt høyde for ved å ekskludere treff som ikke holdt søkeord som var spesielt knyttet til ekspertene våre, f.eks. «Forsvaret» hvis vedkommende hadde uttalt seg som offiser. En annen feilkilde er at personene kan være kjent for andre ting enn det de er eksperter på. Det er ikke gjort en gjennomgang av innholdet i hvert treff, men siden analysen er basert på Atekst er det bare treff i media som er med.

heller ikke er tilfellet om analysen baseres på antall generelle Google-treff eller på bare nyhetssaker på Google, gjengis ikke disse figurene her. Det er heller ingen signifikant korrelasjon mellom treffsikkerheten og antall treff basert på noen av kildene.[233]



Figur 5.20 Ekspertenes relative treffsikkerhet, basert på antall treff i media.

Nok en gang utgjør ikke de signifikante forskjellene i relativ treffsikkerhet mye i praksis. Som vist i tabell 5.6 treffer ekspertene i media bedre enn andre fagfolk og amatører, uansett treffsikkerhetsmål, men det er snakk om 3–4 prosentpoeng høyere treffprosent og et par prosentpoeng lavere overkonfidens. Treffprosenten til ekspertene i media er litt høyere enn forsvarsforskernes, men Brier-scoren er omtrent identisk og overkonfidensen litt høyere enn deres (17,5 %).

Bruk i media (antall deltagere)	**Brier-score**	**Treffprosent**	**Kalibrering**
Eksperter i media (68)	0,47	54,5 %	18,8 %
Eksperter utenfor media (202)	0,50	51,5 %	20,0 %
Amatører (563)	0,53	50,1 %	21,9 %

Tabell 5.6 Objektiv treffsikkerhet, basert på eksperters bruk i media.

I motsetning til EPJ er det altså ikke funnet noen sammenheng mellom eksperters tilstedeværelse i media og treffsikkerheten deres i FFIs turnering. Tvert imot er ekspertene som har blitt brukt i media *bedre* enn dem som ikke har det, selv om det er for stor usikkerhet til at det kan

[233] Siden antall siteringer ikke er symmetrisk fordelt er korrelasjonene målt ved både Pearsons r og Spearmans r_s. Atekst: $r = 0.04$, $t(53) = 0.31$, p = 0.76; $r_s = 0.07$, $p = 0.61$. Google (alle): $r = –0.01$, $t(53) = –0.07$, p = 0.95; $r_s = 0.13$, $p = 0.36$. Google (nyheter): $r = –0.02$, $t(53) = –0.12$, p = 0.91; $r_s = 0.14$, $p = 0.33$.

påvises en signifikant forskjell og at forskjellene er beskjedne i praksis. Det viktigste er at hvor ofte ekspertene har blitt brukt i media har ingen sammenheng med treffsikkerheten deres.

5.2.5 Diskusjon

De foreløpige resultatene fra FFIs turnering både underbygger og utfordrer tidligere funn om eksperters treffsikkerhet.

For det første sliter ekspertene i FFIs turnering med å slå tilfeldig gjetning, slik som i EPJ. Eksperter var her definert som deltagere med profesjonell erfaring med forsvars- og sikkerhetspolitiske spørsmål og kompetanse på relevante temaer. De foreløpige resultatene synes således å styrke hypotesen om at betydningen av ekspertise er begrenset i forbindelse med prediksjon av internasjonal politikk, også mer spesifikt innenfor forsvars- og sikkerhetspolitikk. Faktisk er alle ekspertgruppene i FFIs turnering betydelig mer overkonfidente enn ekspertene i EPJ.

For det andre er det ikke slik at «hvem som helst» er bedre enn ekspertene, slik EPJ kunne gi inntrykk av. Selv om ekspertene sliter med å slå tilfeldig gjetning, er amatørene enda dårligere. Hva gjelder de delene av etterretningsanalyse og langtidsplanlegging som er avhengig av subjektive vurderinger av fremtidig utvikling, er det derfor mer hensiktsmessig å basere seg på prediksjonene til eksperter enn amatører. Samtidig er eksperters evne til å oppgi høye sannsynligheter til riktige svar og lave sannsynligheter til gale ikke bedre enn apens og forskjellene mellom eksperter og amatører er svært små i praksis. Riktig nok har eksperter, i likhet med deltagerne flest, en mye bedre treffprosent enn apen (33 %), men det er likevel bare snakk om en treffprosent på 55 % på spørsmål med 4–5 svaralternativer. Forskjellene mellom eksperter og amatører er fremdeles små, også ved dette treffsikkerhetsmålet.

For det tredje synes kriteriene som normalt brukes til å skille eksperter – utdanningsnivå, erfaring og kompetanse – lite relevante i prediksjonssammenheng. I motsetning til EPJ er det i FFIs turnering funnet forskjeller i treffsikkerheten til deltagere med ulikt utdanningsnivå. Deltagere med minst 4–5 års høyere utdanning er signifikant mer treffsikre enn dem med mindre utdanning. I tråd med EPJ har det imidlertid ingenting å si om deltagerne har *mer* enn 4-5 års utdanning eller ikke. Igjen er ikke forskjellene mellom deltagere med ulik utdanning store i praksis.

Det mest overraskende funnet er at spisskompetanse *ikke* henger sammen med bedre treffsikkerhet. Det er ingenting som tyder på at en områdeekspert er bedre til å forutsi spørsmål om det samme området enn fagfolk med kompetanse på andre temaer eller regioner, hverken i FFIs turnering eller i EPJ. Det eneste kompetansekriteriet hvor det i FFIs turnering er signifikante forskjeller i treffsikkerheten handler om hvorvidt deltagerne har arbeidserfaring med forsvars- og sikkerhetspolitiske spørsmål eller ikke. Hvor lang erfaring de har eller hvor mange spørsmålsrelevante temaer de har kompetanse på innenfor spiller ingen rolle. Det er også lite som tyder på at tilgang til gradert informasjon øker treffsikkerheten, hverken i EPJ eller i FFIs turnering. Den største relative forskjellen i treffsikkerhet i FFIs turnering er mellom forskere og militært ansatte i den norske forsvarssektoren. Forskerne har den beste Brier-scoren av alle ekspertgruppene analysert i dette delkapittelet. De er bedre enn tilfeldig gjetning, men ikke med stor margin.

Det viktigste funnet er kanskje fraværet av en sammenheng mellom bruken av eksperter i media og treffsikkerheten deres, som var det mest urovekkende resultatet i EPJ. Det er heller ingen korrelasjon mellom hvor mye de er brukt og hvor godt de treffer. Tvert imot treffer ekspertene som har blitt brukt i media bedre enn ekspertene som ikke har det, men denne forskjellen er ikke signifikant. Basert på resultatene til de norske ekspertene i FFIs turnering er det derfor ikke grunnlag for å ha mindre tillit til prediksjoner fra ekspertene i media enn dem som ikke er der.

Et viktig forbehold er både FFIs og EPJs forskning er bare basert på eksperter som selv har valgt å delta. Når det er sagt, samsvarer de overordnede resultatene fra FFIs turnering med funnene fra EPJ, som ble gjennomført på en annen måte, på et annet tidspunkt og med helt andre personer. Resultatene fra begge studier samsvarer også med den generelle forskningen innenfor kognitiv psykologi, der eksperters treffsikkerhet har vist seg å være begrenset også på helt andre områder preget av betydelig usikkerhet – fra å forutsi resultatene av forskningsprosjekter til diagnostisering av psykiske sykdommer og vurdering av tilståelser i politiavhør.[234] Dette gir grunn til å tro at funnene kan være gjeldende utover FFIs og EPJs studier alene.

Oppsummert synes det i prediksjonssammenheng å være lite hensiktsmessig å skille mellom eksperter basert på kriteriene som vanligvis brukes. Det viktigste funnene fra EPJ og GJP var tross alt at de største forskjellene i treffsikkerhet handlet om psykologiske egenskaper.

[234] For en gjennomgang av eksperters treffsikkerhet på en rekke forskjellige fagområder, som psykologi og strafferett, se Cassidy, M. F. og Buede, D. M. (2009), 'Does the accuracy of expert judgment comply with common sense: caveat emptor', *Management Decision*, 47:3, ss. 454–469. For eksempelet med prediksjon av forskningsresultater, se McBride, M. F., Fidler, F. og Burgman, M. A. (2012), 'Evaluating the accuracy and calibration of expert predictions under uncertainty: predicting the outcomes of ecological research', *Diversity and Distributions*, 18:8, ss. 782–794.

5.3 Individuelle variasjoner

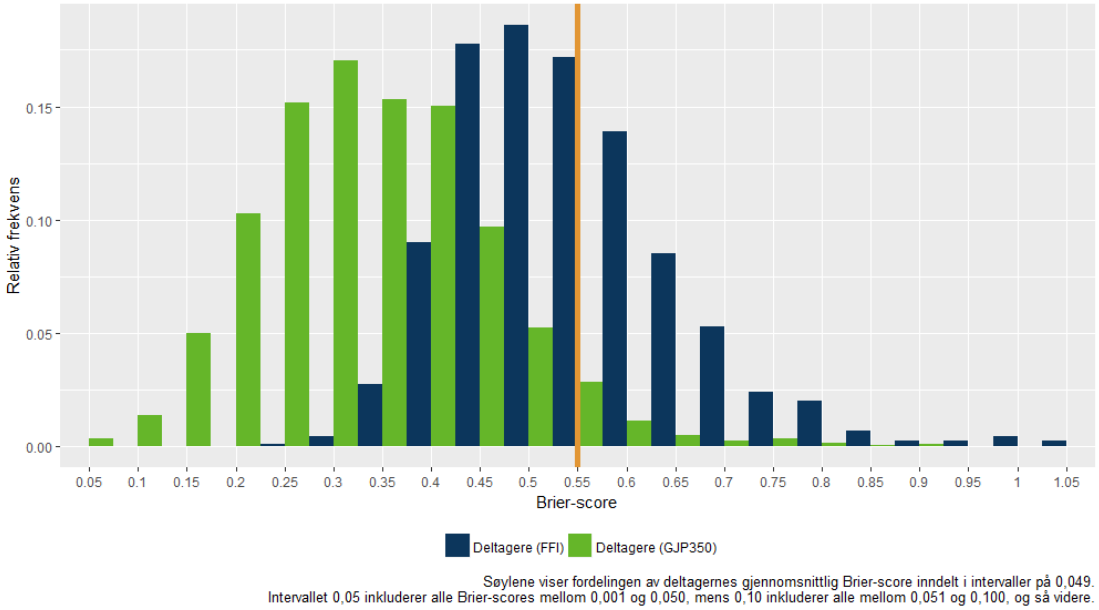
Det tredje forskningsspørsmålet i denne rapporten var: *Finnes det individer som er bedre til å predikere forsvars- og sikkerhetspolitiske utviklinger enn andre?*

For å besvare dette spørsmålet undersøkes det først hvordan treffsikkerheten til deltagerne i FFIs og GJPs turneringer fordelte seg på individuelt nivå. Deretter måles deltagernes treffsikkerhet over tid for å sjekke om de individuelle forskjellene er systematiske eller tilfeldige. Resten av kapittelet analyserer korrelasjoner mellom treffsikkerhet og alle disposisjonelle og innsatsrelaterte variabler. Her diskuteres det også om reve- og pinnsvin-stereotypene fra EPJ kan brukes til å skille mellom gode og dårlige eksperter i FFIs turnering. Til slutt sammenlignes treffsikkerheten ved bruk av ulike prediksjonsspesifikke teknikker og hvorvidt prediksjonsevnen henger sammen med grunnlaget deltagerne følte de hadde for å svare og hvor taktisk de tenkte.

I dette delkapittelet sammenlignes resultatene fra FFIs turnering med både GJP200, som var det opprinnelige datasettet brukt til å måle sammenhenger med individuelle egenskaper, og GJP350, som her brukes til å etterprøve de samme funnene basert på flere deltagere og spørsmål.[235]

5.3.1 Individuell treffsikkerhet

Det første delkapittelet viste at den gjennomsnittlige treffsikkerheten til deltagerne i FFIs turnering er betydelig dårligere enn i GJPs. Disse forskjellene bekreftes når vi ser på fordelingene av vanlige, ikke-standardiserte Brier-scores på individuelt nivå (se figur 5.21).



Figur 5.21 Deltagernes individuelle Brier-scores.

[235] For histogrammer som viser fordelingene av Brier-scores og verdier på alle uavhengige variabler, se kapittel 4 i Beadle (2021), 'Tilleggsdokumentasjon til foreløpige resultater fra FFIs prediksjonsturnering'.

På den ene siden viser figur 5.21 at formen på spredningen av Brier-scores er svært lik på tvers av turneringene. Det betyr at det er omtrent like stor forskjell mellom deltagerne i hver turnering. Formen er også litt høyre-skjev i begge datasett, som betyr at det er litt større spredning blant deltagerne med dårligst treffsikkerhet enn blant deltagerne med best.

På den annen side ligger Brier-scorene i FFIs turnering lenger til høyre enn i GJPs, som reflekterer den relativt dårligere treffsikkerheten til det store flertallet av FFIs deltagere sammenlignet med GJPs. Det er altså ikke noen få deltagere i FFIs turnering som ødelagte snittet for resten. Deltagerne i FFIs turnering er generelt dårligere enn GJPs.

Den oransje linjen viser hvilket intervall apens score faller innenfor. Denne er den samme i FFIs og GJPs turneringer, siden tilfeldig gjetning gav en Brier-score på litt over 0,5 i begge. Søyler til venstre for linjen representerer altså scores som er bedre enn apens, mens søyler til høyre er dårligere. Her bekreftes også forskjellen mellom FFIs og GJPs deltagere sammenlignet med apen. I FFIs turnering er det bare 448 (54 %) av 833 deltagere som er bedre enn tilfeldig gjetning, mens i GJP200 og GJP350 var det hhv. 667 (97 %) og 1670 (95 %) som slo apen.

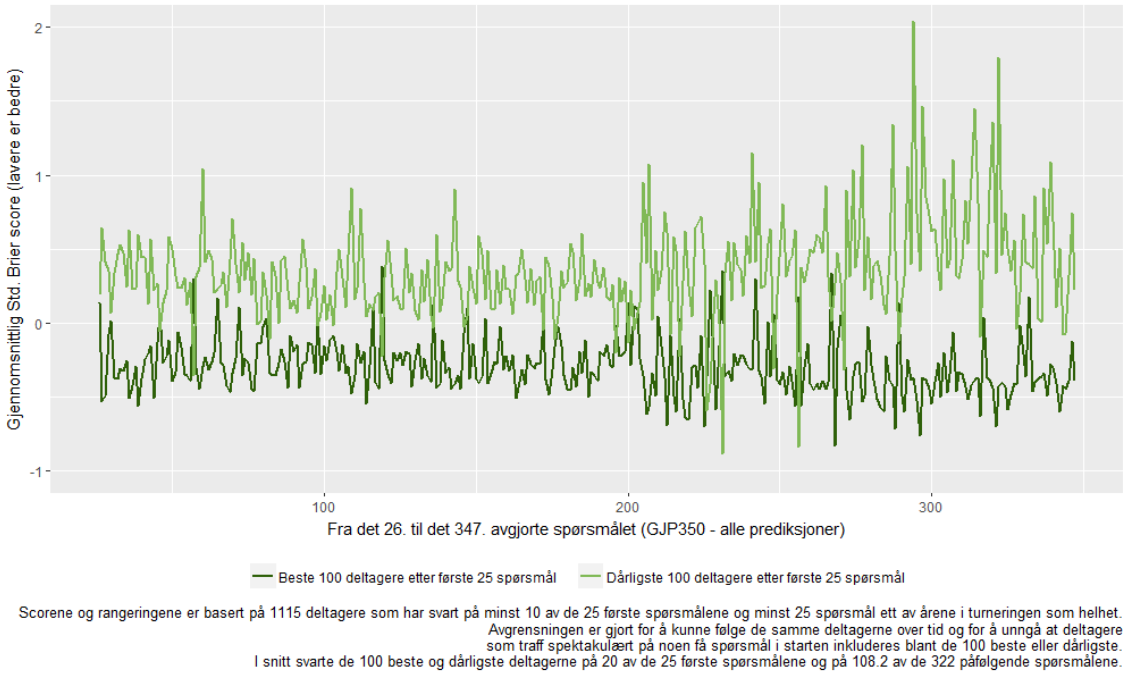
5.3.2 Treffsikkerhet over tid

Det kan imidlertid tenkes at forskjellene mellom deltagerne innad i hver turnering bare skyldes flaks. I så fall vil det være liten konsistens i treffsikkerheten deres over tid. De individuelle forskjellene holder seg imidlertid overraskende stabile i både FFIs og GJPs turneringer.

I GJP ble dette undersøkt ved å identifisere de 100 beste og dårligste deltagerne basert på de 25 første spørsmålene som ble avgjort, og deretter sammenligne treffsikkerheten deres på alle resterende spørsmål.[236] Denne analysen ble bare gjort i GJP200-artikkelen, basert på de to første årene av turneringen og noen færre deltagere enn det som finnes i studiens replikasjonsdatasett. Her har vi tatt utgangspunkt i alle 1751 deltagerne og 347 spørsmålene i replikasjonsdatasettet til GJP350, som også inkluderer resultatene fra det tredje turneringsåret. De 25 første spørsmålene som ble brukt til å identifisere de to deltagergruppene er imidlertid de samme. Her inkluderes bare deltagere som har svart på minst 10 av de første spørsmålene for å unngå at deltagere som traff eller bommet spektakulært på bare noen få spørsmål i starten regnes med.

Figur 5.22 viser de gjennomsnittlige, standardiserte Brier-scorene til de 100 beste (mørk grønn) og 100 dårligste (lys grønn) på de 322 påfølgende spørsmålene i GJP350. Figuren viser at de deltagerne som var best fra starten av traff konsistent bedre enn de dårligste resten av turneringen. Avstanden mellom scorene til de beste og dårligste deltagerne var 0,42 i GJP200 og 0,65 i GJP350.[237] Gapet var altså enda større basert på de tre første årene sammenlignet med bare de to første. Begge forskjellene mellom de beste og dårligste var statistisk signifikante.[238]

[236] I denne rapportens analyse av de 100 beste og dårligste deltagerne i FFIs og GJPs turneringer er det satt et tilleggskrav om at deltagerne måtte ha svart på minst 10 av de 25 første spørsmålene. Dette er gjort for kunne følge de samme over tid og for å unngå å inkludere deltagere som bare traff spektakulært på noen få spørsmål i starten.
[237] I GJP200-artikklen var forskjellen 0,54, som ligger midt mellom snittene basert på replikasjonsdatasettene.
[238] «Beste» vs. «dårligste» i hhv. GJP200: $t(359) = -26.19$, $p < 0.0001$; og GJP350: $t(516) = -29.95$, $p < 0.0001$.

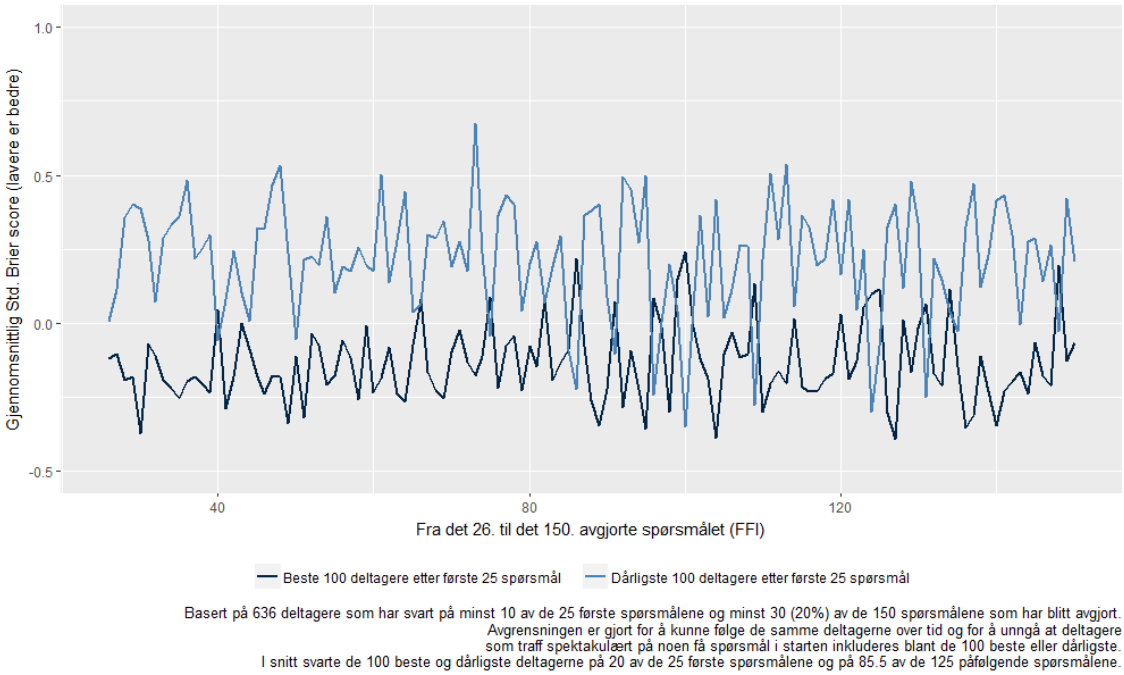


Figur 5.22 Relativ treffsikkerhet, basert på de 100 beste og 100 dårligste deltagerne i GJP.

Reanalysen av GJP bekrefter at det var systematiske forskjeller i den individuelle treffsikkerheten som holdt seg over tid. Samtidig tilhørte de fleste deltagerne i GJP eksperimentgrupper der treffsikkerheten ble forbedret underveis gjennom opplæring og lagarbeid (se underkapittel 5.1.6). Dette gjaldt ikke minst de 60 beste som hvert år ble plukket ut som superforecastere, gitt opplæring og satt i grupper med andre like gode deltagere. Det kan derfor tenkes at deltagere som gjorde det best i starten også fikk mer hjelp til å bli bedre utover i turneringen. Dette er en mulig forklaring på hvorfor gapet mellom de beste og dårligste økte fra det andre til det tredje året.

Siden det ikke ble gjort noen forsøk på å forbedre treffsikkerheten til deltagerne underveis i FFIs turnering, gir datagrunnlaget herfra et sikrere utgangspunkt for å vurdere hvor stabil den individuelle prediksjonsevnen egentlig er. Figur 5.23 viser den samme analysen som over, men basert på de 100 beste (mørk blå) og 100 dårligste (lys blå) deltagerne på de 25 første spørsmålene i FFIs turnering. Resultatene viser det samme mønstret som i GJP: Deltagerne som var best i starten av FFIs turnering fortsatte å treffe statistisk signifikant bedre på de 125 påfølgende spørsmålene som så langt er avgjort.[239] FFIs turnering støtter dermed at individuell treffsikkerhet holder seg stabil over tid, også *uten* noen forbedringstiltak. Dette bekreftes også av at Cronbachs Alpha-verdien av Brier-scorene på spørsmålene i FFIs turnering er 0,87, som tilsier en høy grad av intern konsistens. Til sammenligning lå Alpha-verdiene i GJP på rundt 0,90.[240]

[239] «Beste» vs. «dårligste» i FFIs turnering: $t(220) = -16.71$, $p < 0.0001$.
[240] Cronbach Alpha-verdien oppgis bare i GJP200-artikkelen. Der oppgis den som 0,88, mens basert på replikasjonsdatasettet er den 0,94. Basert på GJP350s datasett øker Alpha-verdien til 0,98, men det er ventet ved flere spørsmål.



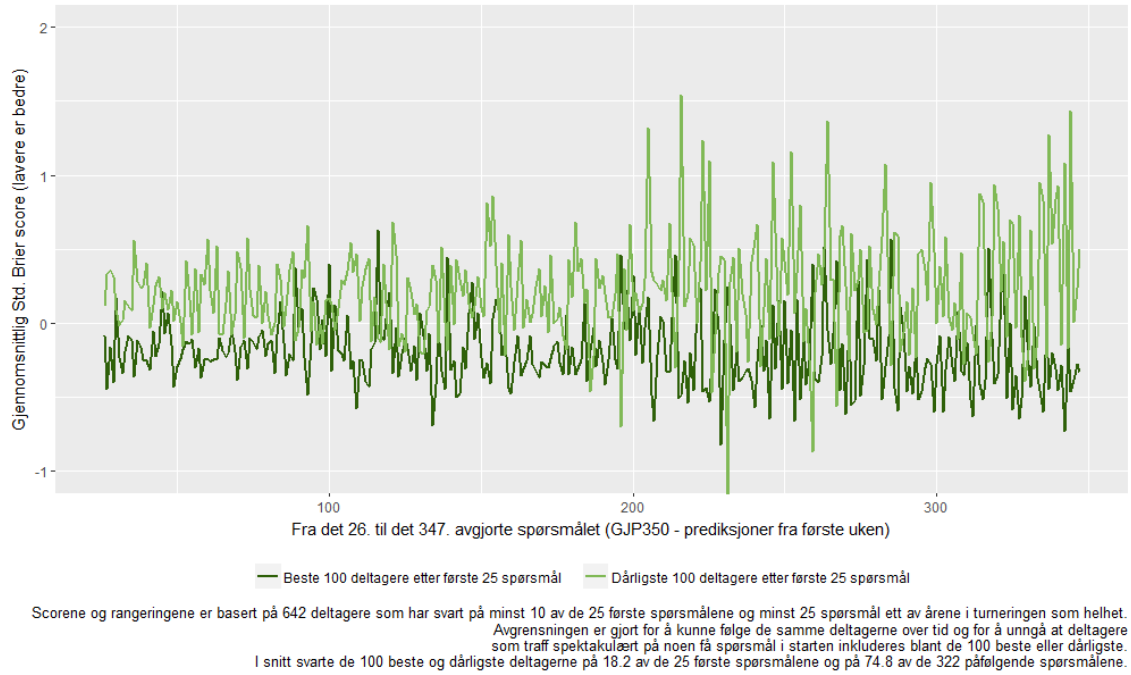
Figur 5.23 Relativ treffsikkerhet, basert på de 100 beste og 100 dårligste deltagerne i FFIs turnering.

Samtidig er avstanden mellom de standardiserte Brier-scorene til de beste og dårligste i FFIs turnering bare 0,35, som er mindre enn gapet i begge GJP-studiene. En mulig forklaring kan være at prediksjonstidspunktet som påvirket snittscoren til deltagerne i GJP generelt (se underkapittel 5.1.5), også kan tenkes å påvirke forskjellene deltagerne imellom.

Figur 5.24 viser derfor de gjennomsnittlige, standardiserte Brier-scorene til de 100 beste og dårligste deltagerne i GJP, men her er treffsikkerheten kun basert på den siste prediksjonen som er registrert i løpet av den første uken etter at spørsmålene ble publisert, slik som i FFIs turnering. Resultatet er at gapet mellom de beste og dårligste deltagerne reduseres. Avstandene mellom de standardiserte Brier-scorene er fortsatt signifikante, men blir mer enn halvert fra 0,42 til 0,16 i GJP200 og redusert med en tredel fra 0,65 til 0,43 i GJP350.[241] Ved likt prediksjonstidspunkt blir altså den relative forskjellen mellom de beste og dårligste mye mindre i GJP200 enn i FFIs turnering, men den er fortsatt litt større i GJP350.

Merk at figur 5.23 og figur 5.24 kan gi inntrykk av at avstanden mellom de beste og dårligste deltagerne er større i FFIs turnering enn i GJP, men dette skyldes at y-aksene har ulike skalaverdier. Alle figurene i dette underkapittelet viser en konsistent avstand mellom de to deltagergruppene, men denne avstanden blir mindre i GJPs figurer når prediksjonstidspunktet tas høyde for.

[241] «Beste» vs. «dårligste» i hhv. GJP200: $t(354) = -10.49$, $p < 0.0001$; og GJP350: $t(570) = -17.52$, $p < 0.0001$.



Figur 5.24 Relativ treffsikkerhet, basert på de 100 beste og 100 dårligste deltagerne i GJP.

Selv om det er statistisk signifikante forskjeller mellom de beste og dårligste deltagerne i både FFIs og GJPs turneringer, sier ikke de standardiserte Brier-scorene noe om hvor store forskjellene var i praksis. For å undersøke hvor godt deltagergruppene faktisk traff og for å kunne sammenligne dem på tvers av turneringene kan vi måle den objektive treffsikkerheten deres.

Tabell 5.7 viser den gjennomsnittlige Brier-scoren, treffprosenten og kalibreringen til de 100 beste og dårligste deltagerne i GJPs (grønn) og FFIs (blå) turneringer.[242] Dette er de samme tre målene som tidligere har blitt brukt til å diskutere hvor godt deltagerne og ekspertene treffer.

De to øverste radene viser treffsikkerheten til de beste og dårligste i GJP, basert på alle prediksjoner, slik studien selv beregnet den. De to neste radene viser scorene til GJPs beste og dårligste, basert på kun prediksjoner fra den første uken, siden dette reduserte gapet mellom turneringene ved sammenligninger av alle deltagerne. Deretter følger to rader med scorene til deltagerne i eksperimentgruppen som hverken fikk opplæring eller ble satt i grupper med andre, basert på prediksjoner fra den første uken, siden dette er de eneste deltagerne fra GJP som kan sammenlignes direkte med FFIs. De siste to radene viser treffsikkerheten til de beste og dårligste deltagerne i FFIs turnering. Innenfor hvert treffsikkerhetsmål skilles det mellom binære, kategoriske og ordinale spørsmål, siden typen spørsmål viste seg å ha betydning for sammenligninger av treffprosentene i de to turneringene.

[242] Basert på de hhv. 125 og 322 resterende spørsmålene i FFIs og GJPs turneringer etter de 25 første som i utgangspunktet ble brukt til å identifisere de 100 beste og dårligste deltagerne analysert her.

Deltagerutvalg	Brier-score	Treffprosent	Kalibrering
GJP – 100 beste	0,24 - Binære: 0,24 - Kategoriske: 0,24 - Ordinale: 0,17	84,3 % - Binære: 86,0 % - Kategoriske: 75,8 % - Ordinale: 75,6 %	–1,5 % - Binære: –1,9 % - Kategoriske: 1,6 % - Ordinale: –1,6 %
GJP – 100 dårligste	0,43 - Binære: 0,42 - Kategoriske: 0,49 - Ordinale: 0,31	64,6 % - Binære: 66,3 % - Kategoriske: 54,5 % - Ordinale: 57,6 %	10,1 % - Binære: 10,0 % - Kategoriske: 11,5 % - Ordinale: 7,2 %
GJP – 100 beste – bare første uken	0,37 - Binære: 0,36 - Kategoriske: 0,39 - Ordinale: 0,31	75,9 % - Binære: 77,9 % - Kategoriske: 72,6 % - Ordinale: 56,4 %	–0,6 % - Binære: –0,7 % - Kategoriske: –5,8 % - Ordinale: 2,8 %
GJP – 100 dårligste – bare første uken	0,50 - Binære: 0,48 - Kategoriske: 0,62 - Ordinale: 0,38	59,8 % - Binære: 62,4 % - Kategoriske: 47,0 % - Ordinale: 42,3 %	12,9 % - Binære: 12,2 % - Kategoriske: 14,4 % - Ordinale: 14,1 %
GJP – 50 beste – bare første uken – predikerte uten hjelp	0,43 - Binære: 0,43 - Kategoriske: 0,47 - Ordinale: 0,35	69,8 % - Binære: 72,2 % - Kategoriske: 60,9 % - Ordinale: 49,0 %	3,5 % - Binære: 3,3 % - Kategoriske: 3,6 % - Ordinale: 6,1 %
GJP – 50 dårligste – bare første uken – predikerte uten hjelp	0,50 - Binære: 0,50 - Kategoriske: 0,64 - Ordinale: 0,35	60,0 % - Binære: 61,7 % - Kategoriske: 48,8 % - Ordinale: 49,9 %	12,1 % - Binære: 12,3 % - Kategoriske: 11,8 % - Ordinale: 8,2 %
FFI – 100 beste – bare første uken – predikerte uten hjelp	0,46 - Binære: 0,41 - Kategoriske: 0,75 - Ordinale: 0,35	57,4 % - Binære: 74,9 % - Kategoriske: 50,4 % - Ordinale: 49,6 %	14,3 % - Binære: 8,9 % - Kategoriske: 18,7 % - Ordinale: 16,0 %
FFI – 100 dårligste – bare første uken – predikerte uten hjelp	0,63 - Binære: 0,67 - Kategoriske: 0,97 - Ordinale: 0,43	44,8 % - Binære: 58,7 % - Kategoriske: 38,4 % - Ordinale: 38,2 %	31,4 % - Binære: 25,3 % - Kategoriske: 36,1 % - Ordinale: 33,7 %

Tabell 5.7 Objektiv treffsikkerhet, basert på de 100 beste og dårligste deltagerne i FFIs og GJPs turneringer, gitt ulike prediksjonstidspunkter og eksperimentgrupper.

For det første traff både GJPs beste og dårligste deltagerne i utgangspunktet bedre enn tilfeldig gjetning, som i GJP innebar en Brier-score på 0,51 og en treffprosent på 47 %. Når treffsikkerheten kun baseres på den første ukens prediksjoner, blir imidlertid Brier-scoren til de dårligste omtrent lik apens, mens treffprosenten fortsatt er bedre. I praksis betyr dette at de dårligste del-

tagerne forble bedre enn apen til å forutsi hva som ville skje, men ikke til å vurdere usikkerhetene ved dem. GJPs beste deltagere var derimot mye bedre enn tilfeldig gjetning og nesten helt perfekt kalibrerte, uansett prediksjonstidspunkt. Treffprosentene deres tilsvarer en evne til å forutsi riktig utfall på minst tre av fire spørsmål. I tillegg var de beste deltagerne litt underkonfidente, som betyr at de kunne vært enda sikrere i sine sannsynlighetsestimater. De dårligste hadde derimot en overkonfidens på 10–13 %, som er på samme nivå som ekspertene i EPJ.

For det andre bekrefter tabell 5.7 at gapet mellom FFIs og GJPs turneringer består, også når vi bare sammenligner de beste og dårligste deltagerne fra hver. Den bekrefter samtidig at gapet reduseres når Brier-scorene baseres på samme prediksjonstidspunkt, men det forblir likevel en betydelig forskjell. Faktisk treffer FFIs beste deltagere fortsatt bare litt bedre enn GJPs dårligste. Bildet er imidlertid annerledes når vi ser på treffprosent. Ved likt prediksjonstidspunkt er de beste og dårligste deltagernes treffprosenter på binære spørsmål omtrent helt like i begge turneringer. Treffprosenten på kategoriske og ordinale spørsmål er likevel lavere i FFIs turnering. I tillegg var det en større andel spesielt kategoriske spørsmål i FFIs turnering, som i begge turneringer var assosiert med dårligere objektiv treffsikkerhet. Dette bidrar til at treffprosenten til FFIs beste og dårligste deltagere blir generelt lavere enn GJPs, men at forskjellene i evnen til å forutsi hvilket utfall som er riktig ikke er så store som det samlede snittet kan gi inntrykk av.

For det tredje er faktisk FFIs beste deltagere omtrent like treffsikre som de beste deltagerne i GJP som heller ikke fikk hjelp underveis og når scoren baseres på samme prediksjonstidspunkt. Her kan vi imidlertid bare sammenligne med de 50 beste og dårligste fra GJP, fordi det var bare 158 deltagere i denne eksperimentgruppen som svarte på minst 10 av de 25 første spørsmålene.[243] Her blir derfor forskjellene mellom de beste og dårligste GJPs små. Likevel er FFIs dårligste deltagere betydelig dårligere enn de dårligste i GJP, uansett treffsikkerhetsmål. En mulig forklaring kan være at alle deltagerne i GJP fikk en innføring i Brier-score, slik at de kanskje var mer bevisst enn FFIs på å unngå svært høye sannsynlighetsestimater på svar som de kunne bli hardt straffet for hvis de bommet. Faktisk er FFIs deltagere generelt mye dårligere kalibrerte enn GJPs. Selv om treffprosenten til FFIs beste deltagere kan måle seg med GJPs beste, utviser de en høyere overkonfidens enn GJPs dårligste. FFIs turnering deltagere oppgav en altfor høy sannsynlighet for svarene de trodde var riktig sammenlignet med hvor ofte de traff.

For det fjerde treffer FFIs beste deltagere bedre enn alle ekspertgruppene i forrige delkapittel. Til sammenligning hadde forskerne, som traff best blant ekspertene, en lavere treffprosent på 54 % og en høyere overkonfidens på 18 %. Dette tilsier at det kan være lurere å basere seg på prediksjonene til deltagere som allerede har vist seg å treffe godt, uavhengig av om de er eksperter eller ikke, enn å bruke formelle kompetansekriterier til å velge hvem vi skal høre på.

[243] Av 315 deltagere som hverken fikk opplæring eller ble satt i grupper og som oppfylte minstekravet i GJP350, var det bare 158 som både predikerte minst én gang i løpet av den første uken og minst 10 av de 25 første spørsmålene.

5.3.3 Individuelle egenskaper

Ikke bare fant GJP at det var stabile forskjeller i deltagernes prediksjonsevne, men også at treffsikkerheten korrelerte med en rekke individuelle egenskaper, som intelligens, kunnskapsnivå, tenkemåter og hvordan deltagerne oppførte seg i turneringen. For å etterprøve disse funnene ble deltagerne i FFIs turnering målt på de samme individuelle egenskapene.

Tabell 5.8 viser scorene til deltagerne på de uavhengige variablene som har blitt målt i FFIs eller GJPs turneringer.[244] I dette delkapittelet måles korrelasjonene mellom disse scorene og deltagernes treffsikkerhet. Hvordan hver variabel ble målt og scorene i FFIs komplette datasett er beskrevet i kapittel 4. Her er scorene basert på de 833 deltagerne som har svart på minst 20 % av de avgjorte, ikke stilte, spørsmålene, men dette er stort sett de samme deltagerne og verdiene er tilnærmet identiske. Alle verdiene fra GJP er basert på en reanalyse av tilgjengelige datasett. GJPs scores er også svært like med dem oppgitt i artiklene (som i tabellen gjengis i parentes).[245]

Den deskriptive analysen av FFIs og GJPs datasett har allerede vist at deltagerne i de to turneringene var svært like. De var over snittet intelligente og scoret omtrent helt likt på testene av kognitiv kontroll, politisk kunnskapsnivå og aktiv fordomsfri tenkning. Deltagerne i FFIs turnering scoret litt lavere på tallforståelse og kognitiv motivasjon. De beskrev også seg selv som likere reven enn deltagerne i GJP, men scoret likevel litt høyere på kognitiv lukking, som forbindes med nettopp pinnsvintenkning. Deltagerne i FFIs turnering svarte på omtrent like mange spørsmål og brukte omtrent like mange unike sannsynlighetsestimater som deltagerne i GJP200, men svarte på flere spørsmål og brukte flere sannsynlighetsestimater enn deltagerne i GJP350.

Ikke alle de uavhengige variablene som har blitt målt i FFIs turnering ble målt i begge GJPs studiene. Kognitiv lukking, reve- vs. pinnsvintenkning og tid brukt per spørsmål er bare analysert i GJP200-artikkelen, mens kognitiv motivasjon og ønsket om å havne blant de beste er bare med i korrelasjonsanalysen i GJP350-artikkelen. Det er også noen variabler som finnes i datasettene, men som ikke ble rapportert i artiklene. Disse er inkludert i denne rapportens reanalyse. For GJP200 betyr dette at analysen suppleres med antallet unike sannsynlighetsestimater, mens for GJP350 er deltagernes score på rev vs. pinnsvintenkning også inkludert. I GJP350 var antall unike estimater og antall besvarte spørsmål bare med i sammenligningen av superforecasterne med resten, men her er også disse variablene med i analysen av deltagerne generelt. Tiden deltagerne brukte per spørsmål ble bare oppgitt i GJP200-artikkelen, men det finnes ikke data på dette i replikasjonsdatasettet. Her rapporteres derfor bare verdien som er oppgitt i artikkelen.

[244] Merk at antallene unike sannsynlighetsestimater, spørsmål besvart og tid brukt per spørsmål i FFIs turnering er basert på alle spørsmålene deltagerne svarte på i turneringen, ikke bare spørsmålene som har blitt avgjort så langt.
[245] Det er bare GJP200-artikkelen som oppgir snittscorene til alle deltagerne. I GJP350-artikkelen er det bare snittscorene til superforecastere, de nest beste og alle andre deltagerne som oppgis. Når det tas høyde for antallet deltagere i hver av disse gruppene ligger imidlertid de aggregerte snittscorene svært nært dem i replikasjonsdatasettet.

		FFI	GJP200	GJP350
Kognitive evner	Abstrakt resonnering – Ravens (0–12)		8,02 (8,56)	7,91
	Abstrakt resonnering – Shipley-2 (0–25)			18,7
	Abstrakt resonnering – Shipley-2 (0–26)	17,66		
	Kognitiv kontroll – CRT original (0–3)	2,45	2,12 (2,10)	2,13
	Kognitiv kontroll – CRT utvidet (0–4)		3,42 (3,37)	
	Kognitiv kontroll – CRT utvidet (0–18)	14,98		14,9
	Tallforståelse (0–3)		2,69 (2,71)	
	Tallforståelse – Berlin (0–4)	2,74		3,27
Kunnskapsnivå	Politisk kunnskapsnivå - FFI (0–50) - GJP 1. år (0–35) - GJP 2. år (0–50) - GJP 3. år (0–55)	35,41	 28,85 (28,79) 36,72 (36,50)	 28,79 36,75 31,25
	Vokabular – Shipley-2 (0–40)			36,91
Tenkemåter	Aktiv fordomsfri tenkning (1–7)	6,15	5,92 (5,91)	5,94
	Kognitiv lukking (1–7)	3,88	3,34 (3,34)	
	Rev vs. pinnsvin – enkeltpåstand (1–5)		2,37	2,36
	Rev vs. pinnsvin – enkeltpåstand (1–7)	2,79		
	Rev vs. pinnsvin – test (1–7)		3,81 (3,82)	
	Motivasjon – være blant de beste (1–7)	4,98		4,91
	Kognitiv motivasjon (1–7)	5,20		5,82
Oppgavespes. ferdigheter	Antall unike estimater - Alle prediksjoner - Bare første uken	 32,99	 51,16 28,04	 45,32 25,63
	Brier-score forståelse (0–5)	0,90		
Innsats	Antall spørsmål besvart	146	130 (121)	108
	Antall prediksjoner per spørsmål		1,68 (1,58)	2,14
	Tid brukt per spørsmål	1,38	(3,6)	

Tabell 5.8 Deskriptiv analyse av individuelle egenskaper i FFIs og GJPs turneringer.

En potensielt viktig forskjell for korrelasjonsanalysene er at det finnes færre deltagere i FFIs datasett med verdier på alle uavhengige variabler enn i GJPs. Forklaringen er at FFIs deltagere kunne selv velge om de ville ta de kognitive testene som målte de disposisjonelle variablene, mens i GJP måtte alle gjennomføre disse under registreringen. I GJP200 er det likevel bare 600 (75 %) av 801 deltagere som er registrert med scores på alle variablene inkludert i artikkelens korrelasjonsanalyse. Årsaken til at det her mangler en firedel av deltagerne er at de som først registrerte seg det andre året ikke hadde mulighet til å ta testene som bare ble gjennomført det første året (kognitiv lukking og den første testen av politisk kunnskapsnivå). Av samme årsak er andelen deltagere med scores på alle variabler enda lavere i GJP350, som i tillegg til GJP200 inkluderer det tredje året av turneringen hvor det ble rekruttert spesielt mange nye deltagere. I replikasjonsdatasettet til GJP350 er det derfor bare 613 (35 %) av 1751 deltagere som er registrert med en score på alle testene som er med i artikkelens korrelasjonsanalyse.

I FFIs turnering er det 199 (24 %) av 833 deltagere som er registrert med scores på alle de samme kognitive testene som ble målt i GJP200 og GJP350.[246] Samtidig er det 634 (76 %) deltagere som tok minst én av testene.[247] I GJP er imidlertid alle deltagere registrert med en score på den aktuelle variabelen inkludert i de bivariate analysene. Den samme tilnærmingen legges derfor til grunn i denne rapporten. Selv om ikke alle deltagerne i FFIs turnering tok samtlige tester, tok de fleste mange av dem, og innsatsvariablene til alle er registrert automatisk.

Alle korrelasjoner er her målt ved hjelp av Pearsons korrelasjonskoeffisient, *r*, som er den vanligste måten å måle individuelle variasjoner innenfor kognitiv psykologi.[248] Denne koeffisienten kan variere mellom –1 og 1. Tallstørrelser som nærmer seg –1 indikerer en sterk negativ korrelasjon, mens tall nært 1 indikerer en sterk positiv korrelasjon. Siden lavere Brier-scores betyr bedre treffsikkerhet, innebærer *negative* korrelasjoner en korrelasjon med *bedre* prediksjonsevne. Tallstørrelser nær 0 betyr derimot at korrelasjonen er svak eller fraværende.

Tabell 5.9 viser alle parvise korrelasjoner mellom deltagernes treffsikkerhet og de disposisjonelle variablene og innsatsvariablene som ble målt i FFIs turnering. Korrelasjoner med p-verdier under 0.001 er uthevet med fet skrift, fordi det var dette signifikansnivået som ble brukt i GJP.

Den første kolonnen viser korrelasjonene mellom deltagernes standardiserte Brier-scores og hver uavhengige variabel. Resultatene viser at treffsikkerheten varierer sammen med kognitiv kontroll, tallforståelse, kunnskapsnivå, aktiv fordomsfri tenkning, forståelse av scoringssystemet og begge målene av innsats i turneringen. Disse korrelasjonene er også signifikante ved både mye strengere utvalgskriterier og andre korrelasjonsmål. Korrelasjonskoeffisientene forblir

[246] Siden ikke alle kognitive tester ble gjennomført i både GJP200 og GJP350, er det litt flere deltagere i FFIs turnering som gjennomførte testene sammenlignet med hver enkelt studie. 264 (32 %) av FFIs 833 deltagere tok alle de samme kognitive testene som i GJP200, mens 200 (24 %) deltagere gjennomførte alle de samme som i GJP350.
[247] Til sammen er det registrert 194 316 sannsynlighetsestimater på 150 spørsmål fra disse 634 deltagerne. Dette er omtrent halvparten av prediksjonene i GJP200 (424 259) og en firedel av GJP350s (1 070 651). Forskjellene vil imidlertid reduseres når resten av spørsmålene i FFIs turnering blir avgjort, ettersom de samme 634 deltagerne er registrert med 287 965 sannsynlighetsestimater på alle 240 spørsmål. Hvis vi bare inkluderer prediksjoner fra den første uken, slik som i FFIs turnering, reduseres gapet ytterligere. Da består selv det foreløpige datagrunnlaget fra FFIs turnering av nesten dobbelt så mange estimater som GJP200 (108 752) og omtrent like mange som GJP350 (206 847).
[248] Svartdal, F. (2021), 'korrelasjon – psykologi', *Store norske leksikon*.

nemlig omtrent uendret om analysen baseres på kun de 199 deltagerne der vi har scores på alle uavhengig variabler. P-verdiene faller, som er ventet når utvalget blir mindre, men de fleste korrelasjonene som var signifikante på 0.0001-nivå er fortsatt dette på minst 0.01-nivå. De fleste korrelasjonene forblir også signifikante med samme p-verdi-nivå ved bruk av Spearmans r_s, som i motsetning til Pearsons r ikke forutsetter en bestemt fordeling. Vedlegg B inneholder en deskriptiv analyse av scorene til de 199 deltagerne som tok alle testene. Denne viser at de scorer svært likt de 833 deltagerne. Vedlegget inneholder også en sammenligning av alle koeffisienter og p-verdier ved begge utvalg og ved bruk av både Pearsons og Spearmans korrelasjonsmål.

De øvrige kolonnene viser hvilke uavhengige variabler som korrelerer med hverandre. Dette er stort sett de samme variablene som korrelerte i GJP, som styrker reliabiliteten til målingene.[249] Tabell 5.9 viser følgende samvariasjoner:

- Deltagernes scores på alle testene av kognitive evner (abstrakt resonneringsevne, kognitiv kontroll og tallforståelse) korrelerer med hverandre. Bedre kognitiv kontroll og tallforståelse korrelerer også med høyere politisk kunnskapsnivå, mens abstrakt resonneringsevne og kunnskapsnivå gjør det ikke.

- Det er en positiv korrelasjon mellom aktiv fordomsfri tenkning og kognitiv motivasjon, som begge handler om å tenke grundig. Som ventet er det også en negativ korrelasjon mellom disse og behovet for kognitiv lukking, som handler om å hoppe til konklusjoner raskt. Det er således heller ikke overraskende at større kognitiv motivasjon korrelerer med bedre kognitive evner, som krever nettopp dypere tenkning for å score høyt.

- Bedre forståelse av scoringssystemet korrelerer også med bruk av flere unike sannsynlighetsestimater når deltagerne predikerte. En mulig forklaring er at deltagere som forstod at Brier-scoresystemet straffer høye sannsynlighetsestimater på galt svar spesielt hardt, nyanserte sine estimater mer. Det er også en positiv korrelasjon mellom tiden brukt per spørsmål og motivasjonen om å være blant de beste, antallet unike estimater og forståelsen av scoringssystemet. Alle disse variablene kan tolkes som uttrykk for aktiv deltagelse i turneringen.

Hvis flere uavhengige variabler forbundet med samme egenskap korrelerer både med hverandre og med treffsikkerheten – for eksempel abstrakt resonneringsevne og kognitiv kontroll, som er to forskjellige mål på kognitive evner generelt – styrker dette hypotesen om at denne egenskapstypen har betydning for prediksjonsevnen. Hvis det derimot bare er én av flere assosierte variabler som korrelerer med treffsikkerheten, er dette relevant for å kunne skille hvilke tester som kan

[249] Basert på reanalyser av 11 av 14 uavhengige variabler i GJPs datasett som er direkte sammenlignbare med FFIs. Variablene som ikke kan sammenlignes er Shipley-2 Block Patterns, Brier-score forståelse og tid per spørsmål. Alle sammenhenger mellom uavhengige variabler i FFIs turnering uthevet med fet skrift i tabell 5.9 var like signifikante i minst én av de to GJP-studiene, med følgende unntak: I GJP korrelerte ikke CRT original med politisk kunnskap, kognitiv motivasjon eller antall unike estimater, men derimot med aktiv fordomsfri tenkning. I GJP korrelerte CRT utvidet også med politisk kunnskap, men ikke med kognitiv motivasjon. I GJP korrelerte tallforståelse med aktiv fordomsfri tenkning, men ikke med kognitiv motivasjon. I GJP korrelerte reve- vs. pinnsvintenkning med kognitiv lukking, men ikke i FFIs turnering.

brukes for å identifisere personer som er bedre enn andre til å predikere. Dette er også tilfellet med flere av de uavhengige variablene i FFIs turnering og vil derfor diskuteres nærmere.

Tabell 5.10 sammenligner resultatene fra FFIs turnering med hvilke uavhengige variabler som korrelerte med treffsikkerheten i GJP. Fet skrift betyr at korrelasjonene er signifikante på 0.001-nivå og verdier fra GJPs artikler er oppgitt i parentes. Resultatene fra hver bivariate korrelasjonsmåling, inkludert p-verdier, oppgis i fotnotene. Fra GJP rapporteres bare resultatene fra korrelasjonene GJP350, siden dette datasettet inkluderer spørsmålene og deltagerne i GJP200. Resultatene fra GJP200 rapporteres bare hvis variabelen kun er målt i dette datasettet eller det er ulike resultater ved de to datasettene. Manglende koeffisienter i parentes skyldes at disse sammenhengene ikke ble analysert i GJPs artikler, men at dette har blitt gjort basert på datasettene.

Tabell 5.10 viser at nesten alle de samme individuelle egenskapene korrelerer med treffsikkerheten i begge turneringer. Samtidig er det noen unntak som vil diskuteres i dette underkapittelets gjennomgang av sammenhengene med hver uavhengige variabel. Mot slutten av dette underkapittelet nyanseres også funnene fra korrelasjonsanalysen basert på sammenhengene med de prediksjonsspesifikke tenkemåtene deltagerne benyttet i turneringen.

	Std. Brier-score	Shipley-2 Block Patterns	CRT original	CRT utvidet	Tallforståelse	Politisk kunnskapsnivå	Aktiv fordomsfri tenkning	Kognitiv lukking	Rev vs. pinnsvin	Motivasjon – være best	Kognitiv motivasjon	Antall unike estimater	Brier-score forståelse	Antall spørsmål besvart
Std. Brier-score	1.00													
Shipley-2 Block Patterns	-0.07													
CRT original	**-0.18**	**0.30**												
CRT utvidet	**-0.23**	**0.37**	**0.51**											
Tallforståelse	**-0.21**	**0.30**	**0.48**	**0.65**										
Politisk kunnskapsnivå	**-0.20**	-0.02	**0.20**	0.17	**0.20**									
Aktiv fordoms-fri tenkning	**-0.17**	0.13	0.17	**0.29**	0.18	**0.18**								
Kognitiv luk-king	-0.03	0.07	-0.11	-0.07	-0.04	-0.06	**-0.18**							
Rev vs. pinnsvin	-0.07	-0.07	-0.04	-0.01	0.02	0.03	-0.07	0.13						
Motivasjon – være best	**-0.20**	0.07	0.06	0.09	0.12	0.08	0.10	0.09	0.04					
Kognitiv motivasjon	-0.10	0.19	**0.23**	**0.28**	**0.21**	0.14	**0.37**	**-0.35**	-0.01	0.11				
Antall unike estimater	**-0.34**	0.13	**0.18**	**0.20**	0.16	-0.03	0.04	0.02	0.00	0.14	0.12			
Brier-score forståelse	**-0.15**	0.10	0.09	0.09	0.13	0.09	0.00	0.04	0.06	0.06	0.08	**0.21**		
Antall spørs-mål besvart	**-0.20**	0.07	0.08	0.05	0.10	0.10	0.01	0.17	0.07	0.04	-0.02	**0.28**	0.14	
Tid brukt per spørsmål	**-0.35**	0.00	0.10	**0.17**	0.15	-0.01	0.08	-0.06	0.03	0.16	0.09	**0.28**	**0.18**	0.03

Tabell 5.9 Parvise korrelasjoner i FFIs turnering. Fet skrift er signifikant på 0.001-nivå.

		FFI	GJP200	GJP350
Kognitive evner	Abstrakt resonnering – Ravens (0–12)		**–0.17 (–0.23)**	**–0.16 (–0.18)**
	Abstrakt resonnering – Shipley-2 (0–25)			**–0.24 (–0.22)**
	Abstrakt resonnering – Shipley-2 (0–26)	–0.07		
	Kognitiv kontroll – CRT original (0–3)	**–0.18**	**–0.15 (–0.15)**	**–0.16 (–0.16)**
	Kognitiv kontroll – CRT utvidet (0–4)		**–0.22 (–0.14)**	
	Kognitiv kontroll – CRT utvidet (0–18)	**–0.23**		**–0.28 (–0.23)**
	Tallforståelse (0–3)		–0.10 (–0.09)	
	Tallforståelse – Berlin (0–4)	**–0.21**		**–0.22 (–0.16)**
Kunnskapsnivå	Politisk kunnskapsnivå - FFI (0–50) - GJP 1. år (0–35) - GJP 2. år (0–50) - GJP 3. år (0–55)	**–0.20**	 **–0.16 (–0.18)** **–0.17 (–0.20)**	 **–0.14 (–0.12)** **–0.18 (–0.18)** **–0.13 (–0.14)**
	Vokabular – Shipley-2 (0–40)			**–0.10 (–0.09)**
Tenkemåter	Aktiv fordomsfri tenkning (1–7)	**–0.17**	**–0.14 (–0.10)**	**–0.11 (–0.12)**
	Kognitiv lukking (1–7)	–0.03	0.01 (0.03)	
	Rev vs. pinnsvin – enkeltpåstand (1–5)		–0.04 (0.09)	–0.05
	Rev vs. pinnsvin – enkeltpåstand (1–7)	–0.07		
	Rev vs. pinnsvin – test (1–7)		0.11	
	Motivasjon – være blant de beste (1–7)	**–0.20**		**–0.13 (–0.11)**
	Kognitiv motivasjon (1–7)	–0.10	–0.01	–0.06 **(–0.07)**
Oppgavespes. ferdigheter	Antall unike estimater - Alle prediksjoner - Bare første uken	 **–0.34**	 –0.03 **0.13**	 **–0.14** 0.00
	Brier-score forståelse (0–5)	**–0.15**		
Innsats	Antall spørsmål besvart	**–0.20**	**0.14** (0.07)	–0.04
	Antall prediksjoner per spørsmål		**–0.38 (–0.49)**	**–0.18**
	Tid brukt per spørsmål	**–0.35**	**(–0.30)**	

Tabell 5.10 Korrelasjoner mellom deltagernes standardiserte Brier-scores og individuelle egenskaper i FFIs og GJPs turneringer. Fet skrift er signifikant på 0.001-nivå.

5.3.3.1 Kognitive evner

I GJP var det en signifikant korrelasjon mellom deltagernes treffsikkerhet og scorene deres på abstrakt resonneringsevne.[250] Det er derimot ingen slik sammenheng i FFIs turnering.[251] Dette er heller ikke tilfellet om vi bare ser på de 199 deltagerne som tok på alle de kognitive testene eller bruker Spearman på begge deltagerutvalgene.[252] Dette er overraskende, siden det var en signifikant korrelasjon med begge testene av abstrakt resonneringsevne som ble brukt i GJP (*Ravens APM* og *Shipley-2 Abstraction Test*). Selv om testen som ble brukt i FFIs turnering var en annen (*Shipley-2 Block Patterns*), anses denne som et fullverdig alternativt mål på abstrakt resonneringsevne. Det er derfor ingen åpenbar forklaring på dette.

På de øvrige testene av kognitive evner sammenfaller resultatene fra FFIs turnering med GJPs. I GJP korrelerte deltagernes treffsikkerhet med begge målene av kognitiv kontroll – både på den opprinnelige testen med 3 oppgaver og den utvidede versjonen med 18 oppgaver.[253] I FFIs turnering er det også signifikante korrelasjoner med deltagernes scores på begge disse to testene.[254] I GJP korrelerte treffsikkerheten også med tallforståelsen, men det var først etter at testen brukt de to første årene ble erstattet av *Berlin Numeracy Test* i GJP350.[255] I FFIs turnering fikk deltagerne derfor bare Berlin-testen. Også her korrelerer tallforståelsen med treffsikkerheten.[256]

Berlin-testen er imidlertid den eneste uavhengige variabelen der formen på fordelingene av scores er vesentlig forskjellige i FFIs og GJPs turneringer. I GJP350 klarte hele 664 (61 %) av 1082 deltagere alle fire oppgavene riktig, mens i FFI var det bare 139 (35 %) av 395 deltagere. Den sannsynlige forklaringen er at GJP benyttet en adaptiv versjon av Berlin-testen, der deltagerne bare fikk én oppgave og bare fikk en ny hvis de svarte riktig på den foregående, mens alle deltagerne i FFIs turnering fikk alle fire oppgavene samtidig. Hvordan dette ledet til så forskjellige scores er uklart. Det kan tenkes at GJPs deltagere fokuserte mer på hver oppgave når de fikk én om gangen, mens FFIs hadde lettere for å scrolle ned til neste hvis den over var vanskelig. Selv om det var svært få deltagere (7 %) som hadde sett noen av oppgavene før i FFIs turnering, er internkonsistensen på denne testen akkurat innenfor grensen for det akseptable.[257] Hvorvidt deltagerne i GJP hadde sett testene før ble ikke undersøkt og det er umulig å måle internkonsistensen siden ikke alle deltagerne ikke fikk like mange oppgaver. Uavhengig av fordelingene av scores påvises det imidlertid en signifikant korrelasjon i begge turneringer.

[250] GJP: Ravens: $r = –0.16$, $t(1742) = –6.8$, $p < 0.0001$. Shipley-2 Abstraction: $r = –0.24$, $t(1079) = –8.06$, $p < 0.0001$.
[251] FFI: Shipley-2 Block Patterns: $r = –0.07$, $t(374) = –1.29$, $p = 0.197$.
[252] FFI: Shipley-2 Block Patterns (bare 199 deltagere): $r = –0.05$, $t(197) = –0.65$, $p = 0.515$. Spearman: 833 deltagere: $r_s = –0.08$, $p = 0.103$; 199 deltagere: $r_s = –0.05$, $p = 0.515$.
[253] GJP: CRT original: $r = –0.16$, $t(1409) = –6.14$, $p < 0.0001$. CRT extended: $r = –0.28$, $t(1084) = –9.5$, $p < 0.0001$.
[254] FFI: CRT original: $r = –0.18$, $t(393) = –3.66$, $p < 0.001$. CRT utvidet: $r = –0.23$, $t(393) = –4.66$, $p < 0.0001$. Når korrelasjonsanalysen avgrenses til de 199 deltagerne som tok alle kognitive tester faller signifikansnivået ved den opprinnelige CRT-testen til 0.05-nivå, men forblir signifikant på 0.001-nivå ved bruk av Spearmans r_s på alle 833. Her kan Spearman være en bedre egnet test, fordi fordelingene på CRT-testen er spesielt skjeve i begge turneringer, sannsynligvis fordi den bare bestod av tre oppgaver som mange av deltagerne hadde sett før.
[255] GJP: Berlin: $r = –0.22$, $t(1080) = –7.5$, $p < 0.0001$.
[256] FFI: Berlin: $r = –0.21$, $t(393) = –4.33$, $p < 0.0001$.
[257] Alpha-scoren basert på de 395 deltagerne i FFIs turnering som tok Berlin-testen var akkurat 0,6.

Oppsummert støtter resultatene fra FFIs turnering at kognitiv kontroll og tallforståelse henger sammen med treffsikkerheten, men ikke med abstrakt resonneringsevne. Disse tre målene på kognitive evner korrelerer ofte med hverandre. Det gjør de også i FFIs turnering på 0.0001-nivå. Det at deltagernes scores på abstrakt resonneringsevne korrelerer med scorene på kognitiv kontroll og tallforståelse – akkurat som i GJP – tilsier at det trolig ikke er feil ved gjennomføringen av testingen som kan forklare fraværet av en sammenheng med treffsikkerheten i FFIs turnering.

5.3.3.2 Kunnskapsnivå

I likhet med GJP er det også en signifikant korrelasjon mellom treffsikkerheten og scorene på testen av det politiske kunnskapsnivået til deltagerne i FFIs turnering.[258]

En forskjell mellom kunnskapstestene i de to turneringene var at FFIs deltagere fikk muligheten til å svare «vet ikke» på påstandene de fikk, mens GJPs ikke synes å ha fått det. I teorien kan dette ha gjort at deltagere som forsøkte å gjette på påstander de ikke visste svaret på, kan ha fått en høyere score enn dem som var ærlige på hva de kunne og ikke. En måte å kontrollere for deltagernes gjetning på kunnskapstesten på er å beregne scoren annerledes. I stedet for å bare telle antall riktige svar kan vi gi deltagerne 1 poeng for riktig svar, 0 for vet ikke og –1 for feil svar. Med denne måten å beregne deltagernes testscore blir korrelasjonen mellom politisk kunnskapsnivå og Brier-scorene til deltagerne i FFIs turnering sterkere.[259] Dette underbygger hypotesen om en sammenheng mellom deltagernes generelle kunnskapsnivå om internasjonal politikk og treffsikkerheten deres på konkrete spørsmål om hendelser innenfor dette fagfeltet.

Denne hypotesen styrkes ytterligere av at politisk kunnskapsnivå og treffsikkerhet også korrelerer innenfor separate temaer. Kunnskapstesten i FFIs turnering bestod nemlig av ti påstander hver om de fem vanligste temaene deltagerne ble bedt om å predikere: internasjonal politikk, økonomi, Russland, NATO/Europa og USA. Innenfor fire av disse fem kategoriene er det signifikante, positive korrelasjoner mellom antall riktige svar på kunnskapstesten og deltagernes snittscorene på spørsmål innenfor samme tema.[260] Jo flere riktige påstander de klarte om for eksempel internasjonal politikk, jo bedre traff de på spørsmål om dette temaet.

Disse sammenhengene med kunnskapsnivå kan fremstå som motsatte funn av det som ble gjort i forrige delkapittel, der konklusjonen var at eksperter ikke er bedre til å forutsi spørsmål på sine egne områder enn fagfolk med kompetanse på andre temaer eller regioner. De to funnene er imidlertid ikke direkte sammenlignbare. For det første baserte forrige delkapittel seg på ekspertenes *egne* vurderinger av hva de hadde kompetanse på, mens her baseres analysen på hvor mange riktige svar de *faktisk* klarte på kunnskapstester om de aktuelle temaene. For det andre undersøkte forrige delkapittel betydningen av spisskompetanse blant bare eksperter, mens her undersøkes betydning av kunnskap blant alle, inkludert amatørene, som utgjorde to tredeler av deltagerne i turneringen. Faktisk er det ingen signifikante korrelasjoner mellom treffsikkerheten

[258] FFI: Politisk kunnskapsnivå: $r = –0.20$, $t(524) = –4.58$, $p < 0.0001$. GJP: Politisk kunnskapsnivå (1–3. år): $r = –0.14$, $t(1260) = –5.19$, $p < 0.0001$; $r = –0.18$, $t(933) = –5.69$, $p < 0.0001$; $r = –0.13$, $t(1104) = –4.2$, $p < 0.0001$.
[259] FFI: Politisk kunnskapsnivå (uten «vet ikke»): $r = –0.29$, $t(524) = –6.96$, $p < 0.0001$.
[260] Internasjonal politikk: $r = –0.12$, $t(442) = –2.61$, $p < 0.01$. Økonomi: $r = –0.11$, $t(517) = –2.67$, $p < 0.01$. Russland: $r = –0.10$, $t(516) = –2.20$, $p < 0.05$. NATO/Europa: $r = –0.23$, $t(524) = –5.30$, $p < 0.0001$. USA: $r = –0.04$, $t(524) = –0.84$, $p = 0.40$. Kun basert på deltagere som svarte på minst 5 spørsmål på temaet som analyseres.

og antall riktige svar på noen av de fem temaene i kunnskapstesten når analysen avgrenses til bare ekspertene.[261] Forskjeller i kunnskapsnivå synes derfor ikke å ha betydning mellom eksperter, mens både generell og spesialisert kunnskap er en fordel når deltagerne inkluderer amatører.

I GJP hang treffsikkerheten også sammen med vokabular, men korrelasjonen var svakere og mindre sikker enn de fleste andre variablene.[262] Vokabular ble ikke målt i FFIs turnering.

5.3.3.3 Tenkemåter

Av de tre første tenkemåtene som ble målt i GJP var det bare aktiv fordomsfri tenkning som korrelerte med deltagernes treffsikkerhet.[263] Det samme funnet gjøres i FFIs turnering.[264] Korrelasjonen er imidlertid mindre sikker enn ved de andre egenskapene, fordi ved det strengeste deltagerutvalget i FFIs turnering på 199 deltagere, er korrelasjonen med denne variabelen bare signifikant på 0.05-nivå ved Pearson og ikke lenger signifikant ved bruk av Spearman.[265]

Derimot korrelerer ikke behovet for kognitiv lukking og reve- og pinnsvintenking – som var to av de andre kognitive stilene som ble brukt til å skille mellom gode og dårlige eksperter i EPJ – med prediksjonsevnen, hverken i FFIs eller GJPs turneringer.[266] Dette støtter Tetlocks egen observasjon om at skillet mellom rever og pinnsvin ble visket ut i GJP, som han spekulerte i om kunne skyldes at deltagerne ikke var helt anonyme og dermed er mer redde for å ta feil.[267]

I FFIs turnering er det i utgangspunktet heller ingen sammenheng med deltagernes kognitive motivasjon.[268] Som med testen av abstrakt resonneringsevne er dette litt overraskende, fordi deltagernes score på kognitiv motivasjon korrelerer med andre variabler som hver for seg korrelerer med treffsikkerheten (kognitiv kontroll, tallforståelse og aktiv fordomsfri tenkning). I FFIs turnering er imidlertid p-verdien (0.068) akkurat utenfor det vanligste kravet på maks 0.05 for å hevde at korrelasjonen er signifikant. Korrelasjonen med deltagernes kognitiv motivasjon blir riktig nok signifikant på 0.01-nivå om vi bruker Spearman, mens basert på bare de 199 deltagerne havner p-verdien på begge sider av 0.05-grensen, avhengig av korrelasjonsmålet.[269]

Kognitiv motivasjon er også én av to variabler der resultatene fra denne rapportens reanalyse ikke samsvarer med verdiene oppgitt i GJPs artikler. I GJP ble testen av kognitiv motivasjon bare gjennomført de to første årene, men er først rapportert i GJP350-artikkelen om superforecasterne. Her inkluderes den i listen over variabler som korrelerte med treffsikkerheten, men

[261] Internasjonal politikk: $r = –0.14$, $t(135) = –1.63$, $p = 0.11$. Økonomi: $r = –0.07$, $t(158) = –0.93$, $p = 0.35$. Russland: $r = 0.01$, $t(160) = 0.10$, $p = 0.93$. NATO/Europa: $r = –0.09$, $t(162) = –1.18$, $p = 0.24$. USA: $r = –0.03$, $t(162) = –0.39$, $p = 0.70$. Basert på 164 av de 270 ekspertene analysert i delkapittel 5.2.
[262] GJP: Shipley-2 Vocabulary: $r = –0.1$, $t(1078) = –3.43$, $p < 0.001$.
[263] GJP: Aktiv fordomsfri tenkning: $r = –0.11$, $t(1611) = –4.64$, $p < 0.0001$.
[264] FFI: Aktiv fordomsfri tenkning: $r = –0.17$, $t(475) = –3.86$, $p < 0.001$.
[265] FFI: Aktiv fordomsfri tenkning (bare 199 deltagere): $r = –0.18$, $t(197) = –2.53$, $p < 0.05$. Spearman (bare 199 deltagere): $r_s = –0,9$, $p = 0.19$.
[266] GJP200: Kognitiv lukking: $r = 0.01$, $t(664) = 0.37$, $p = 0.711$. Rev–pinnsvin: $r = –0.05$, $t(1626) = –1.94$, $p = 0.053$. FFI: Kognitiv lukking: $r = –0.03$, $t(475) = –0.76$, $p = 0.45$. Rev–pinnsvin: $r = –0.07$, $t(475) = –1.55$, $p = 0.123$.
[267] Tetlock og Gardner (2015), *Superforecasting*, fn. 23, s. 299.
[268] FFI: Kognitiv motivasjon: $r = –0.1$, $t(330) = –1.83$, $p = 0.068$.
[269] FFI: Kognitiv motivasjon (bare 199 deltagere): $r = –0.1$, $t(197) = –1.48$, $p = 0.142$. Spearman (alle 833 deltagere): $r_s = –0.14$, $p < 0.01$. Spearman (bare 199 deltagere): $r_s = –0.15$, $p < 0.05$.

den skiller seg fra alle de andre ved å ha den laveste koeffisienten og ved å være signifikant på 0.002-nivå mens alle de andre var det på 0.001-nivå.[270] I denne rapportens reanalyse er korrelasjonen mellom treffsikkerhet og kognitiv motivasjon bare signifikant på 0.05-nivå i GJP350 og ikke signifikant med deltagerne i GJP200.[271] Det er altså mer usikkert hvor mye deltagernes glede av å engasjere seg i dypere tenkning har å si for treffsikkerheten enn andre egenskaper.

Det er derimot en korrelasjon med deltagernes ønske om å havne blant de beste i både FFIs og GJPs turneringer.[272] Dette tilsier at et høyere konkurranseinstinkt – gitt at denne motivasjonen kan være en proxy for konkurranseinstinkt – er en fordel når det skal predikeres i en turnering.

5.3.3.4 Oppgavespesifikke ferdigheter

I GJP ble betydningen av «presisjon» for treffsikkerheten målt ved det gjennomsnittlige antallet unike prosentvise estimater deltagerne brukte når de predikerte på tvers av alle spørsmål.

I FFIs turnering er det en signifikant korrelasjon med antall sannsynlighetsestimater, der flere estimater korrelerer med bedre treffsikkerhet.[273] Faktisk er dette den variabelen som korrelerer nest sterkest med begge deltagerutvalgene og som er sterkest ved bruk av Spearman-testen. Reanalysen av GJP gir derimot motstridende svar. I GJP200 er det i utgangspunktet ingen sammenheng, men når vi bare ser på estimatene brukt den første uken er funnet det motsatte av FFIs.[274] Her er flere unike estimater forbundet med *dårligere* treffsikkerhet. I GJP350 er funnet derimot det samme som i FFIs turnering når vi inkluderer alle unike estimater brukt gjennom hele turneringen – at antallet estimater korrelerer med *bedre* treffsikkerhet – mens det er ingen slik sammenheng når vi bare inkluderer estimater brukt i løpet av den første ukens prediksjoner.[275]

På den ene siden er korrelasjonen med bruk av flere unike sannsynlighetsestimater blant de sikreste i FFIs turnering. På den annen side gir de motstridende resultatene i GJP grunn til ikke å generalisere funnet. Det er i tillegg betydelige forskjeller i antallet estimater brukt i de ulike utvalgene av GJPs datasett. Deltagerne i GJP brukte i snitt flere unike estimater enn FFIs når alle prediksjoner er med (45–51), mens deltagerne i FFIs brukte flest (33) sammenlignet med estimater fra bare den første uken i GJP (26–28). Siden deltagerne i GJP likevel traff best, uansett hvilke prediksjoner scorene deres baseres på, er det lite som tilsier at antallet estimater kan forklare forskjellene mellom treffsikkerheten til deltagerne i FFIs og GJPs turneringer.

En annen potensielt turneringsrelevant ferdighet var forståelsen av hvordan Brier-scorene deres ble beregnet. En forskjell mellom turneringene var at deltagerne i GJP fikk en innføring i hvordan scoringssystemet fungerte, mens det i FFIs turnering var mer opp til deltagerne selv å sette

[270] Se tabell 3 i Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters'.
[271] GJP200: Kognitiv motivasjon: $r = –0.01$, $t(780) = –0.38$, $p = 0.701$. GJP350: $r = –0.06$, $t(1396) = –2.35$, $p < 0.05$.
[272] FFI: Motivasjon: $r = –0.20$, $t(475) = –4.51$, $p < 0.0001$. GJP: Motivasjon: $r = –0.13$, $t(1080) = –4.39$, $p < 0.0001$.
[273] FFI: Antall unike sannsynlighetsestimater: $r = –0.34$, $t(831) = –10.54$, $p < 0.0001$.
[274] GJP200: Antall unike sannsynlighetsestimater (alle prediksjoner): $r = –0.03$, $t(799) = –0.96$, $p = 0.34$. Antall unike sannsynlighetsestimater (bare første uken): $r = 0.13$, $t(795) = 3.82$, $p < 0.001$.
[275] GJP350: Antall unike sannsynlighetsestimater (alle prediksjoner): $r = –0.14$, $t(1749) = –6.04$, $p < 0.0001$. Antall unike sannsynlighetsestimater (bare første uken): $r = 0$, $t(1721) = –0.02$, $p = 0.982$.

seg inn i Brier-scoren. FFIs deltagere scoret svært lavt på testen av forståelse av scoringssystemet, men denne variabelen korrelerte likevel med treffsikkerheten.[276] Korrelasjonen er signifikant på 0.001-nivå, mens de fleste de andre variablene var signifikante på 0.0001-nivå. Korrelasjonen er likevel fortsatt signifikant på 0.01-nivå når analysen avgrenses til de 199 mest aktive deltagerne. Signifikansnivået faller til 0.05-nivå ved bruk av Spearman på alle 833 deltagere, men korrelasjonen er fortsatt signifikant på 0.01-nivå basert på bare de 199 med scores på alle variabler.[277]

FFIs deltageres dårlige kjennskap til scoringssystemet kan være en del av forklaringen på gapet mellom treffsikkerheten i de to turneringene, men dette er umulig å etterprøve siden deltagernes forståelse av Brier-score ikke ble målt i GJP. En indikasjon på dette er at forskjellen mellom turneringene viskes helt ut når treffsikkerheten måles ved treffprosent, der forståelse av Brier-scoresystemet er irrelevant, når de sammenlignes ut fra prediksjoner fra den første uken og binære spørsmål. Den manglende forståelse av konsekvensene av å oppgi for høye sannsynligheter til hendelser som ikke skjer kan dermed også være en forklaring på den veldig mye høyere graden av overkonfidens blant FFIs deltagere, som aldri fikk en innføring i akkurat dette.

5.3.3.5 Innsats

Den andre variabelen hvor det er avvik mellom GJPs artikler og denne rapportens reanalyse er antallet spørsmål deltagerne svarte på. I GJP200s datasett er det en signifikant korrelasjon med treffsikkerheten, mens i artikkelen er ikke denne variabelen uthevet som signifikant på 0.001-nivå.[278] I begge tilfeller peker imidlertid koeffisienten i samme retning: Jo flere spørsmål deltagere svarte på, jo *dårligere* var treffsikkerheten deres. I GJP350 er det derimot ingen signifikante sammenhenger, hverken i datasettet eller artikkelen.[279]

I FFIs turnering er det derimot en signifikant korrelasjon mellom treffsikkerheten og antall spørsmål deltagerne svarte på.[280] Retningen er imidlertid motsatte av GJP200s: Jo flere spørsmål deltagerne svarte på, jo *bedre* treffsikkerhet. Samtidig forsvinner denne sammenhengen, uansett korrelasjonsmål, hvis vi bare ser på de 199 deltagerne med scores på alle uavhengige variabler.[281] Disse deltagerne svarte også på et mye høyere antall spørsmål enn de 833 deltagerne som bare oppfylte minstekravet.[282] Det er ingen åpenbar forklaring på de motstridende funnene

[276] FFI: Forståelse av Brier-scoresystemet: $r = –0.15$, $t(475) = –3.36$, $p < 0.001$.
[277] Spearmans r_s er kanskje den best egnede testen av Brier-score-forståelsen, fordi fordelingen av scores er svært skjev mot venstre. Spearmans ved hhv. 833 deltagere: $r_s = –0.11$, $p < 0.05$; 199 deltagere: $r_s = –0.19$, $p < 0.01$.
[278] GJP200: Antall spørsmål besvart: $r = 0.14$, $t(799) = 4.04$, $p < 0.0001$. Se tabell 2 i Mellers (2015), 'The Psychology of Intelligence Analysis', s. 8. Det oppgis ingen p-verdi for denne variabelen i artikkelen.
[279] GJP350: Antall spørsmål besvart: $r = –0.04$, $t(1749) = –1.59$, $p = 0.111$. Ikke med i tabell 3 i Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters', s. 275, med variabler som korrelerte med treffsikkerhet.
[280] FFI: Antall spørsmål besvart: $r = –0.20$, $t(831) = –6.02$, $p < 0.0001$. Merk at antall spørsmål besvart er basert på alle 240, ikke bare de 150 avgjorte, fordi variabelen er ment å måle innsatsen i turneringen som helhet.
[281] FFI: Antall spørsmål besvart (199 deltagere): $r = –0.02$, $t(197) = –0.3$, $p = 0.764$. Spearman: $r_s = –0,07$, $p = 0.313$.
[282] I FFIs turnering svarte deltagerne i snitt på 146 (61 %) av 240 spørsmål. Andelen blir høyere når bare ser på de 199 deltagerne, som i snitt svarte på 192 (80 %) av 240 spørsmål. Spredningen i antall besvarte spørsmål er, ikke overraskende, langt mindre blant de 199 mest aktive deltagerne (SD=43) enn blant alle 833 deltagere (SD=62).

fra FFIs og GJPs turneringer. Det handler imidlertid antageligvis ikke om forskjeller i gjennomføringene av turneringene, siden andelen spørsmål deltagerne svarte på er omtrent identisk i de to utvalgene der det er funnet motsatte sammenhenger.[283]

I GJP200 var antallet prediksjoner per spørsmål den variabelen som korrelerte sterkest med treffsikkerheten.[284] Det var også en signifikant korrelasjon i GJP350.[285] Som vist tidligere i dette kapittelet, kom imidlertid de fleste av GJPs prediksjoner mot slutten av spørsmålsperioden – og jo senere deltagerne predikerte, jo bedre traff de. Antallet prediksjoner er derfor kanskje et godt mål på hvor engasjerte deltagerne var i turneringen, men korrelasjonen med treffsikkerheten skyldes trolig heller at det var lettere å treffe når sluttidspunktet nærmet seg enn at det var deltagernes høyere innsats som i seg selv bidro til dette. Hyppigere oppdatering kan derfor ha stor betydning for treffsikkerheten og være et relevant mål for å identifisere hvilke deltagere en bør høre mest på i forbindelse med løpende etterretning, men ikke nødvendigvis i forbindelse med årlige trusselvurderinger eller langtidsplaner for Forsvaret.

Et mer relevant mål er tiden deltagerne brukte per spørsmål. Dette var den variabelen som korrelerte nest sterkest i GJP200.[286] Også i FFIs turnering er det en signifikant korrelasjon med den gjennomsnittlige tiden deltagerne brukte på hvert spørsmål.[287] Faktisk er tiden deltagerne brukte den variabelen som korrelerer sterkest i FFIs turnering, uansett deltagerutvalg.[288]

Korrelasjonen med tid brukt per spørsmål samsvarer også på tvers av turneringene til tross for at den gjennomsnittlige tiden deltagerne i FFIs turnering brukte var under halvparten så lang (1,4 minutter) som i GJPs (3,6 minutter). Dette tilsier at tiden deltagerne i en turnering bruker på spørsmålene er av relativt stor betydning – og at denne sammenhengen i mindre grad avhenger av hvor mye eller lite tid det er snakk om i absolutt forstand, men relativt til de andre deltagerne.

[283] I GJP200 svarte deltagerne i snitt på 130 (62 %) av 211 spørsmål (lik andel som i FFIs turnering), mens i GJP350 var snittet 108 (31 %) av 347. Årsaken er at det kom mange nye deltagere det tredje året, som dermed ikke hadde hatt mulighet til å svare på spørsmål fra de to første årene. Det forklarer hvorfor spredningen i antall spørsmål besvart er mindre i GJP200 (SD=53) enn i GJP350 (SD=78). Det er imidlertid lite som tilsier at de motstridende funnene fra FFIs og GJPs turneringer handler om forskjeller i de uavhengige variablene. I FFIs turnering er det bare en signifikant sammenheng når spredningen i antall spørsmål er *størst*, mens det bare er en signifikant sammenheng i GJP når spredningen er *minst* – og når korrelasjonene første er signifikante går koeffisientene i motsatt retning av hverandre.
[284] GJP200: Antall prediksjoner per spørsmål: $r = –0.38$, $t(799) = –11.63$, $p < 0.0001$.
[285] GJP350: Antall prediksjoner per spørsmål: $r = –0.18$, $t(1749) = –7.78$, $p < 0.0001$.
[286] Tid brukt per spørsmål er bare oppgitt i GJP200s artikkel: $r = –0.30$, $t(694) = –8.28$, $p < 0.001$. Det finnes ikke data på tid brukt per spørsmål i noen av datasettene og denne variabelen ble ikke rapportert i GJP350-artikkelen.
[287] FFI: Tid brukt per spørsmål: $r = –0.35$, $t(820) = –10.73$, $p < 0.0001$.
[288] FFI: Tid brukt per spørsmål (bare 199 deltagere): $r = –0.39$, $t(197) = –6.02$, $p < 0.0001$. Det er også den variabelen som korrelerer nest sterkest ved bruk av Spearman på det minste utvalget.

5.3.3.6 Prediksjonsspesifikke tenkemåter

I tillegg til de disposisjonelle variablene og deltagernes innsats ble det i FFIs turnering også gjort en kartlegging av hvordan deltagerne tenkte mer spesifikt når de predikerte. Deltagerne fikk valget mellom 17 tenkemåter som dekket et spekter av vanlige tilnærminger ved prediksjon (se underkapittel 4.2.4). De ble bedt om å krysse av alle tenkemåtene som var dekkende for hvordan de gikk frem når de fordelte sannsynlighetene sine. I snitt krysset deltagerne av for 5,7 tenkemåter hver.

Et interessant funn er at det er en signifikant korrelasjon mellom hvor mange prediksjonsspesifikke tenkemåter deltagerne krysset av for og treffsikkerheten: Jo flere tenkemåter de brukte, jo bedre traff de.[289] Dette er kanskje ikke overraskende gitt korrelasjonen med tid brukt per spørsmål. Sammenhengen er likevel ikke så enkel at det bare handler om å bruke flest mulige tenkemåter, fordi det er forskjeller i *hvilke* teknikker som korrelerer med bedre prediksjonsevne.

Tabell 5.11 sammenligner treffsikkerheten til hver av de 17 prediksjonsspesifikke tenkemåtene. Den mest relevante kognitive stilen, fallgruven eller metoden som tenkemåtene regnes som eksempler på, er oppgitt i parentes. De to første kolonnene til høyre viser den gjennomsnittlige, standardiserte Brier-scoren til deltagerne som brukte («Ja») og ikke brukte («Nei») hver tilnærming, med antall deltagere i parentes. Den neste kolonnen angir differansen mellom snittscorene til disse. En positiv differanse betyr deltagerne som brukte tenkemåten traff best, mens en negativ betyr at deltagerne som ikke brukte den traff best. For å undersøke om disse differansene er signifikante eller ikke, er det gjennomført t-tester av snittscorene til deltagerne som brukte og ikke brukte hver tenkemåte. Resultatene fra t-testene er rapportert i den siste kolonnen. Disse er også etterprøvd ved Wilcoxon rank-sum-test, der de samme korrelasjonene var signifikante.[290]

For å synliggjøre tenkemåtene som i FFIs turnering hang sammen med bedre treffsikkerhet, er snittscorene til deltagerne som traff best uthevet med fet skrift, hvis scorene er signifikant bedre enn de andre deltagernes, med p-verdier under 0.05. Analysen er basert på 348 deltagere som, utover å oppfylle minstekravet til deltagelse, besvarte en undersøkelse om hvordan de tenkte.

[289] FFI: Antall prediksjonsspesifikke tenkemåter: $r = –0.22$, $t(346) = –4.22$, $p < 0.0001$.
[290] I tillegg ble forskjellene mellom snittscorene til deltagerne som brukte eller ikke brukte *wisdom of the crowd* og sorte svaner også signifikante, men bare på 0.05-nivå og antall deltagere som brukte disse teknikkene var få.

Prediksjonsspesifikke tenkemåter	Ja	Nei	Diff.	t-test
Baserte meg på mageølelsen min. (intuisjon)	–0.04 (233)	**–0.10 (115)**	–0.06	$t(201) = 2.2$, $p < 0.05$
Tok utgangspunkt i en teori eller generell oppfatning jeg hadde av fenomenet fra før, og brukte denne til å vurdere hva som ville skje i dette tilfellet. (deduktiv resonnering)	–0.06 (164	–0.06 (184)	0.00	$t(346) = –0.1$, $p = 0.93$
Tok utgangspunkt i det aktuelle spørsmålet, og tenkte gjennom hva ulike teorier ville sagt om hva som ville skje. (induktiv resonnering)	–0.04 (74)	–0.07 (274)	–0.02	$t(113) = 0.8$, $p = 0.45$
Lette etter informasjon fra flere forskjellige kilder. (aktiv fordomsfri tenkning)	**–0.15 (45)**	–0.05 (303)	0.10	$t(55) = –2.6$, $p < 0.05$
Baserte meg på det første som slo meg som mest sannsynlig. (kognitiv lukking)	–0.01 (133)	**–0.09 (215)**	–0.08	$t(264) = 3.1$, $p < 0.01$
Baserte meg på hvordan utfallene har fordelt seg ved tilfeller av samme type hendelse. (referanseklasser)	**–0.11 (141)**	–0.03 (207)	0.08	$t(330) = –3.5$, $p < 0.001$
Baserte meg på tekstene og ev. figurer som fulgte med spørsmålet, og vurderte utfallene ut fra dette. (ankring)	**–0.10 (229)**	0.01 (119)	0.12	$t(175) = –4.2$, $p < 0.0001$
Estimerte hvor ofte hendelsen eller utfallet hadde skjedd før. (grunnfrekvens)	**–0.13 (142)**	–0.01 (206)	0.12	$t(345) = –5.3$, $p < 0.0001$
Baserte meg på snittet av flere, forskjellige estimater av utfallet. (wisdom of the crowd)	–0.14 (21)	–0.06 (327)	0.08	$t(22) = –1.6$, $p = 0.13$
Baserte meg på et lignende, historisk tilfelle som jeg kjente utfallet av. (bruk av én historisk analogi)	–0.03 (82)	–0.07 (266)	–0.04	$t(138) = 1.5$, $p = 0.15$
Baserte meg på flere lignende, historiske tilfeller med forskjellige utfall. (bruk av flere historiske analogier)	**–0.10 (71)**	–0.05 (277)	0.05	$t(126) = –2$, $p = 0.05$
Tok utgangspunkt i dagens situasjon/ nivå, og justerte min prediksjon deretter. (ankring)	**–0.09 (227)**	–0.00 (121)	0.09	$t(196) = –3.3$, $p < 0.01$

Baserte meg på den siste utviklingen som hadde skjedd i saken, da spørsmålet ble stilt. (tilgjengelighetsheuristikk)	–0.09 (108)	–0.05 (240)	0.04	$t(252) = –1.7$, $p = 0.09$
Fordelte prosentene slik at jeg fikk best mulig score hvis jeg traff, men samtidig unngikk å få en veldig dårlig score hvis jeg bommet. (optimalisering av Brier-score)	–0.09 (92)	–0.05 (256)	0.04	$t(209) = –1.8$, $p = 0.07$
Baserte meg på en fremskrivning av den samme utviklingen som frem til nå. (ekstrapolasjon)	**–0.13 (114)**	–0.03 (234)	0.10	$t(230) = –3.9$, $p < 0.001$
Tenkte på hva som gjorde at jeg bommet/traff på tidligere spørsmål. (post-mortem analyse)	–0.10 (41)	–0.06 (307)	0.04	$t(58) = –1.4$, $p = 0.18$
Tok hensyn til uforutsigbare, overraskende hendelser som kunne påvirke utfallet. (sorte svaner)	–0.10 (61)	–0.05 (287)	0.05	$t(83) = –1.5$, $p = 0.14$
Annet.	0.13 (12)	–0.07 (336)	–0.20	$t(11) = 1.6$, $p = 0.14$

Tabell 5.11 Antall og andel deltagere som brukte ulike prediksjonsspesifikke tenkemåter.

Tabell 5.11 viser at det er flere teknikker, som i eksisterende forskning forbindes med bedre treffsikkerhet, der snittscoren til deltagerne som brukte dem var bedre enn dem som ikke gjorde det. Resultatene fra FFIs turnering bekrefter at intuisjon er en dårlig måte å predikere internasjonal politikk på, ettersom deltagerne som brukte mageølelsen traff signifikant dårligere enn dem som ikke gjorde det. Det samme gjorde deltagerne som baserte seg på det første som slo dem som mest sannsynlig, som er et uttrykk for kognitiv lukking. Selv om korrelasjonsanalysen ikke fant en korrelasjon mellom treffsikkerheten og deltagernes behov for kognitiv lukking, er det her en signifikant forskjell mellom snittscorene til deltagerne som praktiserte det og ikke.

Derimot traff deltagere som baserte seg på grunnfrekvens (hvor ofte samme hendelse har skjedd før) og referanseklasser (hvordan tidligere, lignende utfall har fordelt seg) signifikant bedre enn dem som ikke benyttet disse teknikkene. Disse to tenkemåtene er som nevnt kjennetegn på Kahnemans «utsideperspektiv». I tillegg var referanseklasser den teknikken som i GJPs opplæringsmodul hang tettest sammen med bedre treffsikkerhet. De relativt få deltagerne som lette etter informasjon fra flere kilder og som baserte seg på flere lignende, historiske tilfeller, traff også signifikant bedre enn dem som ikke gjorde det. Disse to teknikkene er praktiske eksempler på aktiv fordomsfri tenkning.

Oppsummert tilsier disse funnene at, selv om deltagerne scoret generelt høyt på aktiv fordomsfri tenkning, var det bare et mindretall som praktiserte det, men de som gjorde det traff også bedre.

Et annet interessant funn er at deltagerne som tok utgangspunkt i dagens situasjon traff signifikant bedre enn dem som ikke gjorde det – enten det var å bruke teksten og figurene som fulgte spørsmålene, justere prediksjonene sine ut fra dagens nivå eller ekstrapolere den samme utviklingen som frem til nå. Alle disse tilnærmingene er eksempler på ankring, der vi risikerer å legge for stor vekt på ett enkelt holdepunkt som vi får presentert, hvis dette ankeret er dårlig. Siden det å basere seg på teksten og figurene som fulgte med spørsmålet var den nest vanligste tilnærmingen (etter mageføIelsen) kan dette tyde på at bakgrunnsinformasjonen deltagerne fikk, som var ankeret i denne sammenheng, ikke bidro til å svekke treffsikkerheten. Ved siden av grunnfrekvens var dette den tilnærmingen der forskjellen var størst mellom snittscoren til dem som brukte den og ikke.

En mulig forklaring på hvorfor deltagerne som baserte seg på dagens situasjon traff bedre er at det er generelt mer kontinuitet enn endring i internasjonal politikk – eller i alle fall i spørsmålene som ble stilt i FFIs turnering. Det siste kan vi måle ved å beregne treffsikkerheten en ville fått dersom en alltid gikk for *status quo*, altså det svaralternativet som representerte ingen endring fra dagens situasjon eller nivå.[291] Dette representerer en annen, relativt enkel algoritme som en kunne forvente at deltagerne skulle slå, ved siden av tilfeldig gjetning.

Her er treffsikkerheten ved *status quo* målt ved to forskjellige tilnærminger: 1) 100 % sannsynlighet på dagens situasjon, som utgjør en bastant prediksjon på videreføring av dagens situasjon, og 2) 75 % sannsynlighet på dagens situasjon, mens resten av prosentene fordeles på de nærmeste svaralternativene, som representerer et mer forsiktig svar på samme utfall.[292]

I FFIs turnering gir disse to *status quo*-tilnærmingene Brier-scores på hhv. 0,66 og 0,53, sammenlignet med deltagernes 0,51 på samme spørsmål.[293] Deltagerne var altså bedre til å anslå sannsynlighetene for riktig utfall enn den mest bastante tilnærmingen, men sliter med å slå en mer forsiktig fordeling av sannsynlighetene på de samme svarene. Dette bekreftes av treffprosentene, der begge *status quo*-algoritmene gir en treffprosent på 54 %, siden de alltid pekte på samme svar, mens deltagernes treffprosent er 51 %. Deltagerne var altså omtrent akkurat like gode til å forutsi riktig utfall som ved en ren videreføring av utviklingen frem til nå. Dette er ikke overraskende siden et flertall av deltagerne sa at de baserte seg på nettopp dagens situasjon.

Problemet med en slik *status quo*-tilnærming er at treffsikkerheten avhenger av hvilke spørsmål som stilles. Mens en slik tilnærming antageligvis vil gi høy treffprosent på regelmessige spørs-

[291] På spørsmål om en bestemt hendelse vil skje fremover, baseres svaret på hva utfallet har vært like langt bakover i tid. Hvis Nord-Korea gjennomførte en prøvesprengning det siste året, blir svaret «ja» på spørsmål om landet vil gjøre dette det neste året. På spørsmål uten data på den bestemte hendelsen kan historikk innenfor samme fenomen brukes. På spørsmål om når Trump reiser på statsbesøk til Storbritannina, blir svaret det tredje året i presidentperioden, fordi begge presidentene før ham kom på statsbesøk det tredje året. På spørsmål om en hendelse vil skje på en bestemt måte, baseres sannsynligheten for dette på frekvensstatistikk, hvis dette er tilgjengelig. På spørsmål om det neste islamistiske terrorangrepet i Europa vil involvere bruk av eksplosiver, og en tredel angrepene de siste årene har involvert dette, oppgis sannsynligheten for denne hendelsen som 33,3 %. På noen spørsmål er det ikke mulig å etablere en *status quo*, som på spørsmål om hvem som vinner Nobels fredspris.

[292] Ved å nedjustere fra 100 % til 75 % blir også straffen ved feil svar mindre, slik at det kontrolleres for at det ikke er Brier-score-systemet i seg selv som er avgjørende for algoritmens relative treffsikkerhet.

[293] Basert på 141 av 150 spørsmål der med et svaralternativ som representerer en videreføring av dagens situasjon.

mål om utviklinger med små svingninger, som befolkningsvekst eller arbeidsledigheten i perioder preget av økonomisk stabilitet. Det er imidlertid først når det oppstår usikkerheter rundt dagens situasjon at det blir interessant å stille spørsmål om den videre utviklingen. Det var for eksempel den økonomiske usikkerheten som følge av covid-19-pandemien som gjorde det relevant å spørre om Norges BNP-utvikling de neste årene i FFIs turnering. Den begrensede treffprosenten til *status quo*-algoritmen illustrerer derfor problemet med å basere seg på dagens situasjon i prediksjon generelt; nemlig at det ikke tas hensyn til nye hendelser eller trendbrudd som gjør det relevant å predikere i første omgang.

Det er derimot ingen signifikant forskjell mellom snittscorene til deltagerne som baserte seg på induktiv eller deduktiv tenkning (som var et viktig skille mellom reve- og pinnsvinekspertene), hvorvidt de baserte seg på snittet av flere forskjellige estimater (også kjent som *wisdom of the crowd*), hvorvidt de baserte seg på ett lignende, historisk tilfelle (i motsetning til flere forskjellige) eller hvorvidt de baserte seg på den aller siste utviklingen som hadde skjedd da spørsmålet ble stilt (som kan være et mål på i hvor stor grad de var påvirket av nyhetsdekningen).

Deltagerne som forsøkte å optimalisere sine Brier-scores traff heller ikke signifikant bedre enn dem som ikke forsøkte dette. Dette til tross for korrelasjonen mellom deltagernes forståelse av scoringssystemet og treffsikkerheten, og at én av fire deltagere tok hensyn til scoringssystemets regler når de fordelte sannsynlighetene sine. Det kan tyde på at aktive forsøk på å svare taktisk ikke nødvendigvis lønte seg. Det utgjorde heller ingen signifikant forskjell for treffsikkerheten at deltagerne gjennomførte *post-mortem*-analyser (reflekterte rundt hvorfor de hadde bommet eller truffet sist) eller forsøkte å ta hensyn til sorte svaner (uforutsigbare, overraskende hendelser). Disse to siste teknikkene var også med i GJPs opplæringsmoduler, men heller ikke der ble det funnet en sammenheng med bedre treffsikkerhet.[294]

FFIs resultater nyanserer dagens forskning om betydningen av kognitive stiler for treffsikkerheten på spørsmål om internasjonal politisk. På den ene siden var det viktigste funnet fra EPJ at en først og fremst kunne skille mellom gode (rever) og dårlige (pinnsvin) eksperter basert på *hvordan* de tenkte. Revene var mer fordomsfri, tenkte induktivt og samlet informasjon fra flere kilder før de predikerte, mens pinnsvinene hadde et større behov for kognitiv lukking, tenkte deduktivt og stolte mye på sin eksisterende kunnskap. På den annen side ble skillet mellom rever og pinnsvin mindre viktig i GJP, der deltagernes behov for kognitiv lukking og hvorvidt de kjente seg mest igjen i reve- eller pinnsvin-måtene å tenke på ikke hang sammen med treffsikkerheten. Dette er heller ikke tilfellet i FFIs turnering. Samtidig viser analysen over at måtene deltagerne tenkte på når de faktisk predikerte likevel er av betydning. I boks 5.3 diskuteres det derfor om reve- vs. pinnsvin-dikotomien likevel kan være relevant for å skille mellom gode og dårlige eksperter, basert på prediksjonsspesifikke i motsetning til generelle tenkemåter.

[294] Tvert imot ble det i GJP funnet en sammenheng mellom bruk av *post-mortem*-analyser og *dårligere* treffsikkerhet. Forklaringen er at i GJP ble effekten av de forskjellige teknikkene målt på spørsmålsnivå og at *post-mortem*-analyser normalt bare ble gjort på spørsmål der deltagerne gjorde det dårlig, ikke på spørsmålene hvor de traff godt.

Er skillet mellom rever og pinnsvin relevant for ekspertene i FFIs turnering?
Av 270 eksperter i FFIs turnering var det 105 som besvarte undersøkelsen om prediksjonsspesifikke tenkemåter. Selv om utvalget er lite er det signifikante forskjeller mellom snittscorene til dem som brukte eller ikke brukte fem av tenkemåtene.

I likhet med deltagerne flest treffer ekspertene som søkte informasjon fra flere kilder, brukte grunnfrekvensen, tok utgangspunkt i dagens situasjon og ekstrapolerte utviklingen signifikant bedre enn dem som ikke gjorde disse tingene.[295] Informasjonsinnsamling er et reveaktig trekk og den teknikken hvor det i FFIs turnering er størst forskjell mellom ekspertenes snittscores. Dette tilsier at selv eksperter, som i utgangspunktet kan mye om temaene de ble spurt om, kan ha noe å hente ved å søke mer informasjon før de uttaler seg om fremtidige utviklinger. I likhet med deltagerne flest treffer også ekspertene som baserte seg på det første som slo dem som mest sannsynlig dårligere enn dem som ikke gjorde dette.[296] Dette er et eksempel på kognitiv lukking, som var et kjennetegn på pinnsvinene i EPJ.

I FFIs turnering er det imidlertid en overraskende korrelasjon mellom ekspertenes score på den generelle testen av kognitiv lukking og treffsikkerheten deres. Her treffer ekspertene med *høyere* score på kognitiv lukking *bedre*.[297] Dette er motsatt av funnet i EPJ, der ekspertene som scoret høyere på kognitiv lukking traff *dårligere*. I praksis utøver likevel de mest treffsikre ekspertene i FFIs turnering en mer reveaktig tilnærming, slik som i EPJ, gjennom bredere informasjonssøk, bruk av grunnfrekvens, justeringer ut fra dagens situasjon og ekstrapolasjon. Reve- og pinnsvin-dikotomien kan derfor være nyttig til å skille mellom god og dårlig prediksjonsadferd, også blant eksperter, men ikke som stereotyper til å identifisere hva slags typer eksperter som i utgangspunktet vil være best.

Boks 5.3 Reve- og pinnsvin-eksperter i FFIs turnering.

[295] Lete etter mer informasjon vs. ikke: $t(21) = -3.4$, $p < 0.01$. Grunnfrekvens vs. ikke: $t(100) = -2.8$, $p < 0.01$. Utgangspunkt i dagens situasjon vs. ikke: $t(55) = -2.1$, $p < 0.05$. Ekstrapolering vs. ikke: $t(81) = -2.3$, $p < 0.05$.
[296] Basere seg på det første som slår en som mest sannsynlige vs. ikke: $t(85) = 2.1$, $p < 0.05$.
[297] Korrelasjon mellom eksperters standardiserte Brier-scores og kognitiv lukking: $r = -0{,}16$, $t(151) = -1{,}99$, $p < 0.05$.

Selv om deltagerne brukte flere teknikker når de predikerte, var det altså ikke vilkårlig hvilke metodiske tilnærminger de tok. Dette reflekteres også i tabell 4.2, som viser hvilke tenkemåter som var de mest brukte.[298] Den første kolonnen angir antall tenkemåter deltagerne krysset av for, med antall deltagere som krysset av for hvert av dem i parentes. Den andre kolonnen oppgir de vanligste tenkemåtene blant deltagerne som brukte det aktuelle antallet tenkemåter, med andelen av disse deltagerne som brukte hver av dem i parentes. Tenkemåter som i FFIs turnering er forbundet med dårligere treffsikkerhet er markert med rød skrift, mens tenkemåter forbundet med bedre er markert i grønt. Grå farge betyr ingen sammenheng med treffsikkerheten.

Antall tenkemåter	**Vanligste tenkemåter (andel deltagere som brukte disse)**
1 (11)	Intuisjon (64 %)
2 (11)	Kognitiv lukking (73 %), Intuisjon (55 %)
3 (31)	Intuisjon (87 %), Kognitiv lukking (36 %), Ankring – spørsmålstekst (36 %)
4 (61)	Ankring – dagens situasjon (67 %), Intuisjon (58 %), Ankring – spørsmålstekst (51 %), Deduktiv resonnering (34 %)
5 (71)	Ankring – spørsmålstekst (73 %), Ankring – dagens situasjon (62 %), Deduktiv resonnering (55 %), Intuisjon (54 %), Kognitiv lukking (37 %)
6 (48)	Ankring – dagens situasjon (77 %), Intuisjon (69 %), Ankring – spørsmålstekst (63 %), Deduktiv resonnering (42 %), Kognitiv lukking (42 %), Referanseklasser (40 %)
7 (34)	Ankring – spørsmålstekst (82 %), Intuisjon (74 %), Ankring – dagens situasjon (71 %), Grunnfrekvens (65 %), Deduktiv resonnering (56 %), Referanseklasser (53 %)
8 (33)	Ankring – spørsmålstekst (88 %), Ankring – dagens situasjon (82 %), Intuisjon (79 %), Referanseklasser (70 %), Grunnfrekvens (67 %), Deduktiv resonnering (61 %)
9 (21)	Ankring – spørsmålstekst (91 %), Referanseklasser (86 %), Ankring – dagens situasjon (86 %), Intuisjon (76 %), Grunnfrekvens (76 %), Tilgjengelighetsheuristikk (67 %)
10 (10)	Ankring – spørsmålstekst (100 %), Grunnfrekvens (90 %), Ankring – dagens situasjon (90 %), Intuisjon (80 %), Én historisk analogi (80 %), Flere historiske analogier (70 %)

Tabell 5.12 Vanligste prediksjonsspesifikke tenkemåter per antall tenkemåter brukt totalt.

[298] Listen avgrenset til opptil ti tenkemåter, som inkluderer 331 (95 %) av alle 348 deltagere. I tillegg var det 7 deltagere som krysset av for 11 tenkemåter, 8 som krysset av for 12 og 1 som krysset av for 13.

Blant deltagerne som brukte relativt få tenkemåter (1–3) var de vanligste intuisjon og kognitiv lukking, som var forbundet med dårligere treffsikkerhet. Jo flere teknikker deltagerne krysset av for, jo vanligere ble imidlertid bruken av tenkemåter forbundet med bedre treffsikkerhet, som ankring, referanseklasser, grunnfrekvens og bruken av flere historiske analogier. Blant deltagerne som krysset av for minst seks forskjellige tenkemåter var det teknikker forbundet med bedre treffsikkerhet som dominerte.

Dette mønstret forklarer også den tidligere nevnte korrelasjonen mellom treffsikkerhet og antall tenkemåter deltagerne brukte: Nesten alle deltagerne baserte seg på mageføelelsen, uansett antall tenkemåter, men dess flere andre tilnærminger de brukte, jo flere og mer treffsikre teknikker ble denne supplert med. Blant de vanligste tenkemåtene finner vi også ankring basert på spørsmålsteksten. Denne skiller seg imidlertid fra de andre tenkemåtene ved at hvorvidt denne bidrar til bedre eller dårligere treffsikkerhet vil avhenge av kvaliteten på informasjonen deltagerne får. Betydningen av denne tenkemåten er derfor mindre generaliserbar enn de andre metodene.

Hvor store er forskjellene mellom de prediksjonsspesifikke tenkemåtene i praksis? Tabell 5.13 viser den objektive treffsikkerheten til deltagerne som brukte eller ikke brukte de ni tilnærmingene der det er en signifikant forskjell mellom deres standardiserte Brier-scores. Treffsikkerhetsmålene til deltagerne som traff best er uthevet i fet skrift.

Den enkelte tenkemåten forbundet med høyest presisjon er leting etter informasjon fra flere forskjellige kilder. Deltagerne som dette skiller seg først og fremst ut med en lavere overkonfidens (14 %) enn dem som ikke gjorde det (19 %). Forskjellen i treffprosent er imidlertid ikke like stor, som betyr at deltagerne som søkte bredere etter informasjon oppgav litt lavere sannsynligheter til det de trodde var riktige svar enn de andre. Det samme mønsteret – der differansen mellom gradene av overkonfidens er større enn forskjellene i treffprosent – gjelder alle de ni prediksjonsspesifikke tenkemåtene analysert her. Dette tilsier at de «riktige» tenkemåtene har bidratt til å gjøre deltagerne litt bedre kalibrerte, som her betød en litt mindre overdreven selvsikkerhet.

På den ene siden er ikke deltagerne som brukte tenkemåtene forbundet med bedre treffsikkerhet så mye bedre enn dem som ikke brukte dem. Forskjellene er i gjennomsnitt bare 0,03 målt i Brier-score, 1,5 prosentpoeng i treffprosent og 2,33 prosentpoeng i overkonfidens. På den annen side er nesten ingen av treffsikkerhetsmålene til deltagerne undersøkt her dårligere enn snittene i turneringen som helhet, der Brier-scoren er 0,52, treffprosenten 51 % og overkonfidensen 21 %.

Dette betyr at deltagerne analysert i tabell 5.13 er basert på et noe skjevt utvalg som består av de relativt sett beste deltagerne i turneringen. Av de 348 deltagerne som besvarte undersøkelsen om prediksjonsspesifikke tenkemåter var 224 (64 %) av dem blant de 50 % beste av alle 833 deltagerne i turneringen som helhet. Det betyr at forskjellene mellom de prediksjonsspesifikke tenkemåtene som er funnet her gjelder selv når vi analyserer et utvalg med en overvekt av de beste deltagerne. Det gir også grunn til å tro at forskjellene mellom teknikkene ville vært større om alle deltagerne i turneringen var med i datagrunnlaget. Betydningen av å bruke de riktige prediksjonsspesifikke tenkemåtene er således antageligvis litt større det analysen her viser.

	Brukte			Brukte ikke		
Prediksjonsspesifikke tenkemåter	**Brier-score**	**Treff-prosent**	**Kalib-rering**	**Brier-score**	**Treff-prosent**	**Kalib-rering**
Baserte meg på magefølelsen min. (intuisjon)	0,50	51,7 %	18,9 %	**0,48**	**53,8 %**	**16,1 %**
Lette etter informasjon fra flere forskjellige kilder. (aktiv fordomsfri tenkning)	**0,45**	**55,4 %**	**14,1 %**	0,50	51,9 %	18,6 %
Baserte meg på det første som slo meg som mest sannsynlig. (kognitiv lukking)	0,51	50,9 %	20,4 %	**0,48**	**53,3 %**	**16,5 %**
Baserte meg på hvordan utfallene har fordelt seg ved tilfeller av samme type hendelse. (referanseklasser)	**0,47**	**53,7 %**	**15,8 %**	0,51	51,5 %	19,5 %
Baserte meg på tekstene og ev. figurer som fulgte med spørsmålet, og vurderte utfallene ut fra dette. (ankring)	**0,47**	**53,2 %**	**16,2 %**	0,53	50,8 %	21,5 %
Estimerte hvor ofte hendelsen eller utfallet hadde skjedd før. (grunnfrekvens)	**0,46**	**54,1 %**	**15,4 %**	0,51	51,2 %	19,8 %
Baserte meg på flere lignende, historiske tilfeller med forskjellige utfall. (bruk av flere historiske analogier)	**0,47**	**53,6 %**	**16,6 %**	0,50	52,1 %	18,4 %
Tok utgangspunkt i dagens situasjon/nivå, og justerte min prediksjon deretter. (ankring)	**0,48**	**53,3 %**	**16,9 %**	0,52	50,7 %	20,0 %
Baserte meg på en fremskrivning av den samme utviklingen som frem til nå. (ekstrapolasjon)	**0,46**	**54,2 %**	**14,7 %**	0,51	51,5 %	19,6 %

Tabell 5.13 Objektiv treffsikkerhet, basert på tenkemåter med signifikant ulik treffsikkerhet.

Den statistiske analysen over kan ikke si noe om *hvorfor* deltagerne tenkte forskjellig når de predikerte; bare hvilke tilnærminger som korrelerte med treffsikkerheten. Siden deltagerne ikke ble bedt om å oppgi hvilke teknikker de brukte på hvert enkelt spørsmål, er det heller ikke mulig å måle den direkte sammenhengen mellom treffsikkerheten og tenkemåter på spørsmålsnivå. Alle deltagerne som besvarte undersøkelsen om prediksjonsspesifikke tenkemåter gav imidlertid kvalitative beskrivelser av hvordan de gikk frem. Basert på disse beskrivelsene er det særlig to aspekter som synes å ha påvirket hvordan de predikerte og som kan undersøkes nærmere.

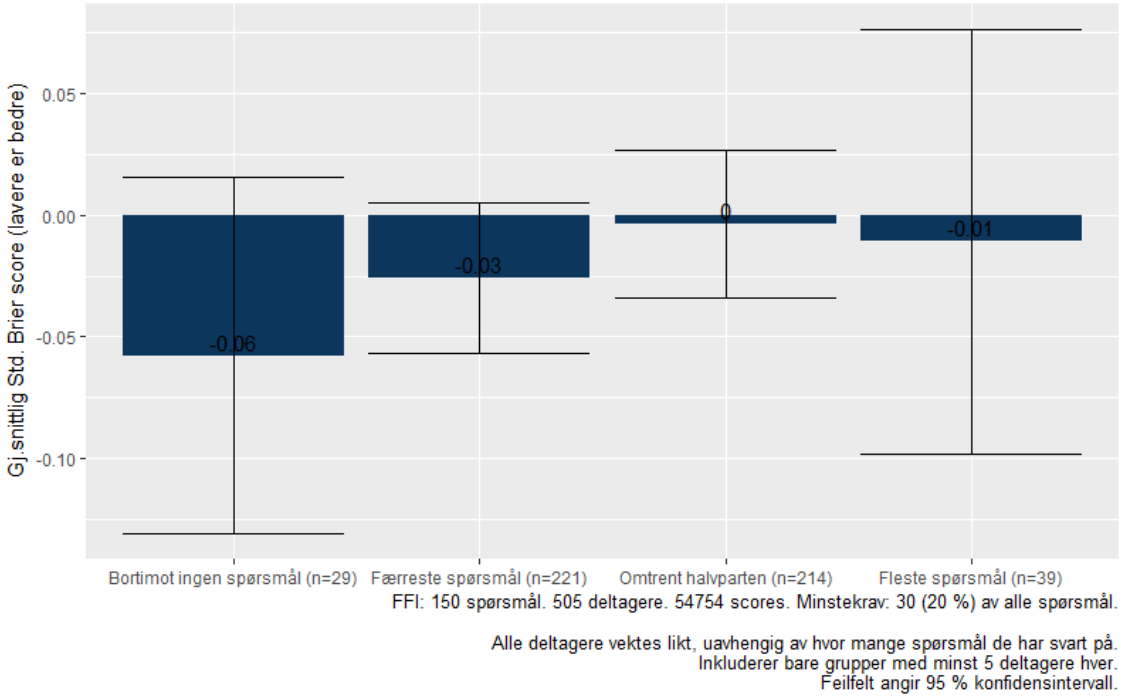
Det første aspektet var *hvor godt grunnlag* deltagerne følte de hadde for å svare på spørsmålene. «Grunnlag» ble her definert som f.eks. bakgrunnskunnskap om temaet. Svært mange av deltagerne beskrev at de baserte seg på mageføelsen når de følte de ikke hadde noe grunnlag for å mene noe om temaet, mens de tenkte seg grundigere om og brukte flere teknikker på spørsmål de følte at de kunne noe om. På spørsmål om hvor stor andel av spørsmålene deltagerne mente at de hadde et godt grunnlag for å svare på, var det imidlertid bare 8 % som svarte «de fleste spørsmålene», mens 42 % svarte «omtrent halvparten av spørsmålene» og 44 % «de færreste spørsmålene».[299] Dette kan forklare hvorfor de mer intuitive teknikkene var de mest brukte.

Figur 5.25 viser den gjennomsnittlige, standardiserte Brier-scoren til deltagerne basert på hvor stor andel av spørsmålene de mente at de hadde godt grunnlag for å svare på. Resultatene viser at deltagerne som mente de hadde dårligst grunnlag (på bortimot ingen eller de færreste spørsmålene) traff i snitt best. Alle konfidensintervallene overlapper imidlertid og ingen av forskjellene i figuren er signifikante.[300] Det kan altså ikke hevdes at deltageres egne oppfattelser av hvor godt grunnlag de har for å svare henger sammen med hvor godt de treffer generelt.[301]

[299] Basert på svar fra 505 av 833 deltagere som svarte på minst 20 % av spørsmålene i turneringen. I tillegg var det to deltagere som svarte «På bortimot alle spørsmålene», men disse er for få til å inkluderes i analysen.

[300] «På de fleste spørsmålene» vs. «På omtrent halvparten av spørsmålene»: $t(48) = –0{,}16$, $p = 0.88$. «På de fleste spørsmålene» vs. «På de færreste spørsmålene»: $t(49) = 0.33$, $p = 0.74$. «På de fleste spørsmålene» vs. «På bortimot ingen av spørsmålene»: $t(66) = 0.84$, $p = 0.40$. «På omtrent halvparten av spørsmålene» vs. «På de færreste spørsmålene»: $t(433) = 1.02$, $p = 0.31$. «På omtrent halvparten av spørsmålene» vs. «På bortimot ingen av spørsmålene»: $t(39) = 1.40$, $p = 0.17$. «På de færreste spørsmålene» vs. «På bortimot ingen av spørsmålene»: $t(40) = 0.82$, $p = 0.42$. Ingen av disse er signifikante ved bruk av Wilcoxon rank-sum-tester heller.

[301] Samtidig var det relativt få deltagere som svarte at de hadde et godt grunnlag på bortimot ingen eller de fleste spørsmålene. Formen på fordelingene av scores på disse to alternativene gjør det også usikkert om forutsetningene for bruk av konfidensintervall eller de statistiske testene er oppfylte.

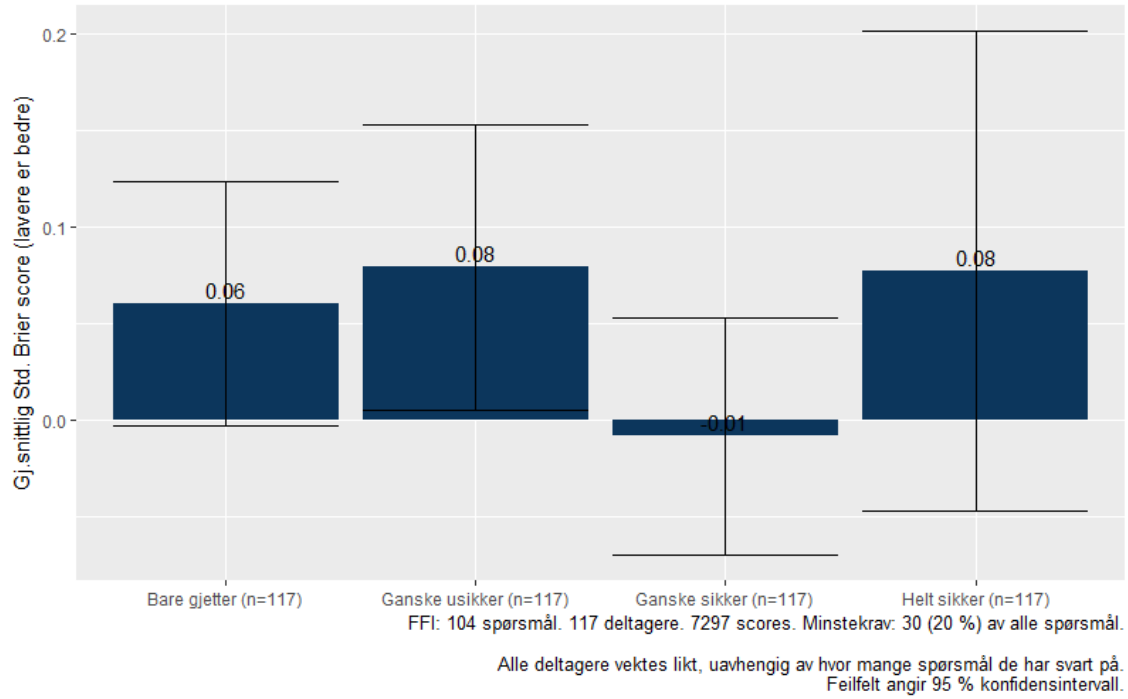


Figur 5.25 Relativ treffsikkerhet, basert på grunnlag for å svare på spørsmålene generelt.

Hvor godt grunnlag deltagerne mente de hadde for å predikere ble også målt på spørsmålsnivå. På hvert spørsmål fikk alle deltagerne en valgfri tilleggsmulighet til å oppgi hvor sikre de følte seg på svarene sine. Her kunne de velge mellom «bare gjetter», «ganske usikker», «ganske sikker» og «helt sikker». Figur 5.26 på neste side viser deltagernes gjennomsnittlige treffsikkerhet på spørsmålene hvor de oppgav samme grad av sikkerhet. Det var 797 av 833 deltagere som oppgav selvsikkerheten på minst ett spørsmål. Til sammen oppgav disse 34 685 svar. Her inkluderes imidlertid bare de 117 deltagere som brukte alle fire alternativene på minst to spørsmål. Dette er nødvendig for å kunne måle relative forskjeller basert på hvor sikre de var, ikke forskjeller mellom deltagerne generelt. Analysen er også avgrenset til 104 av de 150 spørsmålene, fordi muligheten til å oppgi grad av sikkerhet ble introdusert i den 10. spørsmålsrunden.

Figur 5.26 viser ingen åpenbar sammenheng mellom treffsikkerhet og hvor (u)sikre deltagerne var på hvert enkelt spørsmål. Alle konfidensintervallene overlapper. Snittscorene på spørsmål hvor deltagerne var «ganske sikker» er likevel signifikant høyere enn på spørsmål hvor de var «ganske usikker». Dette er imidlertid bare på 0.05-nivå ved bruk av Wilcoxon-testen, mens ved t-testen faller p-verdien akkurat utenfor denne grensen.[302] Ingen av de andre snittene er signifikant forskjellige.[303] Det er spesielt interessant at snittscorene er tilnærmet like på spørsmålene hvor deltagerne «bare gjetter» og på spørsmålene hvor de var «helt sikker».

[302] «Ganske usikker» vs. «Ganske sikker»: $t(116) = 1.92$, $p = 0.06$. Fordelingene av scores er litt skjeve til samme side, som tilsier at Wilcoxon-testen kanskje er den beste egnede.
[303] «Bare gjetter» vs. «Ganske usikker»: $t(116) = –0.44$, $p = 0.66$. «Bare gjetter» vs. «Ganske sikker»: $t(116) = 1.72$, $p = 0.08$. «Bare gjetter» vs. «Helt sikker»: $t(116) = –0.24$, $p = 0.81$. «Ganske usikker» vs. «Helt sikker»: $t(116) = 0.03$, $p = 0.98$. «Ganske sikker» vs. «Helt sikker»: $t(116) = –1.28$, $p = 0.21$.



Figur 5.26 Relativ treffsikkerhet, basert på grunnlag for å svare på spørsmålsnivå.

En annen måte å måle sammenhengen mellom treffsikkerhet og hvor sikre deltagerne følte seg på hvert spørsmål er å rangere de fire alternativene – bare gjetter (1), ganske usikker (2), ganske sikker (3) og helt sikker (4) – for så å beregne en gjennomsnittlig «selvsikkerhetsscore» for hver deltager og måle hvordan denne korrelerte med Brier-scoren. Her inkluderes bare deltagere som har oppgitt selvsikkerhet på minst 20 % av de 104 spørsmålene, som utgjør 569 av 833 deltagere. Det er imidlertid ingen signifikant korrelasjon mellom deltagernes gjennomsnittlige selvsikkerhet og treffsikkerhet på tvers av de samme spørsmålene.[304]

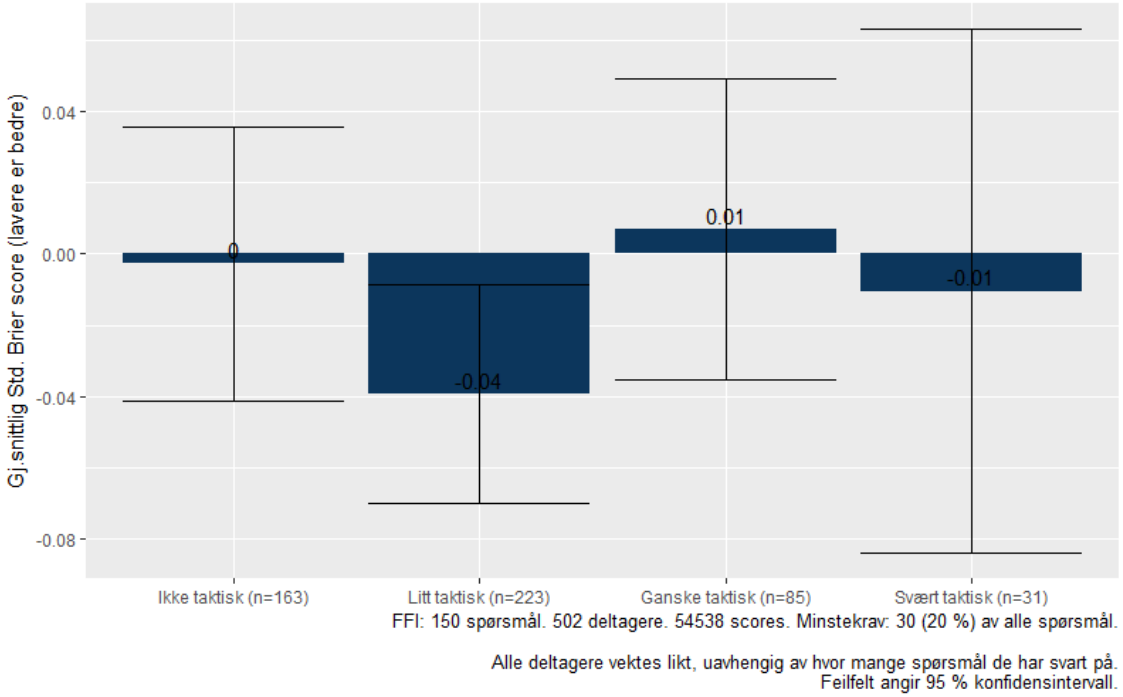
Selv om mange av deltagerne bare gjettet på spørsmål som de ikke følte de kunne noe om, mens de tenkte seg grundigere om på spørsmål de følte de hadde eller burde ha peiling på, er det altså ingen sammenheng mellom treffsikkerheten deres og hvor stor andel av spørsmål de mente de hadde et godt grunnlag for å svare på. Det kan tyde på at deltagere som var ganske sikre traff bedre enn deltagere som var ganske usikre på hver enkelt spørsmål, men denne sammenhengen er relativt usikker og det er ingen forskjell mellom dem som var bare gjettet og var helt sikre. Fraværet av en klar sammenheng mellom treffsikkerhet og hvor godt grunnlag deltagerne følte de hadde for å svare er også i tråd med rapportens tidligere observasjoner om den begrensede betydningen av temaspesifikk kompetanse for fagfolks treffsikkerhet (se underkapittel 5.2.2).

Det andre aspektet som ifølge deltagerne påvirket hvordan de predikerte, var *hvor taktisk de tenkte*. «Taktisk» ble her definert som hvor mye deltagerne vektla konkurranseaspektet, f.eks. for å få en best mulig score/plassering, i motsetning til hvor sannsynlig de «egentlig» trodde at

[304] Korrelasjon med selvsikkerhetsscore: $r = 0.03$, $t(567) = 0.69$, $p = 0.49$.

utfallene var. På direkte spørsmål om hvor taktisk de tenkte når de besvarte spørsmålene i turneringen svarte to tredeler at de tenkte «litt», «ganske» eller «svært» taktisk, mens én tredel svarte «ikke taktisk».[305] Det er imidlertid ikke gitt at deltagerne som tenkte mer taktisk traff bedre.

Figur 5.27 viser deltagernes gjennomsnittlige, standardiserte Brier-scores, basert på de fire alternativene de fikk når de ble spurt om hvor taktisk de tenkte. Alle konfidensintervallene overlapper. Det er en signifikant forskjell mellom deltagerne som svarte «litt taktisk» og «ganske taktisk» ved bruk av Wilcoxon-testen på 0.05-nivå, men ikke ved t-testen.[306] Her er det verdt å merke seg at deltagerne som bare tenkte litt taktisk faktisk var bedre enn deltagerne som tenkte ganske taktisk. Det er derimot ingen signifikante forskjeller mellom noen av de andre alternativene.[307] Det er altså ingen mønster som viser at jo mer taktisk deltagerne tenkte, jo bedre traff de. I beste fall kan resultatene tilsi at en forsiktig taktisk tilnærming er bedre enn ingen eller en offensiv tilnærming.

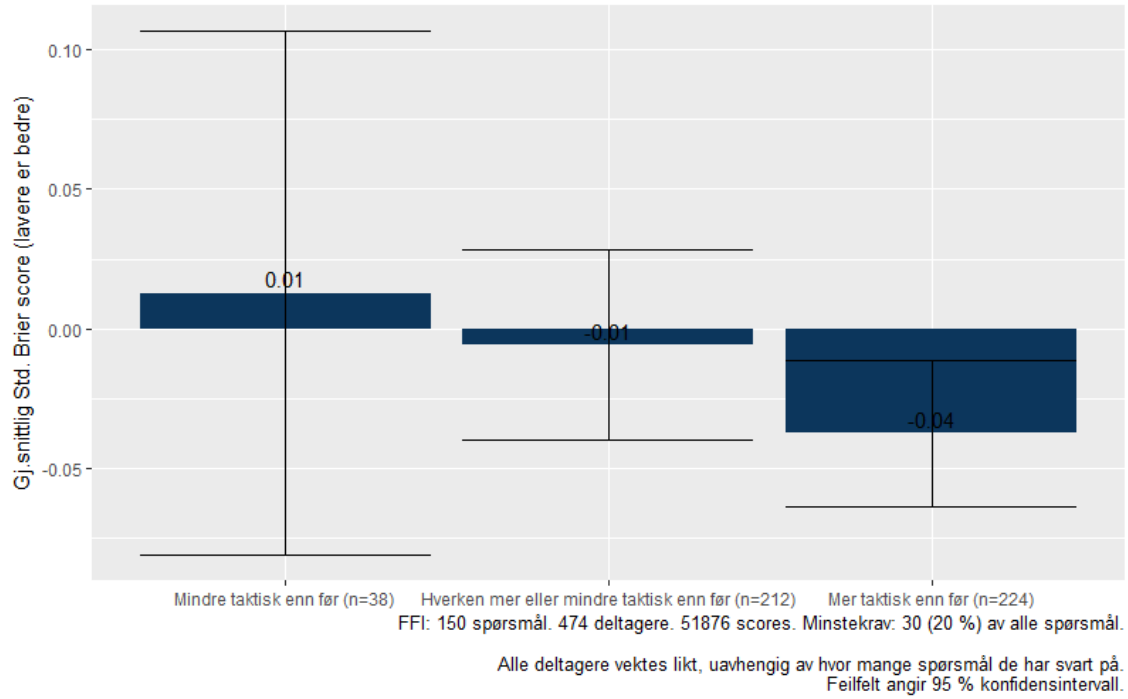


Figur 5.27 Relativ treffsikkerhet, basert på grad av taktisk tenkning.

[305] Basert på svar fra 502 av 833 deltagere som svarte på minst 20 % av spørsmålene i turneringen.
[306] «Ganske taktisk» vs. «Litt taktisk»: $t(179) = 1.75$, $p = 0.08$. Fordelingene av scores er litt skjeve til samme side, som tilsier at Wilcoxon-testen kanskje er den beste egnede.
[307] «Svært taktisk» vs. «Ganske taktisk»: $t(52) = –0.42$, $p = 0.68$. «Svært taktisk» vs. «Litt taktisk»: $t(42) = 0.73$, $p = 0.47$. «Svært taktisk» vs. «Ikke taktisk»: $t(49) = –0.19$, $p = 0.85$. «Ganske taktisk» vs. «Ikke taktisk»: $t(208) = 0.34$, $p = 0.73$. «Litt taktisk» vs. «Ikke taktisk»: $t(336) = –1.46$, $p = 0.14$.

Fordi undersøkelsen om graden av taktisk tenkning ble gjennomført halvveis ut i turneringen, ble deltagerne også spurt om de tenkte *mer* eller *mindre* taktisk enn før. Av deltagerne som svarte på dette spørsmålet var det nesten ingen som svarte at de tenkte mindre taktisk, mens rundt halvparten svarte mer taktisk enn før og nesten like mange ingen av delene.[308]

Figur 5.28 viser de gjennomsnittlige, standardiserte Brier-scorene basert på hvorvidt deltagerne svarte mer eller mindre taktisk enn før. Snittscoren til deltagerne som tenkte «mer taktisk enn før» var bedre enn dem som ikke gjorde det, men ingen av forskjellene er statistisk signifikante.[309] Dette samsvarer med deltagernes egne beskrivelser, der mange av dem forklarer at de av taktiske årsaker endret måten de fordelte prosentene sine underveis, men at de hadde svært ulike erfaringer med hvorvidt denne forskjellen medførte bedre eller dårligere treffsikkerhet.



Figur 5.28 Relativ treffsikkerhet, basert på endring i grad av taktisk tenkning.

[308] Basert på svar fra 502 av 833 deltagere som svarte på minst 20 % av spørsmålene i turneringen.
[309] «Mer taktisk enn før» vs. «Hverken mer eller mindre taktisk enn før»: $t(402) = -1.45$, $p = 0.15$. «Mer taktisk enn før» vs. «Mindre taktisk enn før»: $t(43) = -1.04$, $p = 0.30$. «Hverken mer eller mindre taktisk enn før» vs. «Mindre taktisk enn før»: $t(48) = -0.38$, $p = 0.71$.

5.3.4 Diskusjon

For det første støtter FFIs turnering ett av de viktigste funnene fra både EPJ og GJP, nemlig at det er systematiske forskjeller i hvor gode enkeltpersoner er til å predikere. Denne evnen holder seg også stabil over tid. Et nytt funn fra FFIs turnering er at forskjellen mellom deltagerne bestod selv om treffsikkerheten deres ikke ble forsøkt påvirket underveis. Dette styrker konklusjonen fra GJP om at treffsikkerhet er en individuell evne – og at det basert på få spørsmål er mulig å skille personer som mest sannsynlig vil treffe bedre enn andre også på sikt.

For det andre støtter FFIs turnering de fleste, men ikke alle, tidligere funn om hvilke individuelle egenskaper som henger eller ikke henger sammen med bedre treffsikkerhet. Tabell 5.14 viser alle egenskapene som har blitt undersøkt i både FFIs og GJPs turneringer, fordi det finnes tidligere forskning eller teoretisk belegg for å anta at de henger sammen med treffsikkerhet. Her er alle egenskapene som korrelerer med deltagernes treffsikkerhet i begge turneringer uthevet i fet skrift. De øvrige egenskapene korrelerer enten i ulike retninger eller bare i én eller ingen av turneringene. Egenskaper som bare ble målt i én turneringer er i parentes.

Kognitive evner	*Kunnskap*	*Tenkemåter*	*Oppgavespesifikke evner*	*Innsats i turneringen*
Abstrakt resonneringsevne **Kognitiv kontroll** **Tallforståelse**	**Politisk kunnskap** (Vokabular)	**Aktiv fordomsfri tenkning** Kognitiv lukking Rev vs. pinnsvin **Motivasjon – være best** Kognitiv motivasjon	Unike sannsynlighetsestimater (Brier-score forståelse)	Spørsmål besvart (Prediksjoner per spørsmål) **Tid brukt per spørsmål**

Tabell 5.14 Individuelle egenskaper som korrelerer med treffsikkerheten. Fet skrift er signifikant i samme retning på 0.001-nivå i både FFIs og GJPs turneringer.

Til tross for at deltagerne i FFIs og GJPs turneringer består av personer som ble rekruttert til forskjellige turneringer på ulike tidspunkter i to forskjellige land, scorer de svært likt på alle de individuelle egenskapene de har blitt målt på. Både FFIs og GJPs deltagere anses derfor her som representative utvalg av personer som deltar i denne typen prediksjonsturneringer. Det vil si at egenskaper som korrelerer med treffsikkerheten i FFIs og GJPs datasett kan forventes å gjøre det samme i andre turneringer. Slik generalisering forutsetter imidlertid at de statistiske testenes forutsetninger er oppfylte og at korrelasjonene er statistisk signifikante. I dette delkapittelet er alle analyser derfor etterprøvd med tester og korrelasjonsmål med forskjellige forutsetninger

uten at resultatene endrer seg nevneverdig.[310] Signifikansnivået er også satt til 0.001, som er relativt strengt. Forenklet forklart betyr dette at det må være mindre enn 0,1 % sannsynlig at korrelasjonen er tilfeldig for at vi skal anse den som signifikant. Til sammenligning er grensen som normalt brukes i samfunnsvitenskapelig forskning 0.05, altså en 5 % sannsynlighet for tilfeldige sammenhenger.

Korrelasjonene med egenskapene uthevet i tabell 5.14 fremstår derfor som relativt sikre statistisk sett. De representerer også de mest robuste funnene, fordi det bare er egenskapene med signifikante korrelasjoner på 0.001-nivå i *både* FFIs og GJPs turneringer som her er uthevet. De samme individuelle egenskapene korrelerer også omtrent like mye og på stort sett samme signifikansnivå med deltagernes Brier-scores, treffprosent og kalibrering, slik at de følgende resultatene ikke er avgrenset til relativ prediksjonsevne, men også handler om absolutt treffsikkerhet.

Av de tre kognitive evnene som korrelerte med høyere treffsikkerhet i GJP gjør to av dem – kognitiv kontroll og tallforståelse – det samme i FFIs turnering. Dette tilsier at evnene til å unngå å falle for de mest intuitive, men gale, svarene, og til å forstå tallkonsepter som prosenter og sannsynligheter, henger sammen med bedre prediksjonsevne. I FFIs turnering er det imidlertid ingen signifikant korrelasjon med deltagernes abstrakte resonneringsevne, som handler om evnen til å trekke slutninger fra enkeltobservasjoner til mer generelle prinsipper. Dette er kjernen i induktiv tenkning, som kjennetegnet revenes tilnærming til prediksjon, i motsetning til pinnsvinenes deduktive applikasjon av teoriene de allerede kunne fra før på alle nye hendelser.

Abstrakt resonneringsevne brukes også ofte som et mål på intelligens. At deltagernes score på denne testen ikke korrelerte med treffsikkerheten i FFIs turnering er derfor også relevant for den bredere litteraturen om intelligens og prestasjonsevne. På den ene siden mener noen forskere at det er eksperters eksepsjonelle kognitive evner som gjør dem så gode, for eksempel den usedvanlige høye intelligensen til de aller flinkeste sjakkspillerne i verden.[311] På den andre siden mener andre at, over en viss terskel, er det mengdetrening som betyr mest.[312] I GJP ble det funnet delvis støtte for det siste synet, nemlig at trening økte treffsikkerheten. Dette ble undersøkt

[310] De statistiske sammenhengene i FFIs turnering holder seg også ved mindre utvalg og alternative korrelasjonsmål. I FFIs turnering forblir nesten alle sammenhenger signifikante på minst 0.01-nivå ved bruk Pearsons *r* om analysen avgrenses til de 199 deltagerne hvor det foreligger data på alle uavhengige variabler. Unntakene er den opprinnelige, korteste testen av kognitiv kontroll og aktiv fordomsfri tenkning, som fortsatt er signifikante på 0.05-nivå, og antall spørsmål besvart, der sammenhengen ikke lenger er signifikant. Nesten alle sammenhengene er også signifikante på minst 0.01-nivå ved bruk av Spearmans r_s på alle 833 deltagerne som analysen i utgangspunktet er basert på. Unntakene er korrelasjonen med kognitiv motivasjon som ved Spearmans blir signifikant på 0.01-nivå, mens p-verdien faller til 0.05-nivå ved Brier-score-forståelse. Ved bruk av Spearmans på de 199 deltagerne som tok alle testene er det også de samme tre variablene som ikke lenger er signifikante på 0.01-nivå som ved bruk av Pearsons.
[311] Plomin, R., Shakeshaft, N. G., McMillan, A. og Trzaskowski, M. (2014), 'Nature, nurture, and expertise', *Intelligence*, 45, ss. 46–59.
[312] Ericsson, K. A. (2014), 'Why expert performance is special and cannot be extrapolated from studies of performance in the general population: A response to criticisms', *Intelligence*, 45, ss. 81–103.

ved å sammenligne korrelasjonene mellom deltagernes abstrakte resonneringsevne og treffsikkerheten på hhv. de 50 første og 50 siste spørsmålene. Her fant de at korrelasjonen ble svakere utover i turneringen, når deltagerne hadde fått øvd seg mer.[313]

I FFIs turnering er det imidlertid ingen signifikant korrelasjon mellom deltagernes abstrakte resonneringsevne og treffsikkerheten, uansett om vi sammenligner scorene deres på de 30, 40 eller 50 første og siste publiserte spørsmålene som så langt er avgjort.[314] Det er altså ingen korrelasjon, selv ikke helt i starten av turneringen, der deltagerne ikke hadde hatt mulighet til å øve seg særlig mye. Resultatene sår derfor tvil om betydningen av intelligens for prediksjonsevne, i alle fall innenfor rammene av den treningen de fikk i FFIs turnering, som var lik GJPs.

I likhet med GJP korrelerer høyere score på testen av kunnskap om internasjonal politikk med bedre treffsikkerhet i FFIs turnering. Dette er i tråd med funnene fra delkapittel 5.2, der fagfolk traff bedre enn amatører. Samtidig viser det seg at jo flere riktige svar deltagerne klarte om bestemte temaer, som internasjonal politikk, jo bedre traff de også på spørsmål innenfor samme tema. Dette tilsier en sammenheng med temaspesifikk kunnskap som forrige delkapittel ikke fant hos ekspertene, der det ikke var forskjeller mellom eksperter som predikerte innenfor eller utenfor sine egne kompetanseområder. Det viser seg imidlertid ikke å være temaspesifikke korrelasjoner mellom kunnskapsnivå og treffsikkerhet når vi bare ser på ekspertene. Oppsummert tilsier dette at relevant kunnskap ikke har betydning for forskjellene mellom eksperter, men korrelerer med bedre treffsikkerhet når deltagerne også inkluderer amatører.

Funnene fra den statistiske analysen av både FFIs og GJPs datasett gir bare delvis støtte til tidligere funn om betydningen av tenkemåter. Bare to av de fem kognitive stilene, som det var grunn til å tro kunne henge sammen med treffsikkerheten, gjør det i begge turneringer. På den ene siden korrelerer mer aktiv fordomsfri tenkning med bedre treffsikkerhet i FFIs turnering, mens behovet for kognitiv lukking og deltagernes vurderinger av hvor reve- eller pinnsvinaktig de tenker ikke gjør det. Dette er akkurat de samme funnene som i GJP. På den annen side utfordrer resultatene fra både FFIs og GJPs turnering hovedfunnene fra EPJ, der alle disse tre variablene skilte de gode og dårlige ekspertene. Ønsket om være blant de beste ble aldri målt i EPJ, men korrelerte med treffsikkerheten i begge turneringer. Av alle tenkemåtene er imidlertid denne variablene vanskeligst å kunne se for seg at kan være relevant i andre sammenhenger, siden fremtidsstudier sjeldent gjennomføres som en konkurranse. I tillegg er det også tvil om betydningen av kognitiv motivasjon, som handler om gleden av å engasjere seg i aktiviteter som krever tenkning, som prediksjonsturneringer. Mens det i GJPs artikler var en litt mindre signifikant korrelasjon med deltagernes score på kognitiv motivasjon enn med de andre variablene, er signifikansen svakere eller helt borte i reanalysen av GJPs egne datasett og i FFIs turnering.

[313] Mellers mfl. (2015), 'The Psychology of Intelligence Analysis', s. 8. Ifølge denne artikkelen faller korrelasjonen mellom treffsikkerhet og Ravens-scorene fra –0.22 basert de 50 første spørsmålene til –0.10 basert på de 50 siste. Dette er ikke etterprøvd her, men artikkelen oppgir lignende resultater ved de 40 og 30 første og siste spørsmålene.
[314] Korrelasjoner mellom Shipley-2 Block Patterns-scorene og deltagernes standardiserte Brier-scores basert på hhv. 50 første spørsmål: $r = –0.04$, $t(320) = –0.76$, $p = 0.45$; 50 siste spørsmål: $r = –0.01$, $t(320) = –0.26$, $p = 0.80$; 40 første spørsmål: $r = –0.07$, $t(312) = –1.29$, $p = 0.20$; 40 siste spørsmål: $r = –0.02$, $t(312) = –0.43$, $p = 0.67$; 30 første spørsmål: $r = –0.09$, $t(296) = –1.50$, $p = 0.13$; 30 siste spørsmål: $r = –0.01$, $t(296) = –0.21$, $p = 0.83$. Kun basert på deltagere som svarte på minst ett av de første og siste spørsmålene som sammenlignes her.

Samtidig kaster spørreundersøkelsene fra FFIs turnering nytt lys over sammenhengene mellom treffsikkerhet og hvordan deltagerne tenker når de predikerer i motsetning til hvordan de tenker til vanlig. Selv om det ikke er en korrelasjon med deltagernes generelle behov for kognitiv lukking, er det en signifikant forskjell mellom snittscorene til deltagerne som praktiserte dette og ikke, der deltagerne som baserte seg på det første som slo dem som sannsynlig traff dårligere enn dem som ikke gjorde dette. Selv om deltagerne scoret generelt høyt på aktiv fordomsfri tenkning, var det bare et mindretall som praktiserte dette gjennom å samle informasjon fra flere kilder, men de som gjorde det traff også bedre. Det samme mønsteret finner vi i dette delkapittelets separate analyse av ekspertene i FFIs turnering, der de beste oppførte seg mer reveaktige, selv om de ikke nødvendigvis scoret mer reveaktige på de testene av generelle tenkemåter. Scorene på tester av generelle tenkemåter og reve- vs. pinnsvin-stereotypene kan derfor ikke brukes til å skille mellom gode og dårligere deltagere eller eksperter på forhånd. Det er hvorvidt de praktiserer disse tilnærmingene når de faktisk predikerer som betyr noe.

Deltagerne valgte også ulike tilnærminger til prediksjon ut fra hvor godt grunnlag de følte de hadde for å svare på spørsmålene. Deltagerne gjettet oftere på spørsmål som de ikke følte de kunne noe om, mens de tenkte seg grundigere om på spørsmål de følte de hadde eller burde ha peiling på. I tillegg tenkte et flertall av deltagerne taktisk når de fordelte sannsynlighetene sine, og halvveis ut i turneringen tenkte over halvparten av deltagerne mer taktisk enn før. Det er likevel ingen mønstre som viser at jo mer taktisk deltagerne tenkte, jo bedre traff de.

Fraværet av en sammenheng mellom treffsikkerheten og deltagernes vurderinger av hvor godt grunnlag de hadde for å svare kan også kobles til observasjonene fra GJP, der en stor andel av de beste deltagerne kom fra andre fagfelt enn samfunnsvitenskap, spesielt fra fysikk, biologi og programvareutvikling.[315] Disse fagområdene var overrepresenterte, selv om spørsmålene de fikk stort sett falt innenfor samfunnsvitenskapelige disipliner, spesielt statsvitenskap og samfunnsøkonomi. En mulig forklaring som ble trukket frem i GJP var at de beste deltagerne delte en høy interesse for problemløsningsoppgaver og spørsmålet om hvordan probabilitetsteori kan hjelpe oss med å bli bedre, og at denne interessen antagelig er mer utbredt innenfor naturvitenskapelig fagfelt enn samfunnsvitenskapene. I tillegg er det mulig at samfunnsvitenskapene er mindre tilgjengelige for prediksjon, fordi de handler om sosial fenomener og menneskelig adferd. Samfunnsvitere er antageligvis derfor også mindre vant til, og har mindre trening med, å predikere.

Det er imidlertid lite som støtter en slik forklaring basert på FFIs datasett. Av de 321 deltagerne vi kjenner fagfeltet til, oppgir flesteparten at de har høyere utdanning innenfor naturvitenskapelige og tekniske fag (47 %), nest flest innenfor samfunnsfag og juridiske fag (22 %) og tredje flest innenfor økonomiske og administrative fag (11 %).[316] Det er altså en overvekt av «realfagsfolk» i FFIs turnering. Hvis vi bare ser på de 50 beste av disse 321 deltagerne, er andelene fra de tre fagfeltene nesten like (hhv. 52 %, 24 % og 6 %). Det er altså ingen større andel realfagfolk enn samfunnsvitere blant de beste i FFIs turnering. Det er heller ingen statistisk signifikante forskjeller mellom snittscorene til deltagerne fra disse tre fagfeltene. Dette sammenfaller

[315] 'Edge Master Class 2015: A Short Course in Superforecasting', del 2.
[316] Basert på 321 deltagere som, i tillegg til å oppfylle turneringens minstekrav, svarte på undersøkelsen om hvordan de tenkte når de predikerte og spørsmålet om hvilket fagfelt de eventuelt hadde høyere utdanning innenfor.

for øvrig med fraværet av en signifikant forskjell mellom korrelasjonen med abstrakt resonneringsevne tidlig og sent i FFIs turnering, slik som i GJP, der det hevdes at en gradvis svakere, men fortsatt signifikant korrelasjon skyldes at deltagerne fikk mer trening i å predikere.

I stedet peker resultatene fra FFIs turnering på et sett med felles måter å tenke på som synes å være viktigere for treffsikkerheten. Tabell 5.15 lister opp alle de prediksjonsspesifikke tenkemåtene der det er en signifikant forskjell mellom deltagerne som brukte og ikke brukte dem ved signifikansnivå 0.05. Det viktigste er å unngå å basere seg på magefølelsen eller det første som faller en inn som det mest sannsynlige. De mest treffsikre tilnærmingene er å ta utgangspunkt i dagens situasjon, lete etter informasjon fra flere kilder, tenke over lignende historiske tilfeller med forskjellige utfall og bruke metoder som grunnfrekvens, referanseklasser og ekstrapolasjon til å fremskrive den videre utviklingen. Dette gir også grunn til å tro at deltagerne kunne ha truffet bedre om de brukte mer treffsikre metoder, uavhengig av hvor godt grunnlag de selv følte de hadde for å forutsi utviklingen.

Bruk	**Ikke bruk**
Lette etter informasjon fra flere forskjellige kilder. (aktiv fordomsfri tenkning) Baserte meg på hvordan utfallene har fordelt seg ved tilfeller av samme type hendelse. (referanseklasser) Baserte meg på tekstene og ev. figurer som fulgte med spørsmålet, og vurderte utfallene ut fra dette. (ankring) Estimerte hvor ofte hendelsen eller utfallet hadde skjedd før. (grunnfrekvens) Baserte meg på flere lignende, historiske tilfeller med forskjellige utfall. (bruk av flere historiske analogier) Tok utgangspunkt i dagens situasjon/ nivå, og justerte min prediksjon deretter. (ankring) Baserte meg på en fremskrivning av den samme utviklingen som frem til nå. (ekstrapolasjon)	Baserte meg på magefølelsen min. (intuisjon) Baserte meg på det første som slo meg som mest sannsynlig. (kognitiv lukking)

Tabell 5.15 Prediksjonsspesifikke tenkemåter som korrelerer med bedre treffsikkerhet. Alle er signifikant forskjellige enn bruk/ikke-bruk på 0.05-nivå i FFIs turnering.

Samtidig er det en rekke teknikker, som ofte benyttes i prediksjonssammenheng, der det *ikke* er en signifikant forskjell mellom scorene til deltagerne som brukte dem og ikke. Disse inkluderer induktiv og deduktiv resonnering (som var et viktig skille mellom reve- og pinnsvinekspertene), å bruke snittet av flere forskjellige estimater (også kjent som *wisdom of the crowd*), å basere seg på ett lignende, historisk tilfelle (i motsetning til flere forskjellige), basere seg på den aller siste utviklingen som hadde skjedd da spørsmålet ble stilt (som kan være et mål på i hvor stor grad de var påvirket av nyhetsdekningen), *post-mortem*-analyser (reflekterte rundt hvorfor de hadde

bommet eller truffet sist) og å forsøke å ta hensyn til sorte svaner (uforutsigbare, overraskende hendelser). At opplæring i de to siste teknikkene heller ikke korrelerte med treffsikkerheten i GJP, gjør disse funnene enda mer robuste.

Når det gjelder deltagernes oppgavespesifikke evner og innsats i turneringen er det mer usikkert hvor overførbare sammenhengene er fra én turnering til en annen. I FFIs turnering korrelerer antallet unike sannsynlighetsestimater med bedre prediksjonsevne, mens i GJP korrelerer variabelen med både bedre og dårligere treffsikkerhet. Det er også usikkert i hvor stor grad korrelasjonen med deltagernes forståelse av scoringssystemet kan generaliseres fra FFIs turnering, siden signifikansnivået varierer med deltagerutvalg og korrelasjonstest og siden denne variabelen aldri ble målt i GJP.

Det er imidlertid verdt å merke seg at flere uavhengige variabler korrelerer med hverandre, men ikke med treffsikkerheten. Det er for eksempel en signifikant korrelasjon mellom deltagernes forståelse av scoringssystemet og hvor mange unike sannsynlighetsestimater de brukte. Deltagerne med bedre forståelse brukte relativt mange sannsynlighetsestimater, antageligvis fordi de tenkte at dette ville gi dem bedre scores. Det var likevel en sterkere korrelasjon mellom treffsikkerheten og antall sannsynlighetsestimater enn med scoren deres på Brier-score-forståelse. Tre av de disposisjonelle variablene (abstrakt resonneringsevne, aktiv fordomsfri tenkning og kognitiv motivasjon) korrelerer også relativt sterkt med flere andre evner som hver for seg varierer med treffsikkerheten, men som ikke eller i liten grad korrelerer med treffsikkerheten selv. Det er altså ikke alle assosierte egenskaper som ser ut til å henge sammen med bedre prediksjonsevne.

Når det kommer til betydningen av deltagernes innsats er det ett funn som kan generaliseres til andre turneringer og ett som ikke kan det. Selv om flere spørsmål besvart korrelerer med bedre treffsikkerhet i FFIs turnering, forsvinner denne samvariasjonen om vi bare ser på de mest aktive deltagerne som svarte på et likere antall spørsmål. I tillegg er korrelasjonen motsatt i GJP, der et høyere antall spørsmål besvart betyr dårligere treffsikkerhet. Tiden deltagerne brukte på spørsmålene er derimot en av tydeligste korrelasjonene i både FFIs og GJPs turnering: Jo mer tid deltagerne brukte per spørsmål, jo bedre var treffsikkerheten. Dette var den eneste variabelen som var signifikant på 0.0001-nivå ved begge deltagerutvalg og tester brukt i FFIs turnering, og korrelasjonen er den eller én av de sterkeste i begge turneringer. At korrelasjonen er så sterk til tross for at FFIs deltagere brukte under halvparten så mange minutter (1,4) som GJPs (3,6), tilsier at den ikke handler om absolutt tidsbruk, men relativ tidsbruk i forhold til andre deltagere. Dette funnet vil derfor trolig også gjelde i andre turneringer. Samtidig viste spørreundersøkelsene fra FFIs turnering at det også er viktig hva denne tiden brukes til.

Oppsummert viser den statistiske analysen at det finnes en rekke disposisjonelle variabler som kan brukes til å «screene» deltagere for å identifisere dem som har best utgangspunkt for å treffe bedre enn andre. Testene av kognitive evner og kunnskapsnivå fremstår som særlig relevante. Testene av generelle tenkemåter, som aktiv fordomsfri tenkning og kognitiv lukking, har lite å si hvis ikke deltagerne faktisk praktiserer dem. Tester av deltagernes tenkemåter kan derimot brukes til å filtrere bort deltagere som *ikke* har forutsetningene for å tenke på de riktige måtene.

Innsatsvariablene, som antall unike sannsynlighetsestimater og tid brukt per spørsmål, er umulige å måle på forhånd. Betydningen av disse sammenfaller likevel med de kvalitative undersøkelsene, der fellesnevneren blant deltagerne som traff best var at de gjør grundigere vurderinger.

Når det er sagt, er det ikke sikkert at de relative forskjellene som henger sammen med ulike individuelle egenskaper har så mye å si for treffsikkerheten i praksis. På den ene siden er den gjennomsnittlige treffprosenten til deltagerne i FFIs turnering bare 51 %, mens Brier-scoren tilsvarer den samme som det en ville fått ved lik fordeling av sannsynlighetene på alle svaralternativer, slik som apen. På den annen side treffer de 100 beste deltagerne i FFIs turnering omtrent like godt som de beste, sammenlignbare deltagerne i GJP ved like forhold. Den gjennomsnittlige treffprosenten til de beste deltagerne i FFIs turnering er 57 %. Treffprosenten er 75 % på de binære spørsmålene, mens Brier-scoren deres er litt bedre enn tilfeldig. Selv om dette ikke høres særlig imponerende ut, er treffsikkerheten til de 100 beste deltagerne i FFIs turnering, som bestod av både fagfolk og amatører, litt bedre enn snittscorene til den beste av ekspertgruppene analysert i delkapittel 5.2. Dette tilsier at det finnes et potensial for å rekruttere personer som treffer systematisk bedre enn andre (inkludert eksperter) blant deltagerne i FFIs turnering.

Problemet er at selv de 100 beste deltagerne i FFIs turnering er altfor selvsikre. Med en overkonfidens på 14 % er de langt mer overkonfidente enn selv de dårligste deltagerne i GJP. FFIs turnering deltagere oppgav generelt en altfor høy sannsynlighet for svarene de trodde var riktig sammenlignet med hvor ofte de traff. Denne tendensen til å hevde at en hendelse vil skje, men som ikke gjør det, er spesielt farlig i forsvars- og sikkerhetspolitiske spørsmål, og begrenser derfor den potensielle verdien i å samle og aggregere prediksjoner fra FFIs deltagere.

Selv om det er mulig å rekruttere individer som er systematisk bedre enn andre til å predikere og gi dem metoder som er mer treffsikre enn andre, er det altså ikke gitt at dette er *godt nok*. Det viktigste funnet fra GJP var imidlertid identifiseringen av superforecastere, som var veldig gode til å predikere, uansett treffsikkerhetsmål. Dette reiser spørsmålet om det finnes en tilsvarende god gruppe deltagere i FFIs turnering som også traff godt nok til å utgjøre en forskjell i praksis.

5.4 Norske superforecastere

Det fjerde spørsmålet som analyseres her er: *Hva kjennetegner individene som er best til å predikere forsvars- og sikkerhetspolitikk?*

For å besvare dette spørsmålet identifiseres de aller beste deltagerne – «superforecasterne» – i FFIs turnering. Her sammenlignes disse med de nest beste og resten av deltagerne i FFIs turnering og med superforecasterne i GJP. Det viktigste funnet fra GJP var at superforecasterne hadde en distinkt profil som skilte seg fra resten. Det samme undersøkes også her, basert på de samme variablene som i delkapittel 5.3 ble brukt til å måle systematiske individuelle forskjeller.

I FFIs turnering er det ikke mulig å bruke det samme kriteriet for identifisering av superforecastere som i GJP. I GJP ble deltagerne først rangert ut fra treffsikkerheten etter det første året. Deretter ble de 5 beste fra totalt 12 eksperimentgrupper plukket ut som superforecastere.[317] Disse ble fordelt på 5 nye superforecasterlag med 12 medlemmer hver, som til sammen utgjorde rundt 60 deltagere. Disse utgjorde en egen eksperimentgruppe det andre året. Etter at det andre året var avsluttet ble det identifisert 60 nye superforecastere, slik at denne eksperimentgruppen utgjorde rundt 120 superforecastere etter at det tredje året var over.

I denne rapportens reanalyse består GJPs superforecastere av alle deltagerne som er både markert som dette i prosjektets fullstendige datasettet og inkludert i GJP350s replikasjonsdatasett. Totalt utgjør dette 127 deltagere. I GJPs publikasjoner er superforecasterne omtalt som «de 2 % beste» i turneringen, men denne andelen må være basert på alle som besvarte minst ett spørsmål i løpet av de tre første årene, som i GJP350s datasett inkluderer 3365 deltagere. Av de 1751 deltagerne som oppfylte minstekravet i GJP350 utgjør de 127 superforecasterne 7 %.

I FFIs turnering består superforecasterne av de 60 beste deltagerne totalt sett. FFIs turnering ble ikke gjennomført med ett år av gangen, men over tre sammenhengende år, der mange av spørsmålene ikke ble avgjort i løpet av det inneværende året og deltagerne ikke ble inndelt i eksperimentgrupper. Det er derfor ikke hensiktsmessig å identifisere nye grupper av superforecastere underveis. De 60 beste deltagerne i FFIs turnering utgjør også 7 % av de 833 deltagerne som oppfyller denne studiens minstekravet, altså samme andel av deltagermassen som i GJP.

Som i de øvrige kapitlene om individuelle forskjeller er treffsikkerheten og rangeringen basert på deltagernes gjennomsnittlige, standardiserte Brier-scores på alle spørsmålene de svarte på.[318]

[317] Disse bestod av ulike kombinasjoner av deltagere som predikerte alene, i prediksjonsmarkeder, i grupper, fikk vite hva andre predikerte og fikk trening i scenariotenkning, probabilistisk tenkning eller ingen av delene. Det var tre grupper med deltagere som predikerte helt alene (uten noe trening, med probabilistisk trening eller scenariotrening), tre grupper som kunne se hva andre predikerte (uten trening, med probabilistisk trening eller scenariotrening), to grupper i prediksjonsmarkeder (med og uten trening), to grupper som var arbeidet i grupper (med og uten scenariotrening) og én gruppe med superforecastere som alltid arbeidet i grupper og fikk trening. I tidligere delkapitlers analyser av GJPs eksperimentgrupper er ulike typer trening slått sammen og prediksjonsmarkeder utelatt.

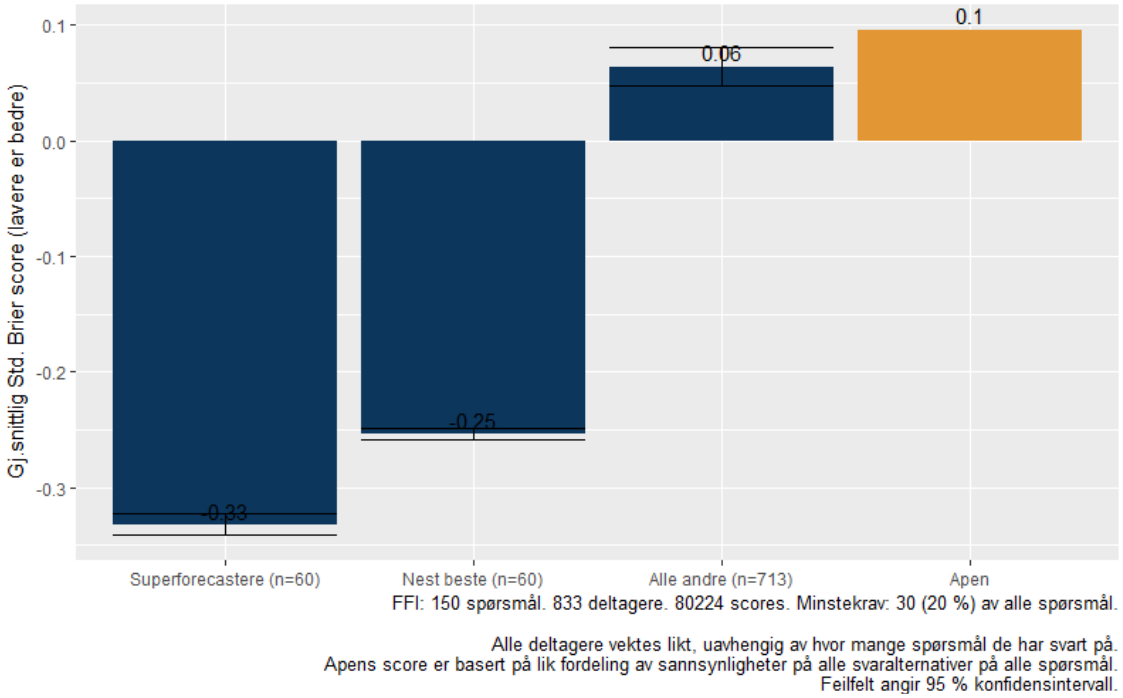
[318] For histogrammer som viser fordelingen av hver gruppes Brier-scores og verdier på alle uavhengige variabler, se kapittel 5 i Beadle (2021), 'Tilleggsdokumentasjon til foreløpige resultater fra FFIs prediksjonsturnering'.

5.4.1 Gjennomsnittlig treffsikkerhet

Det første spørsmålene er om det i det hele tatt finnes en gruppe «superforecastere» i FFIs turnering. Her sammenlignes derfor de de samme tre gruppene som i GJPs superforecaster-studie:[319]

1) De 60 beste deltagerne, som utgjør «superforecasterne».
2) De 60 nest beste deltagerne.
3) Alle resterende deltagere, som utgjør 713 personer.

Figur 5.29 viser de standardiserte Brier-scorene til de 60 beste, 60 nest beste og resten av deltagerne i FFIs turnering. Som i GJP er treffsikkerheten til den første gruppen signifikant bedre enn begge kontrollgruppene.[320] Faktisk er den standardiserte Brier-scoren deres (–0,36) nesten helt lik superforecasterne i GJPs (–0,37). Det betyr at FFIs superforecastere treffer relativt sett like mye bedre enn de andre deltagerne som superforecasterne gjorde i GJPs.[321] Scoren deres er også mye bedre enn forsvarsforskernes (–0,1), som var den relativt sett mest treffsikre av alle gruppene undersøkt i delkapittelet om eksperter (se delkapittel 5.2). Dette understreker hvor mye mer hensiktsmessig det synes å være å basere seg på enkeltpersoner som har truffet godt, uansett bakgrunn, i stedet for å anta at eksperter treffer best, basert på formell kompetanse og erfaring.



Figur 5.29 Relativ treffsikkerhet, basert de 60 beste, 60 nest beste og alle andre deltagere.

[319] Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters', ss. 269–270.
[320] FFIs superforecastere vs. hhv. nest beste: $t(86) = -15.39$, $p < 0.0001$; resten: $t(582) = -41.38$, $p < 0.0001$. Fordelingen av de tre gruppenes scores er ikke overraskende skjeve. Derfor kan Wilcoxon-testen være bedre egnet, men den gir samme funn. Verdiene til alle andre er imidlertid ikke skjeve til samme side som de to beste gruppene.
[321] De nest beste og øvrige deltagerne sammenlignes ikke på tvers av turneringene, siden de nest beste i GJPs studie ble kom fra deltagere i grupper, mens de nest beste i FFIs turnering var de nest beste totalt sett.

Men hvor gode er de norske superforecasterne i praksis? Tabell 5.16 sammenligner den objektive treffsikkerheten til superforecasterne, de nest beste og resten av deltagerne i FFIs turnering.

Deltagergruppe (antall deltagere)	**Brier-score**	**Treffprosent**	**Kalibrering**
Superforecastere (60)	0,36	61,8 %	5,4 %
Nest beste (60)	0,40	57,5 %	8,3 %
Alle andre (713)	0,54	49,3 %	23,6 %

Tabell 5.16 Objektiv treffsikkerhet, basert på de beste, nest beste og alle andre deltagere.

For det første er superforecasternes Brier-score rundt 30 % bedre enn snittet til deltagerne i FFIs turnering som helhet (0,52) og ved tilfeldig gjetning (0,51) på de samme spørsmålene. Denne scoren tilsvarer en prediksjon på 58 % på det riktige av to mulige utfall mot 49 % i turneringen som helhet. Treffprosenten deres er også 20 % høyere enn turneringssnittet (51 %) og nesten dobbelt så høy som apens (33 %). Den største forskjellen er kalibreringen, der superforecasterne har en betydelig lavere grad av overkonfidens (5 %) enn i turneringen generelt (21 %).

For det andre scorer superforecasterne gjennomgående og signifikant bedre enn de nest beste og alle andre deltagere i turneringen på alle de objektive målene av treffsikkerhet.[322] Det er faktisk litt større forskjell mellom scorene til superforecasterne og de nest beste deltagerne enn mellom deltagerne med og uten forsvars- og sikkerhetspolitisk erfaring, som var én av få variabler hvor det var en signifikant forskjell i treffsikkerheten basert på ekspertisekriterier.[323] Samtidig er det bare snakk om noen få prosentpoeng forskjell i treffprosent og kalibrering, og både superforecasterne og de nest beste er bedre enn alle ekspertgruppene, inkludert forsvarsforskerne.[324]

For det tredje er treffsikkerheten til de til sammen 120 deltagerne bestående av superforecastere og de nest beste deltagerne i FFIs turnering også bedre enn scorene til de «100 beste deltagerne» analysert tidligere (se underkapittel 5.3.2).[325] Forklaringen er at scorene til de 100 beste deltagerne baserte seg på de resterende 125 spørsmålene som ble avgjort etter de 25 første som ble brukt til å identifisere dem i første omgang. Hensikten var da å undersøke om treffsikkerheten deres var stabil over tid og konklusjonen var at det er mulig å identifisere deltagere som kommer til å treffe systematisk bedre enn andre, basert på bare 25 spørsmål. I dette delkapittelet er de beste deltagerne derimot identifisert retrospektivt, basert på alle 150 avgjorte spørsmål. Det at treffsikkerheten til de 120 beste deltagerne etter 150 spørsmål er høyere enn scorene til de 100 beste deltagerne etter 25 spørsmål er kanskje ikke overraskende, men tilsier samtidig at det kan være nødvendig å basere seg på flere spørsmål for å finne deltagerne som treffer aller best.

[322] FFIs superforecastere vs. hhv. nest beste: Brier-score: $t(117) = -5.98$, $p < 0.0001$, treffprosent: $t(117) = 5.35$, $p < 0.0001$, kalibrering: $t(116) = -2.65$, $p < 0.01$; alle andre: Brier-score: $t(165) = -29.61$, $p < 0.0001$, treffprosent: $t(89) = 20.60$, $p < 0.0001$, kalibrering: $t(107) = -21.91$, $p < 0.0001$.
[323] Erfaring vs. ingen erfaring: Brier-score: 0,50 og 0,53. Treffprosent: 52 % og 50 %. Kalibrering: 20 % og 22 %.
[324] Forsvarsforskere: Brier-score: 0,46. Treffprosent: 54 %. Kalibrering: 18 %.
[325] Beste 100 deltagere: Brier-score: 0,46. Treffprosent: 57 %. Kalibrering: 14 %.

Det mest interessante funnet er likevel hvor godt kalibrerte FFIs superforecastere er. Det finnes få studier som sier noe om hvor liten over- eller underkonfidensen må være for at den skal være «god» i prediksjonssammenheng, men overkonfidensen på bare 2 % i GJP ble ansett som svært liten.[326] Det er heller ikke mulig å komme så mye nærmere perfekt kalibrering enn dette. Til sammenligning var den gjennomsnittlige overkonfidens til ekspertene i EPJ 12 %, mens overkonfidensen til forsvarsforskerne i FFIs turnering hele 18 %. Med en overkonfidens på bare 5 % må FFIs superforecastere derfor også anses som godt kalibrert, både objektivt og relativt sett.

Tabell 5.17 viser den objektive treffsikkerheten til FFIs (blå) og GJPs superforecastere (grønn). Her skilles det mellom fire ulike definisjoner av superforecastere i GJPs turnering for å sammenligne med FFIs på egne og likest mulige grunnlag. Innenfor hvert treffsikkerhetsmål skilles det også mellom binære, kategoriske og ordinale spørsmål, siden typen spørsmål har vist seg å ha betydning for sammenligninger av de to turneringene.

Deltagerutvalg	Brier-score	Treffprosent	Kalibrering
FFI – superforecastere (60) – bare første uken – predikerte uten hjelp	0,36 - Binære: 0,26 - Kategoriske: 0,62 - Ordinale: 0,31	61,8 % - Binære: 82,8 % - Kategoriske: 54,2 % - Ordinale: 51,7 %	5,4 % - Binære: –0,5 % - Kategoriske: 8,0 % - Ordinale: 8,2 %
GJP – superforecastere (127)	0,18 - Binære: 0,18 - Kategoriske: 0,27 - Ordinale: 0,15	86,3 % - Binære: 88,3 % - Kategoriske: 84,0 % - Ordinale: 77,7 %	–2,4 % - Binære: –1,8 % - Kategoriske: –5,4 % - Ordinale: –0,8 %
GJP – superforecastere (127) – bare første uken	0,28 - Binære: 0,28 - Kategoriske: 0,30 - Ordinale: 0,25	78,4 % - Binære: 80,8 % - Kategoriske: 80,4 % - Ordinale: 65,7 %	–1,4 % - Binære: –0,6 % - Kategoriske: –9,6 % - Ordinale: –0,6 %
GJP – superforecastere (18) – predikerte uten hjelp	0,18 - Binære: 0,17 - Kategoriske: 0,33 - Ordinale: 0,13	87,1 % - Binære: 88,9 % - Kategoriske: 77,7 % - Ordinale: 79,7 %	0,3 % - Binære: –0,1 % - Kategoriske: 2,3 % - Ordinale: 1,8 %
GJP – superforecastere (17) – bare første uken – predikerte uten hjelp	0,29 - Binære: 0,29 - Kategoriske: 0,44 - Ordinale: 0,15	78,8 % - Binære: 80,9 % - Kategoriske: 65,4 % - Ordinale: 71,8 %	2,2 % - Binære: 2,9 % - Kategoriske: 2,8 % - Ordinale: –0,9 %

Tabell 5.17 Objektiv treffsikkerhet, basert på superforecastere i FFIs og GJPs turneringer, gitt ulike prediksjonstidspunkter, eksperimentgrupper og spørsmålstyper.

[326] Moore mfl. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition'.

Ved første øyekast fremstår superforecasterne i FFIs turnering som langt dårligere enn i GJPs. Mens Brier-scoren til FFIs superforecastere er 0,36 (første rad), er GJPs bare 0,18 når den er basert på alle prediksjoner (andre rad) og 0,28 når den er basert kun på prediksjoner registrert i løpet av den første uken (tredje rad). Som tidligere halveres gapet mellom FFIs og GJPs turneringer når de vurderes ut fra likt prediksjonstidspunkt, men det er fortsatt en betydelig forskjell. Det samme mønstret sees ved treffprosenten, der GJPs superforecasteres snittscore ligger gjennomgående høyere enn FFIs.

Forskjellene mellom FFIs og GJPs superforecastere er imidlertid ikke like på alle spørsmålstyper. På binære spørsmål, som er den mest direkte sammenlignbare typen på tvers av turneringene, er både Brier-scoren og treffprosenten til FFIs superforecastere signifikant bedre enn GJPs ved samme prediksjonstidspunkt.[327] Forskjellene er imidlertid små i praksis. Begge turneringenes superforecasteres Brier-score tilsvarer en prediksjon på rundt 63–64 % på riktig svar og 37–36 % på galt svar på et ja/nei-spørsmål, og en treffprosent som tilsvarer en evne til å velge det riktige utfallet fire av fem ganger i snitt. Som ellers i turneringen er likevel FFIs superforecastere dårligere enn GJPs på ordinale spørsmål og veldig mye dårligere på kategoriske, uansett mål. Det er uklart hva som er årsaken til dette mønstret, men det skyldes ikke forskjeller i antallene svaralternativer per spørsmål, siden de var omtrent like i begge turneringer.

I denne sammenheng er det verdt å merke seg at forskjellene i GJP mellom Brier-scorene på binære og kategoriske spørsmål blir mindre når de baseres på prediksjoner fra den første uken enn når de baseres på alle, mens gapet mellom treffprosentene holder seg svært stabilt. Dette betyr at superforecasterne ikke fant det like enkelt å oppgi høyere sannsynligheter til riktig svar på spørsmål med to mulige utfall enn på spørsmål med flere i starten av et spørsmål, som når de kunne oppdatere prediksjonene sine underveis. Det er derfor grunn til å anta at forskjellen mellom FFIs superforecasteres treffsikkerhet på binære og kategoriske eller ordinale spørsmål ville blitt mindre om de også hadde kunnet predikere gjennom hele spørsmålsperioden.

Vi ser også at prediksjonstidspunktet, som gjennom hele denne rapporten har hatt stor betydning for variasjoner i GJPs deltageres Brier-scores og treffprosenter, ikke leder til de samme forskjellene i superforecasternes kalibrering. Selv om treffprosenten til GJPs superforecastere faller med åtte prosentpoeng når den baseres på prediksjoner fra bare den første uken, er det bare snakk om ett prosentpoengs endring i kalibreringen deres. Det gir dermed ingen grunn til å anta at FFIs superforecastere ville blitt særlig bedre kalibrerte om de også hadde fått muligheten til å oppdatere sine prediksjoner helt frem til spørsmålene ble avgjort.

GJPs egendefinerte superforecastere er imidlertid ikke direkte sammenlignbare med FFIs, selv ved likt prediksjonstidspunkt. Treffsikkerheten deres er nemlig basert på snittet over alle tre årene – altså også *etter* at de hadde blitt trukket ut som superforecastere, fått opplæring og blitt satt i grupper med andre, som hver for seg var tiltak som eksperimentene viste bidro til å øke

[327] Forskjellene mellom FFIs og GJPs superforecasteres scores på binære spørsmål, gitt likt prediksjonstidspunkt, er også signifikante på 0.05-nivå: Brier-score: $t(178) = -2.21$, $p < 0.05$. Treffprosent: $t(150) = 2.03$, $p < 0.05$.

prediksjonsevnen. Siden GJPs superforecastere bestod av de 5 beste deltagerne fra 12 ulike eksperimentgrupper, betyr dette i tillegg at bare 5 av 60 superforecastere ble rekruttert fra gruppen av deltagere som predikerte uten noen hjelpemidler, slik alle FFIs superforecastere gjorde.

For å sammenligne superforecastere på tvers av turneringene må vi kun ta utgangspunkt i deltagerne i GJP som predikerte alene hele tiden. Av de 318 deltagerne som hverken fikk opplæring eller ble satt på grupper med andre var det 259 av dem som forble i denne eksperimentgruppen gjennom hele turneringen.[328] De øvrige 59 deltagerne ble senere overført til andre eksperimentgrupper og gjør dem grunnleggende forskjellige fra FFIs. Av de 259 som alltid predikerte alene kan vi plukke ut 18 superforecastere som representerer de beste 7 %. Dette er den samme andelen deltagere som de 60 superforecasterne i FFIs turnering utgjør av alle 833 deltagerne. Disse nye superforecasterne fra GJP er også relativt sett like mye bedre enn resten av deltagerne de sammenlignes med, som det FFIs superforecastere er i sin turnering.[329]

Den gjennomsnittlige Brier-scoren til disse, direkte sammenlignbare superforecasterne fra GJP er 0,18 når den baseres på alle prediksjoner (fjerde rad) og 0,29 når den kun baseres på prediksjoner fra den første uken etter at spørsmålene ble publisert (femte rad). Som over er treffsikkerheten deres også høyere enn FFIs, men gapet halveres ved likt prediksjonstidspunkt. Igjen er også Brier-scoren og treffprosenten til FFIs superforecastere litt bedre enn GJPs på binære spørsmål ved likt prediksjonstidspunkt, mens FFIs fortsatt treffer dårligere på kategoriske og ordinale. Det lave antallet av denne typen superforecastere i GJPs datasett og fordelingen av scorene deres gjør imidlertid at forutsetningene for å kunne teste hvorvidt forskjellene mellom snittscorene deres og FFIs superforecasteres er statistisk signifikante, ikke er oppfylte.

Oppsummert finnes det utvilsomt en gruppe deltagere i FFIs turnering som er signifikant bedre enn resten og som i tillegg kan måle seg med GJPs superforecastere. Under likest mulige forhold treffer FFIs superforecastere minst like godt som GJPs, inkludert dem som fikk opplæring og samarbeidet med andre. Denne gruppen superforecastere fra GJP er trukket frem som bevis for at det er mulig å forutsi internasjonal politikk svært presist og de traff bedre enn amerikanske etterretningsanalytikere med tilgang på informasjon. Det er da interessant å observere at treffsikkerheten til de «ensomme» superforecasterne i GJP, fra eksperimentgruppen som predikerte alene hele tiden, er omtrent helt lik scorene til GJPs «vanlige» superforecastere ved begge prediksjonstidspunkter. Denne rapportens analyse viser med andre ord at det er tilnærmet ingen praktisk forskjell mellom treffsikkerheten til de aller beste deltagerne som ikke fikk *noe* hjelp i GJP og de aller beste som fikk *alle* mulige forbedringstiltak i samme turnering.

[328] Brier-scorene og treffprosentene til disse 259 deltagerne er hhv. 0,38 og 70 % (0,37 og 71 % ved alle 318 deltagere) basert på alle prediksjoner og 0,45 og 64 % (0,43 og 65 % ved alle 318 deltagere) basert på første uken.
[329] Den standardiserte Brier-score til disse superforecasterne i GJP er i snitt –0,33 mot –0,36 i FFIs turnering.

5.4.2 Individuelle egenskaper

Av alle tiltakene som ble gjort i GJP var det identifisering av superforecastere som fremholdes som det mest effektive. Denne rapportens analyser underbygger dette ved å vise at det i FFIs turnering fantes en gruppe deltagere som var like gode som GJPs beste, så lenge de predikerte på samme måte. Dette gir grunn til optimisme rundt mulighetene for å treffe bedre på spørsmål om internasjonal politikk, så lenge vi klarer å rekruttere de «riktige folkene».

For å kunne finne potensielle superforecastere på forhånd må vi imidlertid vite hva vi skal lete etter. Her undersøkes det derfor om superforecasterne i FFIs turnering skiller seg fra resten av deltagerne gjennom bestemte karaktertrekk, slik superforecasterne gjorde i GJP. Ved å sammenligne superforecasterne på tvers av turneringen kan vi også måle om FFIs «ensomme», som treffer godt uten forbedringstiltak underveis, skiller seg fra GJPs «fremdyrkede».

5.4.2.1 Disposisjonelle variabler og innsats

Tabell 5.18 viser scorene til superforecasterne, de nest beste og alle andre deltagere i FFIs turnering på alle de samme disposisjonelle variablene og innsatsvariablene som ble undersøkt i delkapittel 5.3, fordi eksisterende forskning tilsier at de kan henge sammen med treffsikkerhet.

Fet skrift på verdiene til de nest beste eller alle andre deltagerne indikerer at det er en statistisk signifikant forskjell mellom snittscorene deres og superforecasternes på 0.01-nivå. Dette er det samme signifikansnivået som ble brukt i en tilsvarende sammenligning i GJPs superforecaster-artikkel.[330] Siden fordelingen av verdiene på disse variablene er relativt skjeve og langt fra normalfordelte er Wilcoxon signed-rank-testen benyttet i stedet for t-testen, men resultatene er de samme.[331] Vedlegg B gir en deskriptiv analyse av scorene til hver av de tre deltagergruppene.

Helt til høyre oppgis scorene til superforecasterne i GJP350, basert på alle variabler som finnes i det tilgjengelige datasettet og som er sammenlignbare med FFIs turnering.[332] Også her betyr fet skrift at scorene til GJPs superforecastere var signifikant forskjellige fra FFIs på 0.01-nivå.

[330] Mellers mfl. (2015), 'Identifying and Cultivating Superforecasters', s. 274.
[331] Den eneste variabelen der testene gir ulike svar er ved sammenligningen av FFIs og GJPs superforecastere på tallforståelse. Ved Wilcoxon-testen er forskjellen signifikant på 0.01-nivå, mens t-testen bare er signifikant på 0.05-nivå.
[332] Verdiene oppgitt i GJP350-artikkelen gjengis ikke her, fordi de er tilnærmet identiske med reanalysens resultater.

		FFI			GJP
		Super-forecastere	**Nest beste deltagere**	**Alle andre deltagere**	**Super-forecaster**
Kognitive evner	Shipley-2 Block Patterns (0–26)	19,19	19,10	17,30	-
	CRT original (0–3)	2,84	2,65	**2,37**	**2,45**
	CRT utvidet (0–18)	16,42	16,03	**14,66**	16,62
	Tallforståelse (0–4)	3,37	3,24	**2,59**	3,69
Kunn-skap	Politisk kunnskapsnivå (0–50)	38,27	37,24	**34,88**	38,43 (2. året)
Tenkemåter	Aktiv fordomsfri tenkning (1–7)	6,32	6,12	6,13	**6,02**
	Kognitiv lukking (1–7)	3,77	3,89	3,89	-
	Rev vs. pinnsvin – enkeltpåstand (1–7)	2,63	2,56	2,83	-
	Motivasjon – være blant de beste (1–7)	5,78	5,26	**4,86**	5,58
	Kognitiv motivasjon (1–7)	5,43	5,35	**5,14**	**5,96**
Oppgavespes. ferdigheter	Antall unike sannsynlighetsestimater (bare første uken)	51	40	**31**	47
	Brier-score forståelse (0–5)	1,91	**0,74**	**0,79**	-
Innsats	Andel (og antall) spørsmål besvart	69 % (167)	67 % (160)	**60 % (143)**	61 % (210)
	Tid brukt per spørsmål (antall minutter)	1,83	1,65	**1,32**	-

Tabell 5.18 Individuelle variasjoner i FFIs og GJPs turneringer. Fet skrift indikerer signifikant forskjell fra superforecasterne på 0.01-nivå.

For det første viser tabell 5.18 at det bare er én av de disposisjonelle og innsatsrelaterte variablene hvor det er en signifikant forskjell mellom superforecasterne og de nest beste deltagerne i FFIs turnering. Dette gjelder kjennskapen til hvordan treffsikkerheten i turneringen ble beregnet, der superforecasterne hadde en langt bedre forståelse av scoringssystemet enn de nest beste deltagerne (med snittscores på hhv. 1,9 vs. 0,7 av 5 riktige). Mens underkapittel 5.3.3 viste at betydningen av denne variabelen var mindre sikker enn de andre når det kom til variasjoner i individuell treffsikkerhet generelt, viser resultatene herfra at Brier-score-forståelsen er den eneste som kan skille superforecasteres individuelle egenskaper fra de nest beste deltagernes. Det kan derfor tenkes at det er nettopp superforecasternes bedre forståelse av scoringssystemet

som gav dem «det lille ekstra» som bringer dem helt til topps. Siden denne variabelen bare ble kartlagt i FFIs turnering, er det imidlertid ikke mulig å etterprøve funnet basert på GJPs datasett. Funnet tilsier likevel at GJP manglet en potensielt viktig variabel i sine analyser av superforecasterne.

For det andre viser tabell 5.18 at FFIs deltagere, i likhet med GJPs, skiller seg signifikant fra alle resterende deltagere på de fleste variabler. Unntakene er abstrakt resonneringsevne, aktiv fordomsfri tenkning, kognitiv lukking og skillet mellom reve- og pinnsvintenkning. Dette er i tråd med funnene fra underkapittel 5.3.3, der tre av disse variablene ikke korrelerte med forskjeller i individuell treffsikkerhet i det hele tatt, mens sammenhengen med aktiv fordomsfri tenkning var mindre robust enn andre variabler i både FFIs og GJPs turneringer. Dette understreker hvordan generelle tester av tenkemåter heller ikke fremstår som spesielt nyttige for å identifisere de aller beste deltagerne. Superforecasterne bruker også mer tid per spørsmål, som var en av de sterkeste sammenhengene med treffsikkerheten i korrelasjonsanalysen, men det er bare snakk om noen sekunders forskjell i snitt fra de nest beste og et halvt minutt fra resten.

For det tredje viser tabell 5.18 at FFIs superforecastere er svært like GJPs på nesten alle variabler. Scorene er tilnærmet identiske på kognitiv kontroll, tallforståelse og politisk kunnskapsnivå. Superforecasterne i GJP hadde litt, men signifikant, større kognitiv motivasjon, altså vilje til å engasjere seg i oppgaver som krever dypere tenkning, mens FFIs bruker flere forskjellige sannsynlighetsestimater og har besvart en høyere andel av spørsmålene. Forskjellene er imidlertid marginale. Det er også en signifikant forskjell mellom superforecasterne på den korteste CRT-testen, men scorene herfra er ikke pålitelige da testens internkonsistens lå under tilfredsstillende nivå i begge turneringer. Den siste signifikante forskjellen gjelder scorene på aktiv fordomsfri tenkning, der korrelasjonen med treffsikkerhet er mindre sikker enn ved andre variabler i begge tilfeller. I sum deler superforecasterne i FFIs turnering altså de samme kjennetegnene som GJPs.

5.4.2.2 Prediksjonsspesifikke tenkemåter

Selv om generelle tenkemåter som i tidligere forskning er forbundet med bedre treffsikkerhet ikke synes å være avgjørende for de individuelle forskjellene mellom deltagerne i FFIs turnering, viste underkapittel 5.3.3 at de deltagerne som faktisk anvendte metodene assosiert med disse tenkemåtene når de skulle predikere, traff bedre enn dem som ikke brukte dem.

For å undersøke om dette også gjelder for de aller beste deltagerne viser tabell 5.19 hvor stor andel av superforecasterne som brukte hver av de 17 prediksjonsspesifikke tenkemåtene sammenlignet med andelene av de nest beste og alle andre deltagerne. Antall deltagere som krysset av for hver tenkemåte er oppgitt i parentes. De fem tenkemåtene der superforecastere skiller seg mest fra kontrollgruppene er uthevet med fet skrift. I snitt er de største forskjellene mellom andelen superforecastere og kontrollgruppene på disse fem tenkemåtene rundt 25 prosentpoeng.

Prediksjonsspesifikke tenkemåter	Superfore-castere (48)	Nest beste (31)	Alle andre (269)
Baserte meg på mageføelsen min. (intuisjon)	**43,8 % (21)**	67,7 % (21)	71 % (191)
Tok utgangspunkt i en teori eller generell oppfatning jeg hadde av fenomenet fra før, og brukte denne til å vurdere hva som ville skje i dette tilfellet. (deduktiv resonnering)	47,9 % (23)	48,4 % (15)	46,8 % (126)
Tok utgangspunkt i det aktuelle spørsmålet, og tenkte gjennom hva ulike teorier ville sagt om hva som ville skje. (induktiv resonnering)	18,8 % (9)	9,7 % (3)	23 % (62)
Lette etter informasjon fra flere forskjellige kilder. (aktiv fordomsfri tenkning)	**31,2 % (15)**	6,5 % (2)	10,4 % (28)
Baserte meg på det første som slo meg som mest sannsynlig. (kognitiv lukking)	27,1 % (13)	29 % (9)	41,3 % (111)
Baserte meg på hvordan utfallene har for-delt seg ved tilfeller av samme type hen-delse. (referanseklasser)	**54,2 % (26)**	48,4 % (15)	37,2 % (100)
Baserte meg på tekstene og ev. figurer som fulgte med spørsmålet, og vurderte utfallene ut fra dette. (ankring)	79,2 % (38)	74,2 % (23)	62,5 % (168)
Estimerte hvor ofte hendelsen eller utfallet hadde skjedd før. (grunnfrekvens)	**68,8 % (33)**	45,2 % (14)	35,3 % (95)
Baserte meg på snittet av flere, forskjel-lige estimater av utfallet. (wisdom of the crowd)	14,6 % (7)	3,2 % (1)	4,8 % (13)
Baserte meg på et lignende, historisk til-felle som jeg kjente utfallet av. (bruk av én historisk analogi)	14,6 % (7)	19,4 % (6)	25,7 % (69)
Baserte meg på flere lignende, historiske tilfeller med forskjellige utfall. (bruk av flere historiske analogier)	27,1 % (13)	22,6 % (7)	19 % (51)
Tok utgangspunkt i dagens situasjon/ nivå, og justerte min prediksjon deretter. (ankring)	75 % (36)	74,2 % (23)	62,5 % (168)

Baserte meg på den siste utviklingen som hadde skjedd i saken, da spørsmålet ble stilt. (tilgjengelighetsheuristikk)	35,4 % (17)	29 % (9)	30,5 % (82)
Fordelte prosentene slik at jeg fikk best mulig score hvis jeg traff, men samtidig unngikk å få en veldig dårlig score hvis jeg bommet. (optimalisering av Brier-score)	29,2 % (14)	29 % (9)	25,7 % (69)
Baserte meg på en fremskrivning av den samme utviklingen som frem til nå. (ekstrapolasjon)	**54,2 % (26)**	51,6 % (16)	26,8 % (72)
Tenkte på hva som gjorde at jeg bommet/traff på tidligere spørsmål. (post-mortem analyse)	14,6 % (7)	12,9 % (4)	11,2 % (30)
Tok hensyn til uforutsigbare, overraskende hendelser som kunne påvirke utfallet. (sorte svaner)	20,8 % (10)	29 % (9)	15,6 % (42)
Annet.	0 % (0)	3,2 % (1)	4,1 % (11)

Tabell 5.19 Antall og andel deltagere som brukte ulike prediksjonsspesifikke tenkemåter.

Resultatene gjenspeiler det samme settet av prediksjonsspesifikke tenkemåter som ble trukket frem i analysene av deltagerne generelt. Det er en mye større andel av superforecasterne som lette etter informasjon fra flere kilder og brukte referanseklasser, grunnfrekvens og ekstrapolering når predikerte, mens de brukte mageføelsen langt sjeldnere enn begge de to andre kontrollgruppene. Selv om nesten halvparten av superforecasterne også baserte seg på intuisjon, er dette likevel en av tenkemåtene hvor forskjellen er størst sammenlignet med de andre deltagerne.

Superforecasterne tenkte derimot ikke vesentlig mer deduktivt eller induktivt enn andre deltagere, som også er i tråd med funnene fra forrige delkapittel, der det ikke var noen statistisk signifikante forskjeller mellom treffsikkerheten til deltagerne som brukte eller ikke brukte disse tenkemåtene. Forskjellene er også relativt små hva gjelder bruken av bakgrunnsinformasjonen som fulgte spørsmålene og det å ta utgangspunkt i dagens situasjon, som tilsier at superforecasterne var mindre avhengig av dette for å treffe bedre enn resten.

En interessant observasjon er at andelen deltagere som forsøkte å optimalisere sine Brier-scores, ved å fordele prosentene sine slik at de fikk best mulig score og samtidig begrense dette scoringssystemets spesielt harde straff ved høye sannsynlighetsestimater på galt svar, er tilnærmet lik for alle tre gruppene. Enten var det bare superforecasterne som forstod hvordan dette kunne gjøres, eller så var de andre bare dårligere til å angi prosenter til riktig svar. Gitt forskjellene i både forståelsen av Brier-scoresystemet og treffsikkerheten deres er det naturlig å tenke at det var en kombinasjon av disse faktorene som skilte superforecasterne fra resten.

De til dels store forskjellene i andelene metoder brukt skyldes imidlertid ikke at superforecasterne tenkte på flere forskjellige måter enn de andre deltagerne når de predikerte. I snitt krysset superforecasterne av for 6,6 prediksjonsspesifikke tenkemåter, de nest beste 6,0 og resten av deltagerne 5,5.[333] Disse forskjellen er svært små i praksis, gitt at deltagerne kunne velge mellom 18 alternativer. Dette understreker konklusjonen fra forrige delkapittel, nemlig at det først og fremst er *hvilke* tenkemåter superforecasterne brukte som skiller dem fra resten.

5.4.3 Diskusjon

For det første finnes det en gruppe deltagere som er signifikant bedre enn resten av deltagerne og dermed kan kalles «superforecastere» i FFIs turnering. Under like forhold treffer FFIs superforecastere faktisk like godt som GJPs. Dette er overraskende siden alle GJPs superforecastere fikk både opplæring i hvordan predikere best mulig og ble satt på grupper med andre like gode superforecastere, som hver for seg var tiltak med en beviselig forbedring i treffsikkerheten. At deltagere som predikerer alene kan treffe like godt som GJPs aller beste underbygges ytterligere av at vi også finner en tilsvarende gruppe superforecastere innenfor GJPs egen eksperimentgruppe med deltagere som hverken fikk opplæring eller ble satt på grupper med andre. Disse «ensomme» superforecasterne i GJP treffer også stort sett like godt som GJPs «fremdyrkede».

Den overraskende høye treffsikkerheten til alle de tre gruppene av superforecastere sammenlignet her reiser et spørsmål om det kanskje finnes en slags «øvre grense» for hvor godt det er mulig å bli til å predikere internasjonal politikk. Alle superforecasterne har en treffprosent på mellom 80 % og 90 % på binære spørsmål. Det er heller ikke mulig å bli særlig mye bedre kalibrerte enn de allerede er på binære spørsmål, uansett prediksjonstidspunkt. Selv om FFIs superforecastere treffer dårligere på kategoriske og ordinale spørsmål, treffer de ensomme superforecasterne i GJP betydelig bedre på denne typen spørsmål også, som viser at det likevel er mulig å treffe relativt godt på spørsmål med flere svaralternativer uten forbedringstiltak.

Et argument mot at det finnes en øvre grense for internasjonal politisk prediksjonsevne er at, om vi avgrenser analysene til bare de 2 % beste i stedet for de 7 % beste deltagerne som har blitt målt her, blir snittscorene og treffprosentene deres enda litt bedre. Da stiger for eksempel treffprosenten på binære spørsmål fra 81 % til 87 % blant GJPs egne superforecastere og fra 83 % til 86 % i FFIs turnering.[334] På den ene siden viser dette at det er mulig å øke den aggregerte treffsikkerheten ytterligere ved bruk av enda strengere seleksjonskriterier. På den annen side demonstreres det nok en gang at – selv blant de aller beste – er det i praksis ingen forskjell mellom GJPs superforecastere som ble dyrket frem gjennom ulike forbedringstiltak og FFIs superforecastere som predikerte helt alene uten noe ekstern innblanding.

[333] FFIs superforecastere vs. hhv. de neste beste: $t(70) = 0.94$, $p = 0.35$, og alle andre: $t(62) = 2.52$, $p = 0.01$.
[334] 2 % av de totalt 1751 og 833 deltagerne i GJP350s og FFIs datasett utgjør hhv. 35 og 17 superforecastere. Ved binære spørsmål og prediksjoner fra den første uken faller Brier-scorene til disse aller beste superforecasterne i GJP fra 0,28 til 0,20, mens treffprosenten stiger fra 81 % til 87 %. Til sammenligning faller de tilsvarende Brier-scorene i FFIs turnering fra 0,26 til 0,22, mens treffprosenten stiger fra 83 % til 86 %.

For det andre skiller FFIs superforecastere seg fra resten av deltagerne på samme måte som superforecasterne i GJP og scorer overraskende likt de amerikanske på alle variabler. Superforecastere synes altså å tilhøre en generell gruppe individer med høyere prediksjonsevne enn andre. De scorer bedre på alle de individuelle egenskapene som hang sammen bedre treffsikkerhet generelt: kognitiv kontroll, tallforståelse, politisk kunnskapsnivå, ønske om å havne blant de beste deltagerne, antall unike sannsynlighetsestimater, antall spørsmål besvart og tid brukt per spørsmål. I tillegg til disse egenskapene av generell betydning, er det her også en signifikant forskjell mellom superforecasterne og de andre deltagernes kognitive motivasjon. I underkapittel 5.3.3 ble det funnet tegn til en sammenheng mellom denne variabelen og treffsikkerhet, men den var ikke signifikant. Superforecasterne scorer imidlertid signifikant høyere på gleden av å engasjere seg i dypere tenkning enn deltagerne flest. Dette støtter antagelsen om at denne egenskapen også kan ha betydning for høyere prediksjonsevne.

Alle forskjellene er imidlertid svært små på de aller fleste variablene. Superforecasterne er altså ikke personer med superevner; de er bare generelt litt bedre enn resten. De la heller ikke ned en mye større innsats. Selv om egenskapen som korrelerer sterkest med treffsikkerheten er tiden deltagerne brukte på spørsmålene, bruker superforecasterne under 2 minutter per spørsmål i snitt og bare 30 sekunder mer enn deltagerne i turneringen generelt. Unntaket er scoren deres på forståelsen av scoringssystemet, der superforecasterne scorer både betydelig og signifikant enn selv de nest beste. Dette leder til en ny hypotese om at bedre forståelse av spillereglene kan være noe av det viktigste som kjennetegner de aller beste i prediksjonsturneringer. Denne variabelen ble imidlertid ikke undersøkt i GJP og dens relativt betydning i forhold til andre variabler må derfor undersøkes nærmere i senere analyser.

For det tredje bekreftes det at det finnes noen prediksjonsspesifikke tilnærminger som henger sammen med bedre treffsikkerhet. Superforecasterne stolte mindre på mageføIelsen, lette mer etter informasjon fra flere kilder og brukte oftere metoder forbundet med «utsideperspektivet». I tillegg nyanseres koblingene til tid brukt per spørsmål og det totale antallet tenkemåter deltagerne brukte, som hver for seg korrelerer med bedre treffsikkerhet. Superforecasterne brukte hverken mye mer tid eller mange flere teknikker enn de andre deltagerne. Tvert imot var det hvilke tenkemåter superforecasterne valgte å bruke tiden sin på som skilte dem fra resten.

Dette delkapittelets analyser viser samtidig at FFIs datasett har begrensninger når det gjelder mulighetene for å generalisere funnene om de norske superforecasterne. Antallet superforecastere og andelen av disse som er registrert med scores på alle uavhengige variabler er betydelig mindre enn GJPs totale antall superforecastere. Dette skyldes både forskjellen i måten de uavhengige variablene ble målt på og at det bare var mulig å identifisere superforecastere én gang i løpet av FFIs turnering. Forskjellene mellom de beste og nest beste deltagerne i FFIs turnering er sjeldent statistisk signifikante, slik som i GJP. Mønstret i forskjellene er imidlertid det samme som i GJPs på alle variablene. Det er derfor grunn til å tro at forskjellene ville blitt signifikante også i FFIs turnering om antallet deltagere i hvert utvalg var større. Samtidig består FFIs turnering av langt flere deltagere som predikerte alene enn i GJPs. FFIs turnering representerer derfor det største datasettet for å kunne etablere en «baseline» for hvor godt de aller beste delta-

gerne i en turnering klarer å predikere, uten at det iverksettes forbedringstiltak underveis. Resultatene viser at de aller beste treffer relativt godt, selv om de predikerer helt alene over flere år, og at de deler et sett med egenskaper som kan bidra til å identifisere dem på forhånd.

Dette reiser et siste spørsmål om hvorvidt nytten av tiltak for å oppnå en litt høyere presisjon utover en allerede god treffsikkerhet er verdt innsatsen det krever. Dette delkapittelets funn tyder på at den største effekten av eventuelle tiltak vil være å løfte den gjennomsnittlige treffsikkerheten til deltagerne generelt, mens det er relativt lite å hente for deltagere som i utgangspunktet er best. Reanalysen av GJPs datasett har også vist at forskjellen mellom deltagere som fikk forbedringstiltak og ikke blir mindre når treffsikkerheten kun baseres på prediksjoner fra starten av spørsmålsperioden. Dette tilsier en begrenset effekt av forbedringstiltak, også for deltagere generelt. Det å identifisere «de riktige folkene» kan derfor være enda viktigere i situasjoner der en ikke kan oppdatere sine prediksjoner underveis, som vil være tilfellet i de fleste situasjoner hvor det skal gjøres analyser av utviklingen i Norges forsvars- og sikkerhetspolitiske omgivelser.

6 Implikasjoner

Selv om det er umulig å vite helt sikkert hva som vil skje i fremtiden, er alle trusselvurderinger, fagmilitære råd og innspill til Forsvarets langtidsplanlegging avhengig av å gjøre antagelser om den videre utviklingen. Hvordan vil forholdet mellom Norge og Russland utvikle seg? I hvilke situasjoner vil det russiske regimet være villig til gå å bruke militærmakt mot naboland? Hvor sikre kan vi være på alliert støtte fra USA? Vil nye teknologier skape helt andre trusler enn dem vi forbereder oss på i dag? Hvor store vil de norske forsvarsbudsjettene bli fremover?

I praksis er alle svar på disse spørsmålene prediksjoner av fremtiden, selv om de sjeldent omtales og kanskje heller ikke anerkjennes som dette. Svarene vi legger til grunn er likevel helt avgjørende for utviklingen av norsk forsvars- og sikkerhetspolitikk og dermed landets reelle nasjonale sikkerhet. Fraværet av empiri om fremtiden gjør imidlertid at vi aldri vil vite om antagelsene vi gjør er riktige eller ikke. For å treffe best mulig tilsier funnene fra denne rapporten at det er helt avgjørende hvem vi hører på.

Dette kapittelet diskuterer først hvor overførbare funnene fra prediksjonsturneringer kan være til prediksjon av internasjonal politikk generelt. Her vises det hvordan de fleste individuelle egenskapene som henger sammen med bedre treffsikkerhet i turneringer trolig også vil gjelde prediksjon i den virkelige verdenen. Deretter argumenteres det for at prediksjonsturneringer kan være et verktøy som kan forbedre treffsikkerheten til prediksjonene som i dag gjøres av enkeltpersoner og små fagmiljøer bestående av kun profesjonelle fagfolk.

6.1 Prediksjon i den virkelige verdenen

Mens delkapittel 5.1 viste at funnene fra FFIs og GJPs turneringer er generaliserbare til prediksjonsturneringer generelt, er det minst like interessant å diskutere hvor overførbare svarene på hvor godt det er mulig å forutsi fremtiden og hvem som treffer best er til den virkelige verdenen.

I utgangspunktet er både temaene og tidsperspektivene som deltagerne ble bedt om å predikere i de to turneringene omtrent de samme som dem en forsøker å forutsi i forbindelse med trusselvurderinger og forsvarsplanlegging. I GJPs turnering ble spørsmålene utarbeidet av den amerikanske etterretningen selv, nettopp for å sikre spørsmål av relevans for amerikansk nasjonal sikkerhet. Spørsmålene i FFIs turnering var også basert på de viktigste temaene for norsk sikkerhet, som krig og konflikt, Russland, USA, Europa, økonomi, teknologi og norsk politikk. Selv om det utenfor etterretningsmiljøer sjeldent ville blitt formulert like konkrete spørsmål som dem deltagerne fikk, er innholdet i dem representative for utviklingene som må vurderes i forsvars- og sikkerhetspolitiske analyser. Mens de fleste spørsmålene i GJP hadde et tidsperspektiv på noen få måneder, som er mest relevant for løpende etterretningsvurderinger, er snittet til FFIs foreløpig avgjorte spørsmål rundt ett år, som er nøyaktig så langt frem de årlige trusselvurderingene forsøker å predikere. Når resten av spørsmålene i FFIs turnering avgjøres vil datagrunnlaget også inneholde av mange spørsmål med flere års perspektiv, som er mest relevant for Forsvarets fireårige langtidsplaner.

Å studere hvordan treffsikkerheten varierte på spørsmålene innad i FFIs og GJPs turneringer kan derfor gi en pekepinn på hvor presist en kan forvente å kunne forutsi samme temaer og tidsperspektiver i den virkelige verdenen. Et overraskende funn er at det ikke er noe som tyder på systematiske forskjeller mellom hvor godt vi kan predikere forskjellige temaer, selv om det ofte hevdes at noen temaer, som økonomi og teknologi, er preget av mer usikkerhet enn andre. Enten er de ikke vanskeligere å predikere disse temaene eller så er usikkerheten ved de andre temaene så store at det ikke er mulig å predikere dem heller noe mer presist. Et annet overraskende funn er fraværet av en sammenheng mellom tidsperspektiv og treffsikkerhet i begge turneringer. Dette utfordrer antagelsen om at det blir vanskeligere å predikere jo lenger inn i fremtiden en ser, i alle fall innenfor tidsperspektivene på et halvt til to år som så langt har blitt undersøkt her.

Samtidig tilsier forskjellene i treffsikkerheten til deltagerne i FFIs og GJPs turneringer at det er vanskelig å tallfeste nøyaktig hvor godt det er mulig å predikere internasjonal politikk. Mens FFIs deltagere er i snitt like dårlige til å predikere som tilfeldig gjetning, var GJPs deltagere langt bedre enn begge. At deltagerne er like på tvers av turneringene og avstanden mellom dem holder seg stabil på tvers av temaer, typer spørsmål og tidsperspektiver, tyder imidlertid på at det er gjennomføringen av turneringene, ikke deltagerne eller spørsmålene i seg selv, som er mest avgjørende for forskjellen i treffsikkerhet. Dette handlet både om måten deltagerne kunne predikere på og hvordan treffsikkerheten deres ble forsøkt forbedret i GJPs turnering, men ikke i FFIs. Når deltagerne i GJP kunne oppdatere sine prediksjoner helt frem til spørsmålet ble avgjort og prediksjonsevnen deres ble forbedret underveis (gjennom opplæring i probabilistisk tenkning og ved å bli satt i grupper), klarte de å forutsi riktig utfall på tre av fire spørsmål med to mulige utfall og var tilnærmet perfekt kalibrerte. De var også betydelig bedre enn FFIs deltagere til å oppgi høye sannsynligheter til riktige svar og lave til de gale, uansett spørsmålstype.

Når vi derimot ser på deltagerne i GJP som ikke ble forsøkt forbedret underveis og baserer treffsikkerheten deres kun på prediksjoner fra starten av spørsmålsperioden, slik som i FFIs turnering, reduseres gapet betydelig. Da har både FFIs og GJPs deltagere en evne til å forutsi riktig utfall på to av tre binære spørsmål, men FFIs deltagere er fortsatt dårligere til å estimere sannsynlighetene til ulike svaralternativer. En mulig forklaring er at alle deltagerne i GJP fikk en innføring i hvordan Brier-scoren (som ble brukt til å måle treffsikkerheten i begge turneringer) ble beregnet, og fikk vite at systemet straffe høye sannsynlighetsestimater på gale svar spesielt hardt. FFIs deltagere fikk ingen slik innføring og scoret svært lavt på forståelsen av scoringssystemet. Igjen handler dette i så fall om forskjeller i gjennomføringen av turneringene.

Svaret på hvor godt det er mulig å forutsi forsvars- og sikkerhetspolitikk er altså avhengig av hvordan predikeringen gjøres. Det er imidlertid også avgjørende hvilke enkeltpersoner sine prediksjoner vi baserer oss på. I begge turneringer finnes det nemlig en gruppe deltagere – superforecastere – som er betydelig mer treffsikre enn resten og omtrent like gode som hverandre på tvers av turneringene, uavhengig av om de ble forsøkt forbedret eller ikke. Både FFIs og GJPs superforecastere klarer å forutse riktig utfall på fire av fem spørsmål med to mulige utfall.

Begge turneringenes superforecastere er også like gode til å oppgi høye sannsynligheter til riktig svar og lave til de gale på slike binære spørsmål. I tillegg er de tilnærmet perfekt kalibrerte, som betyr at en kan anta at de treffer like presist som de predikerer. Dette skyldes ikke at superfore-

casterne var mer forsiktige i sine prediksjoner. Tvert imot oppgav superforecasterne i begge turneringene en snittprediksjon på 80 % sannsynlighet for det som viste seg å bli riktig utfall på de binære spørsmålene. Med en så høy snittprediksjon og treffsikkerhet er det vanskelig å se for seg hvordan personer i den virkelige verdenen kunne truffet så mye bedre enn dette.

Svaret på hvor godt det er mulig å forutsi forsvars- og sikkerhetspolitiske spørsmål synes altså å ligge et sted mellom i verste fall slite med å slå tilfeldig gjetning og i beste fall klare å forutsi riktig utfall på fire av fem spørsmål og samtidig være perfekt kalibrert i sine sannsynlighetsvurderinger av riktig utfall. For å treffe best mulig, uansett måten prediksjoner samles inn på, er vi altså avhengig av å kunne identifisere «de riktige folkene» på forhånd.

Det er heldigvis flere grunner til å tro at mange av de individuelle egenskapene som henger sammen med bedre treffsikkerhet i FFIs og GJPs turneringer også er overførbare til prediksjon i den virkelige verdenen. For det første samsvarer funnene med forskningen fra bedømmings- og beslutningspsykologien. De samme disposisjonelle variablene som korrelerer med bedre treffsikkerhet i begge turneringer – som kognitiv kontroll, tallforståelse, kunnskapsnivå og interesse for dypere tenkning – henger også sammen høyere prestasjonsnivå på andre områder, som hukommelse, reaksjonstid, risikovurdering og evne til å unngå tankefeil.[335] Deltagere som er gode til å predikere er altså høyst sannsynlig også gode på andre ting. Dette understreker hvordan treffsikkerhet er en evne, som vi i likhet med andre evner ikke anser som tilfeldig. I tillegg er de kognitive evnene faste variabler som heller ikke påvirkes av hvordan prediksjoner samles inn.

For det andre er det grunn til å anta at relevansen av de prediksjonsspesifikke tenkemåtene fra FFIs turnering også er overførbar til reelle situasjoner, basert på det deltagerne selv sier. Etter at turneringen var over fikk nemlig alle deltagerne som arbeidet med forsvars- og sikkerhetspolitikk spørsmål om de ville tenkt annerledes hvis de hadde blitt bedt om å predikere de samme spørsmålene i sin virkelige jobb. Her svarte et flertall på 69 % av deltagerne ja.[336] På spørsmål om hva de ville gjort annerledes var det derimot ikke snakk om helt andre tilnærminger enn dem som ble brukt i turneringen. Så godt som alle deltagerne svarte at de ville samlet inn mer informasjon, brukt mer tid på spørsmålene og gjort grundigere analyser – altså mer av de metodene som fungerte, ikke andre tenkemåter enn dem som har blitt kartlagt i denne rapporten. Sammenhengene med mer informasjonsinnsamling og tidsbruk tilsier derfor at disse personene antageligvis ville ha truffet enda bedre i virkeligheten, gitt at de faktisk gjorde disse tingene.

Samtidig har funnene fra FFIs turnering vist at det ikke er alle former for grundigere analyser som bidrar til bedre treffsikkerhet. Bruk av intuisjon eller det første som faller en inn som det mest sannsynlige var forbundet med dårligere treffsikkerhet. Bruk av induktiv eller deduktiv resonnering, som var det grunnleggende skillet mellom reve- og pinnsvinekspertene i EPJ, ser derimot ikke ut til å ha noe si for treffsikkerheten. Denne typen tilnærminger var imidlertid utbredt blant deltagerne. Forsøk på å ta hensyn til uforutsigbare, overraskende hendelser, som ofte anbefales i metodelitteraturen til fremtidsstudier, hadde ingen betydning hverken i FFIs

[335] Se f.eks. Frederick (2005), 'Cognitive Reflection and Decision Making'; Cokely mfl. (2012), 'Measuring Risk Literacy: The Berlin Numeracy Test', og Cacioppo og Petty (1982), 'The Need for Cognition'.
[336] Basert på 112 deltagere som både oppfylte minstekravet og de arbeidet med forsvars- og sikkerhetspolitikk.

eller GJPs turneringer. Derimot traff deltagerne bedre hvis de baserte seg på grunnfrekvens, referanseklasser og flere historiske analogier med ulike utfall.

Et funn med potensielt stor betydning utover FFIs turnering er at hvilke prediksjonsspesifikke tenkemåter en velger å bruke ikke må baseres på hvor godt grunnlag en selv tror at en har for å svare på de aktuelle spørsmålene. Det er nemlig ingen sammenheng mellom hvor mange spørsmål deltagerne mente de hadde et godt grunnlag for å predikere eller hvor sikre de var på prediksjonene på hvert enkelt spørsmål og hvor godt de faktisk traff. Det er heller ingen sammenheng med mellom treffsikkerheten og hvor gode deltagerne selv tenkte at de var til å predikere generelt. Når deltagerne registrerte seg ble nemlig alle bedt om å vurdere sin egen evne til å forutsi forsvars- og sikkerhetspolitiske utviklinger. Basert på en skala fra 1 til 7, der 1 var «svært dårlig» og 7 var «svært god», var median scoren 4. Flertallet svarte altså at de var hverken god eller dårlig til å predikere, mens en tredel mente at de var litt eller ganske god.[337]

Det er imidlertid ingen korrelasjon mellom deltagernes vurdering av egen prediksjonsevne og treffsikkerheten eller forskjeller mellom snittscorene til deltagerne som rangerte evnen ulikt. Deltagerne burde derfor ha benyttet de samme prediksjonsspesifikke tenkemåtene forbundet med bedre treffsikkerhet, uavhengig av hvilket grunnlag de selv mente de hadde. De tenkemåtene som bidro til bedre treffsikkerhet i turneringen er også forbundet med bedre vurderingsevne og treffsikkerhet i helt andre settinger, som når vi skal anslå tiden det vil ta å fullføre prosjekter.[338] I likhet med evnene nevnt over, er det altså også grunn å tro at de spesifikke tilnærmingene identifisert i FFIs turnering også er forbundet med bedre treffsikkerhet i virkeligheten.

De relativt sikre sammenhengene mellom individuelle egenskaper og treffsikkerhet har implikasjoner for rekruttering av personell som skal predikere internasjonal politikk. Personer som treffer bedre enn andre scorer generelt høyere på tester av de disposisjonelle evnene nevnt over. Resultatene viser derimot at personers generelle kognitive stiler ikke er like overførbare til prediksjonsevne. Det hjelper lite at en scorer høyt på tenkemåter forbundet med bedre treffsikkerhet, hvis ikke de også anvendes i praksis. Samtidig må en antageligvis være disponert for de relevante tenkemåtene for å kunne praktisere dem. I prediksjonssettinger kan tester av generelle kognitive stiler derfor være nyttige til å ekskludere individer med dårlige forutsetninger. For å sikre bedre treffsikkerhet må en samtidig oppfordre til å bruk av «de riktige teknikkene».

Treffsikkerhet er imidlertid bare ett av suksesskriteriene i forsvarsplanlegging. Prediksjon vil ha en mindre sentral rolle i utviklingen av scenarioer, der hensikten er å identifisere et spekter av mulige trusler, ikke forutsi hvilke av dem som vil skje. Her kan andre tilnærminger, som det å identifisere uforutsigbare, overraskende hendelser, være mer nyttige enn dem som bidrar til treffsikkerhet. Det anses også som umulig å forutsi presist det viktigste spørsmålet i all forsvarsplanlegging; nemlig, hvordan den neste krigen vil se ut. Til det er krig et for sjeldent sosialt fenomen preget av enda større usikkerhet enn internasjonal politikk generelt. Det viktigste i forsvarsplanlegging er derfor å bomme så lite at det er mulig å tilpasse seg når krigen kommer.

[337] Her svarte 223 (27 %) av 833 deltagere at de var «litt god», 58 (7 %) at de var «ganske god», 427 (51 %) svarte at de var «hverken dårlig eller god», 69 (8 %) svarte at de var «litt dårlig» og 47 (6 %) at de var «ganske dårlig».
[338] For mer om utsideperspektivet, se kapittel 23 i Kahneman (2013), *Tenke, fort og langsomt*.

Samtidig er prediksjon umulig å unngå helt, så lenge det må gjøres vurderinger av hvilke trusler som er mulige eller ikke. Resultatene fra FFIs turnering har også vist at det er mulig å forutsi relevante utviklinger i denne sammenhengen. En fornuftig ambisjon kan derfor være å forutsi det som kan forutsies best mulig og å bomme minst mulig på det som ikke kan predikeres.

6.2 Turneringer som verktøy

De individuelle egenskapene som er minst overførbare til den virkelige verdenen, er dem som er knyttet til deltageres innsats i prediksjonsturneringer, som antall spørsmål besvart, antall prediksjoner per spørsmål og tid brukt per spørsmål. De færreste utrednings- og beslutningsprosesser er konkurranser og personene involvert i disse blir sjeldent målt på eller stilt til ansvar for feilslåtte prediksjoner. Gitt at disse egenskapene er avhengig av turneringssettingen, betyr det at – med mindre vi simulerer den samme turneringssettingen i etterretningstjenestene, forskningsinstituttene og relevante departementer – vil ikke disse variablene være relevante.

Personene som deltok i FFIs eller GJPs turneringer er heller ikke representative for dem som predikerer samme type spørsmål i den virkelige verdenen, der vurderinger av fremtidige internasjonale politiske utviklinger er forbeholdt profesjonelle eksperter. Ekspertene er en mangfoldig gruppe – som akademikere, offiserer og etterretningsanalytikere med svært ulik bakgrunn – men de har til felles at de er selektert på bakgrunn av formell kompetanse eller erfaring. Vi kan derfor ikke bruke turneringsresultatene til å si noe om hvor godt det norske forsvars- og sikkerhetspolitiske miljøet er til å predikere generelt. Samtidig viser resultatene at, hvis vi bare hadde avgrenset oss til bare forsvars- og sikkerhetspolitiske fagfolk i Norge, hadde vi gått glipp av flesteparten av superforecasterne. Rundt 60 % av superforecasterne har nemlig aldri arbeidet med forsvars- og sikkerhetspolitiske spørsmål tidligere og er derfor amatører på dette fagfeltet.

Dette reiser et spørsmål av stor praktisk betydning for hvordan prediksjoner bør genereres: Er det mest hensiktsmessig å basere trusselvurderinger og langtidsplaner på antagelsene til eksperter som arbeider med og har tid til å sette seg grundig inn i temaene, eller å ha hundrevis av personer som bruker noen minutter på å predikere mange forskjellige spørsmål de både har og ikke har kompetanse på? Svaret er ikke åpenbart.

På den ene siden har eksperimentene i GJP vist at det er mulig å forbedre treffsikkerheten til personers sannsynlighetsvurderinger av internasjonale politiske hendelser, gjennom enkle tiltak som korte kurs i probabilistisk tenkning og å sette deltagere sammen i grupper. Disse funnene fremheves som svært lovende, nettopp fordi profesjonelle analyseorganisasjoner vil ha betydelig større muligheter til å trene og organisere sine ansatte enn deltagerne i GJP. Dette tilsier også at det er god grunn til å anta at de samme tiltakene sannsynligvis ville økt treffsikkerheten til fagmiljøene som i dag er ansvarlige for å vurdere utviklingen i Norges strategiske omgivelser.

På den annen side viser denne rapportens sammenligning av resultatene fra FFIs og GJPs turneringer at disse tiltakene ikke nødvendigvis er de mest effektive, hvis målet er å komme frem til de aller mest treffsikre prediksjonene. Selv basert på bare 25 spørsmål er det mulig å identifisere personer som med stor sannsynlighet vil treffe systematisk bedre enn andre på påfølgende

spørsmål. Ikke minst er det mulig å identifisere superforecastere som treffer så godt som det synes å være mulig å forutsi internasjonal politikk, uten noen tiltak.

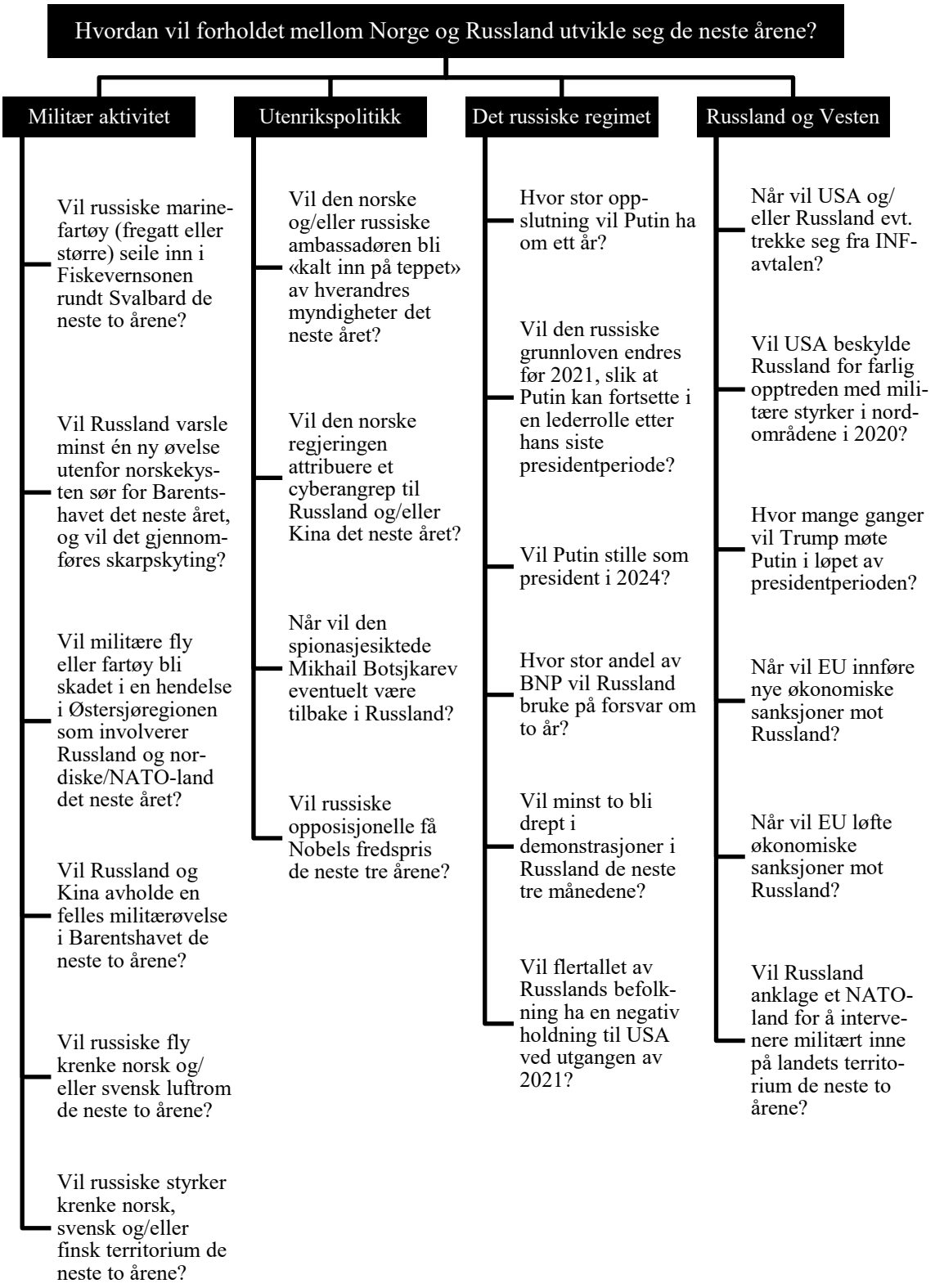
Turneringer representerer derfor et potensielt mer treffsikkert alternativ til dagens avhengighet av prediksjoner fra kun eksperter, som forskningen har vist ikke er spesielt gode, hverken relativt eller objektivt sett. I det minste kan superforecasteres prediksjoner brukes til å etterprøve prediksjonene som uansett gjøres innenfor de små, ofte lukkede, fagmiljøene med ansvar for etterretning og forsvarsplanlegging. I ytterste fall kan en «outsource» sannsynlighetsvurderingene til superforecastere i prediksjonsturneringer. Dette høres radikalt ut, men gitt at de beste deltagerne i både FFIs og GJPs turneringer har en høy treffprosent og er tilnærmet perfekt kalibrerte, er det lite sannsynlig at de vil treffe dårligere enn fagfolkene som brukes i dag.

Til forskjell fra måten det vanligvis predikeres på i dag, representerer turneringer også en metode som motvirker én av de vanligste tankefeilene i forbindelse med prediksjon, nemlig *the wrong side of maybe*-fallgruven. Denne handler om at vi i retrospekt har en tendens til å tolke vage formuleringer, som «kan» og «muligens», slik at de faller på riktig side av *maybe*-grensen (50/50 %). Slike språklige sannsynlighetsbegreper er den vanligste måten å formulere prediksjoner på. De gir imidlertid alltid mulighet til å endre tolkning av hvor sannsynlig vi mente det var etter at vi vet svaret. Dermed risikerer vi å fortsette å predikere galt, uten at det får konsekvenser for fremgangsmåten. Denne faren unngås ved å basere seg på probabilistiske prediksjoner og bruke dem til å måle treffsikkerheten og justere forventningene til hvor godt vi kan anta å treffe.

Hvordan kan turneringer brukes til å støtte etterretningsarbeid og forsvarsplanlegging i praksis? Den største utfordringen er at for å kunne måle treffsikkerheten, må spørsmålene som stilles være falsifiserbare. Det vil si at de må være så konkrete og avgrensede i tid og rom at det er mulig å slå fast om en hendelse faktisk skjer eller ikke, som: «Vil Putin stille som kandidat i Russlands presidentvalg i 2024?». Denne typen små spørsmål kan imidlertid ikke gi direkte svar på store spørsmålene som vi egentlig er mest interessert i, som: «Hvordan vil forholdet mellom Norge og Russland utvikle seg de neste årene?». For å ivareta dette kan vi benytte en metode som heter *question clustering*, der vi lager klynger av mindre, falsifiserbare spørsmål, hvis samlede svar kan gi en indikasjon på det store spørsmålet vi egentlig ønsker å kunne forutsi.[339]

Figur 6.1 viser et eksempel på en klynge med spørsmål som har blitt stilt i FFIs turnering om forholdet mellom Norge og Russland. Her er disse inndelt i fire parametere av særlig betydning: russisk militær aktivitet i Norges nærområder, det utenrikspolitiske forholdet, det russiske regimets utvikling og forholdet mellom Russland og Vesten. De konkrete spørsmålene ba deltagerne forutsi om russisk militær aktivitet ville bryte med etablerte mønstre, om det ville skje politiske hendelser som forsuret forholdet mellom de norske og russiske regjeringene, om Putin ville fortsette som Russlands viktigste leder og hvorvidt spenningen vis-a-vis Norges allierte vil avta eller øke. Bakgrunnen for alle spørsmålene var hendelser i den virkelige verdenen som gjorde at disse mulige hendelsene og utviklingene ble diskutert i fagmiljøer og media. Flere av spørsmålene i figur 6.1 ble stilt flere ganger over flere år, men er her forenklet og gjort tidløse.

[339] For mer om denne teknikken, se Tetlock og Gardner (2015), *Superforecasting*, ss. 261–266.



Figur 6.1 Spørsmålsklynge om utviklingen i forholdet mellom Norge og Russland.

Gitt at det er mulig å forutsi svarene på de små spørsmålene i figur 6.1, er det også mulig å si noe om hvilken retningen forholdet mellom Norge og Russland sannsynligvis vil utvikle seg de neste årene. Etter at 18 av spørsmålene er avgjort så langt, viser resultatene at det ikke er lett, men mulig å forutsi svarene på disse spørsmålene. Med en treffprosent på 87 % klarte superforecasterne i FFIs turnering å forutsi det riktige utfallet på fem av de seks binære spørsmålene de fikk.[340] I tillegg var de sikre i sine prediksjoner på riktig svar, som betyr at de på forhånd klarte å oppgi høye sannsynligheter til det som skjedde og lave til det som ikke skjedde. I praksis oppgav superforecasterne en 76 % sannsynlighet for det utfallet som viste seg å stemme. I tillegg hadde de en underkonfidens på 6 % på binære spørsmål. De hadde dermed grunn til å være enda sikrere i sine prediksjoner. Som i turneringen generelt, er ikke superforecasterne like treffsikre på spørsmål med flere svaralternativer. I snitt klarte de å forutsi riktig utfall på tre av de seks kategoriske spørsmålene og to av de seks ordinale som ble stilt. Dette var likevel bedre enn resten av deltagerne og tilfeldig gjetning, uansett treffsikkerhetsmål.[341]

På den ene siden viser eksempelet at det *er* mulig å forutsi med relativt høy grad av sikkerhet utfall på binære spørsmål som indikerer retningen på utviklingen i Norges forhold til Russland. På den annen side sliter selv de aller beste med å forutsi utfallene på kategoriske og ordinale spørsmål som kan gi de mest nyanserte svarene, fordi en her ikke bare må forutsi *om* noe vil skje eller ikke, men også *hvem*, *når* eller *hvor mye eller lite* som blir riktig svar. Dette kan så tvil om verdien av å bruke turneringer til å predikere denne typen spørsmål. Alternativt kan vi omformulere og lage flere binære spørsmål i stedet for ordinale og kategoriske.

Problemet er at ekspertene som vi normalt ville spurt, treffer mye dårligere på de samme spørsmålene. Til sammenligning var treffprosenten til fagfolkene med Russland-kompetanse i FFIs turnering bare 61 % på de binære spørsmålene. Dette var faktisk dårligere enn resten av deltagere, inkludert amatører, som hadde en treffprosent på 67 % på samme spørsmål. Ikke bare treffer Russland-ekspertene dårligere enn de andre deltagerne, men de har også en svært høy overkonfidens (20 %) og høyere enn resten (15 %). Her er det viktig å nevne at ekspertene og amatørene konkurrerte under helt like forhold, og at de brukte like lang tid per spørsmål og svarte på omtrent like mange spørsmål hver. Dette virker kanskje overraskende, men gjenspeiler det generelle funnet gjort i FFIs turnering, der det ikke var noen signifikant forskjell mellom treffsikkerheten til fagfolk som predikerte innenfor og utenfor sine egne kompetanseområder.

Det er heller ikke noe grunnlag for å anta at ekspertene ville ha truffet mye bedre enn de andre deltagerne i virkeligheten. Forskningen innenfor kognitiv psykologi har vist at eksperters treffsikkerhet er begrenset også på helt andre områder preget av betydelig usikkerhet i den virkelige

[340] Superforecastere: Binære: Brier-score: 0,21, treffprosent: 87 % og kalibrering: –6 %. Kategoriske: Brier-score: 0,69, treffprosent: 46 % og kalibrering: 16 %. Ordinale: Brier-score: 0,60, treffprosent: 27 % og kalibrering: 25 %.
[341] Resten av deltagerne: Binære: Brier-score: 0,51, treffprosent: 65 % og kalibrering: 17 %. Kategoriske: Brier-score: 0,90, treffprosent: 40 % og kalibrering: 32 %. Ordinale: Brier-score: 0,67, treffprosent: 25 % og kalibrering: 38%. Tilfeldig gjetning: Binære: Treffprosent: 50 %. Kategoriske: Treffprosent: 29 %. Ordinale: Treffprosent: 23 %.

verdenen – fra å forutsi resultatene av forskningsprosjekter til diagnostisering av psykiske sykdommer og vurdering av tilståelser i politiavhør.[342] Internasjonal politikk er intet unntak.

Turneringer er altså ikke ingen fasit på utfordringene med prediksjon av internasjonal politikk, men superforecastere kan være de beste vi har. Prediksjonene samlet inn i FFIs turnering kan også brukes til å svare på andre «store» spørsmål av relevans for norsk sikkerhet, som utviklingen i det transatlantiske forholdet, faren for væpnet konflikt i Europa, terrortrusselen, forsvarsøkonomiske premisser og trender i krigføring. Superforecasterne treffsikkerhet på de binære spørsmålene i eksempelet om forholdet til Russland er også representativ for prediksjonsevnen deres i turneringen generelt, men de treffer i snitt betydelig bedre på kategoriske og ordinale spørsmål. I tillegg treffer GJPs superforecastere (inkludert dem som predikerte alene) enda bedre, som betyr at det finnes et reelt forbedringspotensial også på disse spørsmålstypene.

Turneringer er også et potensielt svært kostnadseffektivt tiltak for å forbedre treffsikkerheten på spørsmål om nasjonal sikkerhet, der selv litt bedre presisjon kan ha mye å si. Driftskostnadene ved FFIs turneringer beløper seg til et par hundre tusen kroner fordelt over tre år, som utgjør under halvparten av ett enkelt forskerårsverk. FFI har allerede utviklet verktøyene som trengs for gjennomføre en ny, tilsvarende turnering i fremtiden. Turneringer må heller ikke stekke seg over flere år. Siden individers prediksjonsevne holde seg stabil over tid er det også mulig å samle inn tusenvis av prediksjoner på mange spørsmål samtidig, for eksempel i forkant av nye trusselvurderinger og langtidsplaner for Forsvaret.

Det mest ressurskrevende med prediksjonsturneringer er spørsmålsgenereringen. Her kan en imidlertid bare ta utgangspunkt i de samme parameterne og indikatorene som uansett utvikles i forbindelse med etterretningsanalyser og scenarioutvikling i dag. Her spiller de profesjonelle fagmiljøene en avgjørende rolle, ettersom evnen til å predikere godt ikke er den samme som evnen til å stille «de riktige spørsmålene». Som nevnt er prediksjonsevne er bare én av flere relevante egenskaper i forsvarsplanleggingsprosesser. Det er også nødvendig å vurdere konsekvenser av hendelsene en forutser, utrede mulige løsninger og identifisere hvilke utviklinger det er viktig å følge med på i første omgang. En mulig rollefordeling kan derfor være at eksperter lager spørsmålene, mens predikeringen gjøres av dem som treffer best, uansett hvem de er.

Det er likevel ikke sikkert at det er mulig å predikere internasjonal politikk «godt nok». Det er til slutt politiske og militære ledere som sitter med ansvaret og risikoene forbundet med eventuelle beslutninger som tas på bakgrunn av prediksjoner. Selv om FFIs og GJPs resultater viser at det er mulig å treffe relativt presist, er det ikke sikkert at de aggregerte prediksjonene samlet inn gjennom turneringer anses som gode nok til å bruke dem som beslutningsgrunnlag.

Problemet er at de samme vurderingene og beslutningene tas i dag, uten å vite hvor god treffsikkerheten til prediksjonene de allerede baserer seg på egentlig er. Dette metodiske gapet er bakgrunnen til at det lenge har vært argumentert for å innføre tallfestede sannsynlighetsvurderinger

[342] Cassidy og Buede (2009), 'Does the accuracy of expert judgment comply with common sense: caveat emptor'; McBride mfl. (2012), 'Evaluating the accuracy and calibration of expert predictions under uncertainty'.

i etterretningssammenheng, til hvorfor Tetlock først begynte å måle eksperters treffsikkerhet i EPJ og til at amerikansk etterretning valgte å sponse GJP.

Det viktigste argumentet for tallfesting er å unngå feilbeslutninger. Et eksempel som ofte trekkes frem er USAs mislykkede forsøk på å invadere Cuba i 1961.[343] Bare tre måneder etter at John F. Kennedy ble president iverksatte han en operasjon som skulle styrte Castro-regimet ved å landsette en liten styrke med eksilcubanere i Grisebukta. Det militære lederskapet hadde i forkant sagt at planen hadde en «*fair chance*» for å lykkes. Mannen bak denne vurderingen har senere uttalt at han med dette hadde ment at operasjonen hadde en rundt 30 % sjanse for å lykkes. Kennedy fikk bare kommunisert den språklige formuleringen og tolket denne som en høyere enn 30 % sannsynlighet. Vi kan aldri vite om Kennedys beslutning ville ha vært en annen om han hadde fått en tallfestet sannsynlighet å forholde seg til, men det er mulig han ville ha tenkt seg mer om før han iverksatte en operasjon som viste seg å bli en politisk katastrofe.

Det eksisterer likevel fortsatt en sterk aversjon mot å tallfeste sannsynlighetsvurderinger i etterretningssammenheng og i det hele tatt forsøke å predikere internasjonal politikk generelt.[344] Motargumentene er også gode. Det å tallfeste noe kan skape et inntrykk av at vi kjenner usikkerhetene bedre enn vi gjør. Dette kan lede til feilbeslutninger, fordi vi får større tillit til vurderingene som kommuniseres enn om de samme resultatene hadde blitt kommunisert uten tallverdier. Som Kahnemans forskning har vist, er det å tenke statistisk heller ikke noe som faller mennesker naturlig, og selv forskere på høyt nivå har vanskeligheter med å forstå statistikk og resonnere probabilistisk.[345]

Resultatene fra FFIs og GJPs turneringer utfordrer disse motargumentene og tilsier at vi bør forsøke å tallfeste sannsynlighetsvurderinger likevel. Prediksjoner samlet inn gjennom turneringer kan ikke bare brukes til å kommunisere hvor sikre vi er, men også til å etablere en baseline for hvor mye vi kan stole på dem. Når vi vet hvor godt noen treffer, kjenner vi også usikkerheten ved prediksjonene deres. Det kan for eksempel tenkes at Kennedy ville ha vektlagt prediksjoner på 33 % for å lykkes forskjellig, hvis den ene personen som fremsatte dem normalt traff på nesten alle prediksjonene sine, mens den andre traff på bare halvparten. Selv om forskere generelt sliter med tenke probabilistisk, er det også noen som er gode på det, og de treffer bedre enn andre. Faktisk finnes det enkeltindivider som vi på forhånd kan si at kommer til å treffe relativt godt, men vi kan ikke bruke de samme kriteriene som vi bruker i dag til å identifisere dem.

Nettopp fordi det er vanskelig å forutsi internasjonal politikk, og det er få områder hvor konsekvensene av feilslåtte antagelser er større enn innenfor forsvars- og sikkerhetspolitikk, er det viktig å ha en riktig forståelse av hvor godt vi klarer å predikere og hvem vi bør høre mest på.

[343] For mer om dette eksempelet, se Tetlock og Gardner (2015), *Superforecasting*, ss. 55–56, 193–195 og s. 200.
[344] Nygård, H. M. (2015), 'Prediksjon i Internasjonal politikk', *Internasjonal Politikk*, 73:4, ss. 467–487.
[345] Kahneman (2013), *Tenke, fort og langsomt*, s. 124. Se også Schrodt, P. A. (2013), 'Seven deadly sins of contemporary quantitative political analysis', *Journal of Peace Research*, 51:2, ss. 287–300.

A Kognitive tester

Dette vedlegget inneholder alle de kognitive testene som deltagerne fikk i FFIs turnering. Hvordan testene ble gjennomført er nærmere beskrevet i kapittel 4. Resultatene fra disse danner grunnlaget for scorene på de fleste uavhengige variablene analysert i kapittel 5.

Det finnes ofte flere versjoner av de samme testene, for eksempel med forskjellige antall oppgaver eller påstander brukt til å måle de bakenforliggende egenskapene. Versjonene som er brukt i FFIs turnering er de samme som i GJP. Her oppgis kildene til de spesifikke testversjonene, og det vises til disse for instruksjoner om hvordan scorene beregnes. De fleste kognitive testene måtte imidlertid oversettes til norsk i forbindelse med FFIs turnering. Oversettelsene ble kvalitetssikret av en ekstern person med engelsk som morsmål og som kan norsk flytende. I tillegg til de kognitive testene beskrevet i dette vedlegget fikk deltagerne også en test av abstrakt resonneringsevne (*Shipley-2 Block Patterns*). Denne er opphavsbeskyttet og kan derfor ikke gjengis her.

A.1 Kognitiv kontroll

Kilde: Frederick, S. (2005), 'Cognitive Reflection and Decision Making', *Journal of Economic Perspectives*, 19:4, ss. 25–42.

Du får nå tre oppgaver med varierende vanskelighetsgrad. Svar på så mange du klarer.
Et balltre og en ball koster 1 dollar og 10 cent til sammen. Balltreet koster 1 dollar mer enn ballen. Hvor mye koster ballen? • _____ cent [Riktig svar: 5 cent]
Hvis 5 maskiner bruker 5 minutter på å lage 5 dingser, hvor lang tid bruker 100 maskiner på å lage 100 dingser? • _____ minutter [Riktig svar: 5 minutter]
I en innsjø er det et felt med vannliljer. Hver dag dobles størrelsen på vannliljefeltet. Hvis det tar 48 dager før feltet med vannliljer dekker hele innsjøen, hvor lang tid tar det før feltet dekker halvparten av innsjøen? • _____ dager [Riktig svar: 47 dager]

A.2 Kognitiv kontroll – utvidet

Kilde: Baron, J. Scott, S. Fincher, K. og Metz, S. E. (2015), 'Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)?', *Journal of Applied Research in Memory and Cognition*, 4:3, ss. 265–284.

Du vil nå få 18 oppgaver med varierende vanskelighetsgrad. Svar på så mange du klarer.
Alle ting som røykes er godt for helsen. Sigaretter røykes. Hvis disse to påstandene stemmer, kan vi konkludere fra dem at sigaretter er godt for helsen? • Ja [Riktig] • Nei
Alle laloobayer er rike. Sandy er en laloobay. Hvis disse to påstandene stemmer, kan vi konkludere fra dem at Sandy er rik? • Ja [Riktig] • Nei
Alle bedriftseiere er rike. Bill Gates er en bedriftseier. Hvis disse to påstandene stemmer, kan vi konkludere fra dem at Bill Gates er rik? • Ja [Riktig] • Nei
Alle blomster har kronblader. Roser har kronblader. Hvis disse to påstandene stemmer, kan vi konkludere fra dem at roser er blomster? • Ja • Nei [Riktig]
Et balltre og en ball koster 96 cent til sammen. Balltreet koster 2 cent mer enn ballen. Hvor mye koster ballen? • ___ cent [Riktig svar: 47 cent]
Hvis 1 maskin bruker 10 minutter på å lage 5 dingser, hvor lang tid vil det ta 10 maskiner å lage 600 dingser? • ___ minutter [Riktig svar: 120 minutter]
I en innsjø er det et felt med vannliljer. Hver dag firedobles størrelsen på vannliljefeltet. Hvis det tar 48 dager før feltet med vannliljer dekker hele innsjøen, hvor lang vil det ta for feltet å dekke 1/16-del av innsjøen? • ___ dager [Riktig svar: 46 dager]
Alle katter har pels. Kaniner har pels.

Hvis disse to påstandene stemmer, kan vi konkludere fra dem at kaniner er katter? • Ja • Nei [Riktig]
Alle blekkspruter liker Vitamin A. Wuzzier liker Vitamin A. Hvis disse to påstandene stemmer, kan vi konkludere fra dem at wuzziene er blekkspruter? • Ja • Nei [Riktig]
All tanter er søstre. Noen kvinner er tanter. Hvis disse to påstandene stemmer, kan vi konkludere fra dem at noen kvinner er søstre? • Ja [Riktig] • Nei
Alle bjørner er voldsomme. Noen kosedyr er bjørner. Hvis disse to påstandene stemmer, kan vi konkludere fra dem at noen kosedyr er voldsomme? • Ja [Riktig] • Nei
Hvis det tar 1 sykepleier 5 minutter å måle blodtrykket til 6 pasienter, hvor mange minutter vil det ta 100 sykepleier å måle blodtrykket til 300 pasienter? • ___ minutter [Riktig svar: 2,5 minutter]
Suppe og salat koster $5,01 til sammen. Suppen koster $1,03 mer enn salaten. Hvor mye koster salaten? • ___ dollar [Riktig svar: 1,99 dollar]
Sally lager te. Hver time tredobles konsentrasjonen av te. Hvis det tar seks timer å lage ferdig teen, hvor lang tid vil det ta før teens konsentrasjon er 1/9-del ferdig? • ___ timer [Riktig svar: 4 timer]
Alle pattedyr er sjenerte. Noen shidoer er pattedyr. Hvis disse to påstandene stemmer, kan vi konkludere fra dem at noen shidoer er sjenerte? • Ja [Riktig] • Nei
Alle koner er gifte. Noen kvinner er gifte. Hvis disse to påstandene stemmer, kan vi konkludere fra dem at noen kvinner er koner? • Ja • Nei [Riktig]

Alle hunder er svømmere. Noen reltaer er svømmere. Hvis disse to påstandene stemmer, kan vi konkludere fra dem at noen reltaer er hunder? • Ja • Nei [Riktig]
Alle fisker er svømmere. Noen olympiske atleter er svømmere. Hvis disse to påstandene stemmer, kan vi konkludere fra dem at noen olympiske atleter er fisker? • Ja • Nei [Riktig]

A.3 Tallforståelse

Kilde: Cokely, E. T., Galesic, M., Schulz, E. og Ghazal, S. (2012), 'Measuring Risk Literacy: The Berlin Numeracy Test', *Judgment and Decision Making*, 7:1, ss. 25–47.

Vennligst svar på spørsmålene under. Ikke bruk kalkulator, men du kan notere, f.eks. på papir.
Se for deg at vi kaster en femsidet terning 50 ganger. Basert på disse 50 kastene, hvor mange ganger i gjennomsnitt vil denne femsidete terningen vise et oddetall (1, 3 eller 5)? • ____ av 50 kast [Riktig svar: 30 kast]
Av 1000 personer i en liten by er 500 medlemmer av et kor. Av disse 500 kormedlemmene er 100 menn. Av de 500 innbyggerne som ikke er med i koret, er 300 menn. Hva er sannsynligheten for at en tilfeldig trukket mann er medlem av koret? Oppgi sannsynligheten i prosent: • ____ % [Riktig svar: 25 %]
Se for deg at vi kaster en vektet terning (6 sider). Sannsynligheten for at terningen viser en 6-er er dobbelt så høy som sannsynligheten for hver av de andre tallene. Basert på 70 kast, hvor mange ganger i gjennomsnitt vil terningen vise tallet 6? • _____ av 70 kast [Riktig svar: 20 kast]
I en skog er 20 % av soppene røde, 50 % brune og 30 % hvite. En rød sopp har 20 % sannsynlighet for å være giftig. En sopp som ikke er rød, har 5 % sannsynlighet for å være giftig. Hva er sannsynligheten for at en giftig sopp i skogen er rød? • _____ % [Riktig svar: 50 %]

A.4 Politisk kunnskap

Score basert på antall riktige (0–50). Deltagerne kunne svare «Vet ikke», som telte som feil.

Påstand	Svar
Internasjonal politikk og væpnede konflikter generelt	
1) I 2018 hadde USA og Russland mer enn 90 % av verdens atomvåpen.	Riktig
2) De faste medlemmene i FNs sikkerhetsråd er USA, Kina, Tyskland, Brasil og Russland.	Feil
3) NATOs intervensjon i Libya i 2011 hadde et mandat fra FNs sikkerhetsråd til å bruke militærmakt.	Riktig
4) Både Russland og USA har signert avtalen om ikke-spredning av atomvåpen (NPT).	Riktig
5) Nord-Korea gjennomførte sine fem første atomvåpentester mellom 1991 og 1996.	Feil
6) Kinesiske hangarskip har seilt i Middelhavet de tre siste årene.	Feil
7) Egypt er det mest folkerike av de arabiske landene.	Riktig
8) FN har i dag mellom 190 og 200 medlemsland.	Riktig
9) IAEA er en internasjonal organisasjon som arbeider for å sikre stabile oljemarkeder.	Feil
10) Siden konflikten i Øst-Ukraina begynte i 2014, har færre enn 30,000 mennesker blitt drept.	Riktig
Økonomi	
11) Det tok seks år før verdensøkonomien begynte å vokse igjen etter finanskrisen i 2008.	Feil
12) Det har vært tilnærmet ingen økonomisk vekst i India de siste 10 årene.	Feil
13) I perioden 2015-2019 var den gjennomsnittlige oljeprisen mellom $30 og $80.	Riktig
14) Blandingsøkonomi er et økonomisk system som først og fremst forbindes med sosialistiske land som Cuba og Nord-Korea.	Feil
15) De fem landene i BRICS-gruppen er Brasil, Russland, India, Kina og Sør-Afrika.	Riktig
16) I 2018 var Russlands bruttonasjonalprodukt (BNP) omtrent like stort som Tyskland og Storbritannias til sammen.	Feil
17) Kjøpekraftsparitet (KKP) er verdien av alle varer og tjenester som produseres i løpet av et år i et land.	Feil
18) I perioden 2014-2018 hadde Russland en gjennomsnittlig BNP-vekst på under 2 %.	Riktig
19) Mellom 2013 og 2018 økte Europas import av russisk naturgass.	Riktig
20) I 2018 sto USA for over 70 % av verdens samlede forsvarsutgifter.	Feil
Russland/Arktis	
21) Russlands befolkning er omtrent like stor som USAs.	Feil
22) I 2017 var forventet levealder i Russland over 65 år.	Riktig
23) Ved utgangen av 2020 vil Putin ha sittet flere år ved makten enn Stalin gjorde.	Feil
24) Det dominerende partiet i Russland heter Forent Russland.	Riktig

25) Borej/Dolgorukiy-klassen er en russisk, strategisk ubåt med interkontinentale, ballistiske missiler.	Riktig
26) I 2014 bombet russiske fly mål inne i Ukrainas hovedstad, Kiev.	Feil
27) I 2019 brøt Hviterussland alle diplomatiske bånd til Russland.	Feil
28) Russiske militære fartøy kan lovlig seile gjennom Norges økonomiske sone.	Riktig
29) I 2019 ble de endelige grensene i Arktis avgjort av FNs kontinentalsokkelkommisjon.	Feil
30) De fleste NATO-land har undertegnet Svalbardtraktaten.	Riktig
NATO/Europa	
31) Siden Sovjetunionen ble oppløst i 1991, har antallet medlemsland i NATO omtrent doblet seg.	Riktig
32) I 2018 brukte mer enn halvparten av NATO-landene over 2 % av BNP på forsvar.	Feil
33) NATOs Artikkel 5 forplikter alle medlemsland til å sende minst ett kompani som militær assistanse til det landet som blir angrepet.	Feil
34) Sverige, Finland og Danmark er alle EU-medlemmer.	Riktig
35) I 2017 var forventet levealder i EU over 75 år.	Riktig
36) Det døde færre migranter i Middelhavet i 2018 enn i 2016.	Riktig
37) Partene som inngikk atomavtalen med Iran i 2015 var USA, Storbritannia, Russland, Frankrike, Kina, Tyskland og EU.	Riktig
38) ETA er en separatistgruppe som kjemper for katalansk selvstendighet.	Feil
39) I 2014 stemte et lite flertall i Skottland for å forbli i Storbritannia.	Riktig
40) I 2019 la USA ned veto mot at NATO skulle definere verdensrommet (space) som et eget operasjonsdomene.	Feil
USA	
41) I 2019 hadde USA over 500,000 militært personell i Europa.	Feil
42) I det amerikanske presidentvalget i 2016 fikk Hillary Clinton flest stemmer.	Riktig
43) USA har i motsetning til Russland signert Minekonvensjonen som forbyr antipersonellminer.	Feil
44) USAs forsvarsminister heter Mark Esper/Christopher C. Miller.[346]	Riktig
45) Siden 2017 har ikke USA hatt en frihandelsavtale med Canada og Mexico.	Feil
46) Siden annen verdenskrig har det vært dobbelt så mange amerikanske presidenter fra Det republikanske partiet som fra Det demokratiske partiet.	Feil
47) USA har en militær base på Grønland.	Riktig
48) USA har flere hangarskip enn Russland og Kina til sammen.	Riktig
49) Omtrent halvparten av USAs befolkning er spansk-talende.	Feil
50) Etter at Trump ble president har Kina gått forbi USA som den største donoren til FNs samlede budsjett.	Feil

[346] Endret etter at M. Esper ble erstattet av C. C. Miller i nov. 2020.

A.5 Actively open-minded thinking

Kilde: Haran, U., Ritov, I. og Mellers, B. A. (2013), 'The role of actively open-minded thinking in information acquisition, accuracy, and calibration', *Judgment and Decision Making*, 8:3, ss. 188–201.

På en skala fra 1 (helt uenig) til 7 (helt enig), hvor uenig/enig er du i følgende påstander?

	1 - Helt uenig	2	3	4 - Hverken uenig eller enig	5	6	7 - Helt enig
Å la seg selv bli overbevist av et motargument er et godt karaktertrekk. (1)	☐	☐	☐	☐	☐	☐	☐
Folk burde ta i betraktning bevis som går imot deres egne overbevisninger. (2)	☐	☐	☐	☐	☐	☐	☐
Folk burde justere sine overbevisninger i lys av ny informasjon eller nye bevis. (3)	☐	☐	☐	☐	☐	☐	☐
Å endre din egen oppfatning er et tegn på svakhet. (4)	☐	☐	☐	☐	☐	☐	☐
Intuisjon er den beste guiden i beslutningstaking. (5)	☐	☐	☐	☐	☐	☐	☐
Det er viktig å opprettholde dine egne overbevisninger selv om det fremkommer bevis mot dem. (6)	☐	☐	☐	☐	☐	☐	☐
Man bør overse bevis som er i konflikt med ens egne etablerte overbevisninger. (7)	☐	☐	☐	☐	☐	☐	☐

A.6 Kognitiv lukking

Kilde: Roets, A. og Hiel, A. V. (2011), 'Item selection and validation of a brief, 15-item version of the Need for Closure Scale', *Personality and Individual Differences*, 50:1, ss. 90–94.

På en skala fra 1 (helt uenig) til 7 (helt enig), hvor uenig/enig er du i følgende påstander?

	1 - Helt uenig	2	3	4 - Hverken uenig eller enig	5	6	7 - Helt enig
Jeg liker ikke situasjoner som er usikre. (1)	☐	☐	☐	☐	☐	☐	☐
Jeg misliker spørsmål som kan besvares på mange ulike måter. (2)	☐	☐	☐	☐	☐	☐	☐
Jeg finner at et velordnet liv med vanlig arbeidstid passer meg best. (3)	☐	☐	☐	☐	☐	☐	☐
Jeg føler meg ukomfortabel når jeg ikke forstår årsaken til hvorfor en hendelse skjedde i livet mitt. (4)	☐	☐	☐	☐	☐	☐	☐
Jeg blir irritert når én person er uenig i det alle andre i en gruppe mener. (5)	☐	☐	☐	☐	☐	☐	☐
Jeg liker ikke å gå inn i en situasjon uten å vite hva jeg kan forvente i den. (6)	☐	☐	☐	☐	☐	☐	☐
Når jeg har tatt en beslutning føler jeg meg lettet. (7)	☐	☐	☐	☐	☐	☐	☐
Når jeg står overfor et problem har jeg et sterkt ønske om å finne en løsning svært raskt. (8)	☐	☐	☐	☐	☐	☐	☐
Jeg vil raskt bli utålmodig og irritert hvis ikke jeg finner en løsning på et problem umiddelbart. (9)	☐	☐	☐	☐	☐	☐	☐
Jeg liker ikke å være sammen med folk som er i stand til å gjøre uventede handlinger. (10)	☐	☐	☐	☐	☐	☐	☐
Jeg misliker det når en persons påstand kan bety mange ulike ting. (11)	☐	☐	☐	☐	☐	☐	☐
Jeg finner at det å etablere en fast rutine gjør at jeg setter mer pris på livet. (12)	☐	☐	☐	☐	☐	☐	☐
Jeg setter pris på en tydelig og strukturert levemåte. (13)	☐	☐	☐	☐	☐	☐	☐

På en skala fra 1 (helt uenig) til 7 (helt enig), hvor uenig/enig er du i følgende påstander?

	1 - Helt uenig	2	3	4 - Hverken uenig eller enig	5	6	7 - Helt enig
Jeg søker normalt ikke mange ulike synspunkter før jeg danner min egen oppfatning. (14)	□	□	□	□	□	□	□
Jeg misliker uforutsigbare situasjoner. (15)	□	□	□	□	□	□	□

A.7 Pinnsvin vs. revetenking

Kilde: Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E. og Tetlock, P. (2015), 'The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics', *Journal of Experiment Psychology: Applied*, 21:1, ss. 1106–1115.

På en skala fra 1 (klart mest rev) til 7 (klart mest pinnsvin), hvordan vil du beskrive deg selv?

	1 - Klart mest rev	2	3	4 - Hverken rev eller pinnsvin	5	6	7 - Klart mest pinnsvin
I et kjent essay delte Isaiah Berlin intellektuelle inn i pinnsvin og rever: «Pinnsvinet kan én stor ting og forsøker å forklare så mye som mulig ved å bruke denne teorien eller dette rammeverket. Reven kan mange små ting og er tilfreds med å finne forklaringer på stående fot fra sak til sak.» Når det kommer til å predikere, vil du beskrive deg selv som mest likt et pinnsvin eller en rev?	☐	☐	☐	☐	☐	☐	☐

A.8 Kognitiv motivasjon

Kilde: Cacioppo, J. T. og Petty, R. E. (1984), 'The Efficient Assessment of Need for Cognition', *Journal of Personality Assessment*, 48:3, ss. 306–307.

På en skala fra 1 til 7, vennligst oppgi i hvilken grad følgende påstander passer deg. Velg det alternativet som best beskriver det som er typisk for deg.

	1 - Passer svært dårlig	2 - Passer ganske dårlig	3 - Passer litt dårlig	4 - Passer hverken dårlig eller godt	5 - Passer litt godt	6 - Passer ganske godt	7 - Passer svært godt
Jeg foretrekker komplekse fremfor enkle problemer. (1)	☐	☐	☐	☐	☐	☐	☐
Jeg liker å ha ansvar for situasjoner som krever mye tenking. (2)	☐	☐	☐	☐	☐	☐	☐
Tankearbeid er ikke det jeg synes er mest gøy. (3)	☐	☐	☐	☐	☐	☐	☐
Jeg gjør heller noe som krever lite tankearbeid, fremfor noe som sikkert vil utfordre mine tenkeevner. (4)	☐	☐	☐	☐	☐	☐	☐
Jeg prøver å forutse og unngå situasjoner hvor det er en sjanse for at jeg må tenke grundig gjennom noe. (5)	☐	☐	☐	☐	☐	☐	☐
Jeg finner det tilfredsstillende å tenke grundig og lenge på ting. (6)	☐	☐	☐	☐	☐	☐	☐
Jeg tenker bare så grundig som jeg må. (7)	☐	☐	☐	☐	☐	☐	☐
Jeg foretrekker å tenke på små, daglige prosjekter framfor langsiktige prosjekter. (8)	☐	☐	☐	☐	☐	☐	☐
Jeg liker oppgaver som krever lite tankearbeid når jeg først har lært meg dem. (9)	☐	☐	☐	☐	☐	☐	☐

På en skala fra 1 til 7, vennligst oppgi i hvilken grad følgende påstander passer deg. Velg det alternativet som best beskriver det som er typisk for deg.

	1 - Passer svært dårlig	2 - Passer ganske dårlig	3 - Passer litt dårlig	4 - Passer hverken dårlig eller godt	5 - Passer litt godt	6 - Passer ganske godt	7 - Passer svært godt
Ideen om å bruke tankene mine til å komme meg til topps, appellerer til meg. (10)	☐	☐	☐	☐	☐	☐	☐
Jeg setter stor pris på oppgaver som involverer å finne nye løsninger på problemer. (11)	☐	☐	☐	☐	☐	☐	☐
Å lære nye måter å tenke på, engasjerer meg ikke i særlig stor grad. (12)	☐	☐	☐	☐	☐	☐	☐
Jeg foretrekker at livet mitt er fylt med "puzzles" som jeg må løse. (13)	☐	☐	☐	☐	☐	☐	☐
Abstrakt tenking appellerer til meg. (14)	☐	☐	☐	☐	☐	☐	☐
Jeg ville foretrukket en oppgave som er intellektuell, vanskelig og viktig, fremfor en som er noe viktig, men ikke krever mye tankearbeid. (15)	☐	☐	☐	☐	☐	☐	☐
Jeg føler lettelse heller enn tilfredsstillelse etter at jeg har løst en oppgave som krever mye mental innsats. (16)	☐	☐	☐	☐	☐	☐	☐
For meg er det nok at noe får jobben gjort; jeg bryr meg ikke om hvordan eller hvorfor det virker. (17)	☐	☐	☐	☐	☐	☐	☐
Jeg ender som regel opp med å tenke grundig gjennom ting, selv om de ikke angår meg personlig. (18)	☐	☐	☐	☐	☐	☐	☐

A.9 Motivasjoner for å delta

Tenk over din motivasjon for å delta i FFIs prediksjonsturnering. På en skala fra 1 (helt uenig) til 7 (helt enig), hvor uenig/enig er du i følgende påstander?

	1 - Helt uenig	2	3	4 - Hverken uenig eller enig	5	6	7 - Helt enig
Jeg synes spørsmålene er interessante. (1)	☐	☐	☐	☐	☐	☐	☐
Jeg ønsker å se hvor godt jeg klarer å predikere for min egen del. (2)	☐	☐	☐	☐	☐	☐	☐
Jeg ønsker å se hvor godt jeg klarer å predikere sammenlignet med andre deltagere. (3)	☐	☐	☐	☐	☐	☐	☐
Jeg ønsker å havne blant de beste deltagerne. (4)	☐	☐	☐	☐	☐	☐	☐
Jeg tror svarene mine kan være nyttige for andre. (5)	☐	☐	☐	☐	☐	☐	☐
Jeg føler en plikt til å delta. (6)	☐	☐	☐	☐	☐	☐	☐

A.10 Forståelse av scoringssystemet

Scoren er basert på antall riktige svar (0–5). Riktige svar er krysset av i boksen under. Alternativene nærmest det riktige svaret ble altså ikke vektet. «Vet ikke» telte som feil.

I FFIs prediksjonsturnering måles treffsikkerheten ved hjelp av et system som heter Brier-score. Vi bruker en skala fra 0 til 2, der målet er å få lavest mulig score. En score på 0 betyr «helt riktig» (100 % på riktig svar), mens 2 betyr «helt galt» (100 % på galt svar). Tenk deg at du får spørsmålet: Vil Donald Trump vinne presidentvalget i 2020? Du vil nå bli bedt om å anslå Brier-score din ved forskjellige sannsynlighetsfordelinger og utfall.

1) Du er usikker, og fordeler sannsynligheten helt likt: Ja (50 %). Trump vinner, så riktig svar ble: Ja (50 %). På en skala fra 0 til 2, hvilken Brier-score får du?

☐ 0,25	☒ **0,5**	☐ 1	☐ 1,5	☐ 1,75	☐ Vet ikke

2) Du svarer det samme som over: Ja (50 %). Denne gangen taper imidlertid Trump, så Ja ble feil svar. På en skala fra 0 til 2, hvilken Brier-score får du?

☐ 0,25	☒ **0,5**	☐ 1	☐ 1,5	☐ 1,75	☐ Vet ikke

3) Du har større tro på at Trump vinner, og svarer: Ja (80 %). Trump vinner, så Ja ble riktig svar. Siden du traff så godt, lurer du på hvor mye bedre score du kunne fått, hvis du i stedet hadde svart: Ja (95 %). Hvor mye bedre score tror du at du ville fått om du hadde økt fra 80 % til 95 % på riktig svar?

☒ **Rundt 0,1 bedre Brier-score**	☐ Rundt 0,3 bedre Brier-score	☐ Rundt 0,5 bedre Brier-score	☐ Vet ikke

4) La oss si at situasjonen var omvendt. Du var like sikker på at Trump skulle vinne som over: Ja (80 %). Denne gangen taper imidlertid Trump, så Ja ble feil svar. Nå lurer du på hvor mye dårligere score du kunne fått, hvis du i stedet hadde svart: Ja (95 %). Hvor mye dårligere score tror du at du ville fått om du hadde økt fra 80 % til 95 % på feil svar?

☐ Rundt 0,1 bedre Brier-score	☐ Rundt 0,3 bedre Brier-score	☒ **Rundt 0,5 bedre Brier-score**	☐ Vet ikke

5) For å beregne dine sammenlagte resultater, tar vi utgangspunkt i Brier-scoren din på alle spørsmål. Den sammenlagte scoren kaller vi Accuracy score, men hvordan beregnes denne i FFIs prediksjonsturnering?

☐ Snittet av alle Brier-scorene jeg har fått	☐ Summen av alle Brier-scorene jeg har fått	☐ Snittet av differansene mellom min Brier-score og medianen av deltagernes Brier-score på hvert spørsmål	☒ **Summen av differansene mellom min Brier-score og medianen av deltagernes Brier-score på hvert spørsmål**	☐ Vet ikke

A.11 Prediksjonsspesifikke tenkemåter

Du får nå en liste med ulike måter å tenke på når en skal predikere. Kryss av for de måtene som er mest dekkende for hvordan du tenkte, når du svarte på de månedlige spørsmålsrundene. Du kan krysse av flere alternativer.

- ☐ Baserte meg på mageføIelsen min.
- ☐ Tok utgangspunkt i en teori eller generell oppfatning jeg hadde av fenomenet fra før, og brukte denne til å vurdere hva som ville skje i dette tilfellet.
- ☐ Tok utgangspunkt i det aktuelle spørsmålet, og tenkte gjennom hva ulike teorier ville sagt om hva som ville skje.
- ☐ Lette etter informasjon fra flere forskjellige kilder.
- ☐ Baserte meg på det første som slo meg som mest sannsynlig.
- ☐ Baserte meg på hvordan utfallene har fordelt seg ved tilfeller av samme type hendelse.
- ☐ Baserte meg på tekstene og ev. figurer som fulgte med spørsmålet, og vurderte utfallene ut fra dette.
- ☐ Estimerte hvor ofte hendelsen eller utfallet hadde skjedd før.
- ☐ Baserte meg på snittet av flere, forskjellige estimater av utfallet.
- ☐ Baserte meg på et lignende, historisk tilfelle som jeg kjente utfallet av.
- ☐ Baserte meg på flere lignende, historiske tilfeller med forskjellige utfall.
- ☐ Tok utgangspunkt i dagens situasjon/nivå, og justerte min prediksjon deretter.
- ☐ Baserte meg på den siste utviklingen som hadde skjedd i saken, da spørsmålet ble stilt.
- ☐ Fordelte prosentene slik at jeg fikk best mulig score hvis jeg traff, men samtidig unngikk å få en veldig dårlig score hvis jeg bommet.
- ☐ Baserte meg på en fremskrivning av den samme utviklingen som frem til nå.
- ☐ Tenkte på hva som gjorde at jeg bommet/traff på tidligere spørsmål.
- ☐ Tok hensyn til uforutsigbare, overraskende hendelser som kunne påvirke utfallet.
- ☐ Annet.

B Analyse av foreløpig datasett

Dette vedlegget gir en deskriptiv analyse av datagrunnlaget som de foreløpige resultatene beskrevet i kapittel 5 er basert på. Til forskjell fra kapittel 3 og 4 er dette datagrunnlaget basert på de første 150 spørsmålene som er avgjort så langt og de 833 deltagerne som har svart på minst 20 % av disse. Dette er imidlertid stort sett de samme deltagerne som de 857 analysert tidligere.

For å skille beskrivelsen av det foreløpige datagrunnlaget basert på de 150 første spørsmålene fra den komplette analysen basert på alle 240 spørsmålene, omtales det foreløpige datasettet her som «FFI150», mens verdiene fra GJP200 og GJP350 er de samme som i foregående kapitler.

B.1 Tema

Tabell B.1 viser at forskjellene i den tematiske fordelingen av spørsmålene i FFIs og GJPs turneringer er omtrent helt lik i det foreløpige datagrunnlaget som i det komplette (se underkapittel 3.1.2). I FFIs turnering ble det stilt en høyere andel spørsmål om krig, konflikt og militære operasjoner, Russland, USA og norsk innen- og utenrikspolitikk, mens det i GJP var langt flere spørsmål om spesielt Midtøsten og Nord-Afrika og Øst-Asia.

Kategori	**FFI150**	**GJP200**	**GJP350**
Internasjonal politikk generelt	10 (6,7 %)	9 (4,3 %)	20 (5,8 %)
Krig, konflikt og militære operasjoner	**52 (34,7 %)**	**40 (19,0 %)**	**68 (19,5 %)**
Teknologi	17 (11,3 %)	8 (3,8 %)	18 (5,2 %)
Økonomi	24 (16 %)	**35 (16,6 %)**	**64 (18,4 %)**
Terrorisme	12 (8 %)	0 (0 %)	0 (0%)
Norsk innenrikspolitikk	**27 (18 %)**	0 (0 %)	0 (0%)
Norsk utenrikspolitikk	15 (10 %)	0 (0 %)	0 (0%)
Russland	**28 (18,7 %)**	5 (2,4 %)	16 (4,6 %)
Norden	6 (4 %)	0 (0 %)	0 (0 %)
Europa	**33 (22 %)**	**44 (20,9 %)**	**66 (19,0 %)**
NATO	8 (5,3 %)	1 (0,5 %)	1 (0,3 %)
USA	**29 (19,3 %)**	8 (3,8 %)	20 (5,8 %)
Mellom- og Sør-Amerika	2 (1,3 %)	11 (5,2 %)	24 (6,9 %)
Midtøsten og Nord-Afrika	23 (15,3 %)	**81 (38,4 %)**	**115 (33,1 %)**
Afrika sør for Sahara	1 (0,7 %)	24 (11,4 %)	32 (9,2 %)
Øst-Asia	10 (6,7 %)	**34 (16,1 %)**	**64 (18,4 %)**
Sentral-Asia	0 (0 %)	0 (0 %)	1 (0,3 %)
Sør- og Sørøst-Asia	3 (2 %)	15 (7,1 %)	32 (9,2 %)
Annet (ingen politisk relevans)	0 (0 %)	2 (0,9 %)	3 (0,9 %)

Tabell B.1 Antall og andel spørsmål per tema.

B.2 Type spørsmål

Tabell B.2 viser at rundt 70 % av de avgjorte spørsmålene i FFIs foreløpige datagrunnlag er kategoriske eller ordinale, mot rundt 80 % binære i GJPs. Denne forskjellen er omtrent like stor som ved sammenligninger med det komplette datasettet (se underkapittel 3.1.3). Selv uten den siste tredelen av spørsmålene i FFIs turnering som vil avgjøres senere, består det foreløpige datagrunnlaget også av et høyere antall kategoriske og ordinale spørsmål enn i begge GJP-studiene.

Spørsmålstype	**FFI150**	**GJP200**	**GJP350**
Binære, betingede binære	43 (28,7 %)	148 (70,1 %), 32 (15,2 %)	206 (59,4 %), 72 (20,7 %)
Kategoriske	33 (22,0 %)	15 (7,1 %)	24 (6,9 %)
Ordinale	74 (49,3 %)	16 (7,6 %)	45 (13,0 %)

Tabell B.2 Antall og andel spørsmål per type.

Tabell B.3 viser at fordelingen av avgjorte spørsmål på forskjellige antall svaralternativer også er omtrent lik FFIs komplette datasettet. Det gjennomsnittlige antallet svaralternativer per spørsmål i det foreløpige datagrunnlaget er også like høyt (3,5) og fortsatt høyere enn i GJP (2,3).

Antall svaralternativer	**FFI150**	**GJP200**	**GJP350**
2	43 (28,7 %)	180 (85,3 %)	278 (80,1 %)
3	26 (17,3 %)	10 (4,7 %)	28 (8,1 %)
4	58 (38,7 %)	16 (7,6 %)	31 (8,9 %)
5	21 (14,0 %)	5 (2,4 %)	10 (2,9 %)
6	0	0	0
7	0	0	0
8	1 (0,7 %)	0	0
9	1 (0,7 %)	0	0

Tabell B.3 Antall spørsmål per antall svaralternativer i FFIs turnering.

B.3 Tidsperspektiv

Det gjennomsnittlige tidsperspektivet på de 150 avgjorte spørsmålene i FFIs turnering er 372 dager, som er en betydelig kortere enn snittet på 521 dager i det komplette datasettet. Forklaringen er spørsmålene med lengst tidsperspektiv ennå ikke er avgjort.

Tabell B.4 viser fordelingen av FFIs og GJPs spørsmål på tidsperspektivsintervaller på seks måneder hver. Siden datagrunnlaget her bare inkluderer spørsmål som så langt er avgjort i FFIs turnering, er andelen spørsmål med de korteste tidsperspektivene på 0–18 måneder høyere (85 %) enn i det fullstendige datasettet (69 %) (se underkapittel 3.1.4).

Tidsperspektiv	**FFI150**	**GJP200**	**GJP350**
Opptil 6 måneder (under 183 dager)	37 (24,7 %)	156 (73,9 %)	266 (76,7 %)
6–12 måneder (183–365 dager)	47 (31,3 %)	46 (21,8 %)	70 (20,2 %)
12–18 måneder (366–548 dager)	43 (28,7 %)	8 (3,8 %)	9 (2,6 %)
18–24 måneder (549–731 dager)	9 (6,0 %)	1 (0,5 %)	2 (0,6 %)
24–30 måneder (732–914 dager)	7 (4,7 %)	0 (0 %)	0 (0 %)
30–36 måneder (915–1097 dager)	4 (2,7 %)	0 (0 %)	0 (0 %)
Over 36 måneder (over 1097 dager)	3 (2,0 %)	0 (0 %)	0 (0 %)

Tabell B.4 Antall og andel spørsmål per tidsperspektiv.

B.4 Kjønn, alder og utdanning

Tabell B.5 viser antallet og andelen deltagere med forskjellige utdanningsnivåer i FFIs turnering, GJP og EPJ. Her er andelene deltagere på hvert utdanningsnivå basert på de foreløpige resultatene helt like andelene basert på det komplette datagrunnlaget (se underkapittel 3.2.2).

	FFI150	**GJP350**[347]	**EPJ140**[348]
Kjønn	89 % menn	83 % menn	76 % menn
Alder (snitt)	40 år[349]	40 år	43 år
Ingen høyere utdanning	72 (9 %)	12 (1 %)	-
1–3 års høyere utdanning	192 (23 %)	544 (31 %)	11 (4 %)
4–5 års høyere utdanning	293 (35 %)	710 (41 %)	124 (44 %)
Over 5 års høyere utdanning	276 (33 %)	483 (28 %)	148 (52 %)

Tabell B.5 Deltagernes kjønn, alder og utdanning.

[347] Kjønn og alder er basert på Mellers 2015, Superforecasters, s. 268. Kjønnsfordelingen oppgitt i artikkelen samsvarer replikasjonsdatasettet (84 % menn). Replikasjonsdatasettet inneholder imidlertid ikke nøyaktig alder, bare aldersgrupper. Aldersgruppene med flest deltagere er imidlertid 25–29 (404), 30–34 (388) og 35–39 år (222), som tilsier et lavere snitt enn 40 år, som er oppgitt i artikkelen. Andelene per utdanningsnivå er bare basert replikasjonsdatasettet til GJP350, der det finnes data om høyeste utdanning på 1748 av deltagerne. I artikkelen er det bare oppgitt at 64 % hadde utdanning på mastergradsnivå, som ligger nært de 68 % som her er kategorisert med minst 4–5 års utdanning.
[348] Her er antallet eksperter per utdanningsnivå beregnet ut fra 284 eksperter og andelene som er oppgitt i Tetlock (2005), *Expert Political Judgment*, s. 239ff. Her opplyses det at 96 % hadde masternivå og 52 % doktorgrad. De resterende 4 % som faller utenfor, er kategorisert innenfor bachelornivå, siden det antas at alle ekspertene hadde minst ett års høyere utdanning. Tetlock (2005), *Expert Political Judgment*, s. 239ff. Oppsummert i Tetlock og Gardner (2015), *Superforecasting*, ss. 66–67.
[349] Alder er basert på 2017, som var tidspunktet for oppstart av turneringen og da de fleste registrerte seg for å delta.

B.5 Ekspertise

Tabell B.6 viser at deltagerne i FFIs foreløpige datasett skiller seg svært lite fra deltagerne i det komplette datagrunnlaget på noen av ekspertisekriteriene.

	FFI150	**EPJ**[350]
Bransje/ sektor (andel av deltagere)	408 (49 %) forsvarssektoren, hvorav de fleste var: - 170 (20 %) offiserer - 124 (15 %) forskere - 37 (4 %) befal/grenader/konstabel 100 (12 %) faglige, vitenskap. og tekn. tjenester 63 (8 %) offentlig administrasjon 51 (6 %) informasjon og kommunikasjon 44 (5 %) ikke yrkesaktiv/pensjonert 166 (20 %) fordelt på resterende (opptil 3 % hver)	116 (41 %) akademia 74 (26 %) staten 48 (17 %) tenketanker og stiftelser 23 (8 %) internasjonale organisasjoner 23 (8 %) privat sektor (inkludert media)
Eksperter (andel av deltagere)	270 (32 %) svarte at de arbeidet/hadde arbeidet med forsvars- og sikkerhetspolitikk som en del av jobben (10 års arbeidserfaring i gjennomsnitt)	Alle 284 var profesjonelle eksperter (12 års relevant erfaring i gjennomsnitt)
Ekspertise-områder (andel av eksperter)	138 (51 %) krig, konflikt og militære operasjoner 92 (34 %) internasjonal politikk 81 (30 %) NATO 75 (28 %) teknologi 73 (27 %) Russland 58 (22 %) USA 51 (19 %) norsk utenrikspolitikk 50 (19 %) Midtøsten og Nord-Afrika 50 (19 %) terrorisme 48 (18 %) Europa 45 (17 %) Norden 44 (16 %) økonomi 27 (10 %) norsk innenrikspolitikk Alle resterende regioner (opptil 25/9 % hver)	116 (41 %) områdestudier 68 (24 %) internasjonal relasjoner 34 (12 %) økonomi 31 (11 %) nasjonal sikkerhet og rustningskontroll 26 (9 %) journalistikk 6 (2 %) diplomati 3 (1 %) internasjonal rett
Intervjuet i media (andel av eksperter)	68 (25 %) med relevant arbeidserfaring oppgav at de hadde blitt intervjuet som eksperter i media om forsvars- og sikkerhetspolitiske spørsmål 49 (18 %) sitert eller omtalt minst 10 ganger	173 (61 %) intervjuet av minst ett stort medium 60 (21 %) intervjuet minst 10 ganger

Tabell B.6 Deltagernes yrkesbakgrunn og ekspertise.

[350] Tetlock (2005), *Expert Political Judgement*, s. 239ff. Oppsummert i Tetlock og Gardner (2015), *Superforecasting*, ss. 66–67. Her er antallene beregnet ut fra de prosentvise andelene som er oppgitt for 284 eksperter.

B.6 Disposisjonelle variabler og innsatsvariabler

Tabell B.7 gir en deskriptiv analyse av alle variablene som er brukt i de bivariate analysene i delkapittel 5.3. Det er kun snittene fra tabellen under som er gjengitt i tabell 5.8, der disse sammenlignes med scorene fra GJP.

	Snitt	**SD**	**Min**	**Max**	**α**	**n**
Shipley-2 Block Patterns (0–26)	17,66	4,98	1	26	0,88	376
CRT orig. (0–3)	2,45	0,80	0	3	0,5	395
CRT utv. (0–18)	14,98	2,99	4	18	0,8	395
Tallforst. (0–4)	2,74	1,22	0	4	0,6	395
Politisk kunnsk. (0–50)	35,41	6,62	8	50	0,81	526
Aktiv fordomsfri tenkning (1–7)	6,15	0,58	3	7	0,68	477
Kognitiv lukking (1–7)	3,88	0,81	1,6	6,2	0,82	477
Rev–pinnsvin påstand (1–7)	2,79	1,44	1	7		477
Motivasjon – være blant de beste (1–7)	4,98	1,55	1	7		477
Kognitiv motivasjon (1–7)	5,20	0,70	2,72	6,72	0,84	332
Antall unike estimater (første uke)	32,99	18,61	2	114		833
Brier forståelse (0–5)	0,90	1,21	0	5	0,63	477
Antall spørsmål besvart	146,18	61,86	30	240		833
Tid per spørsmål (minutter)	1,38	0,58	0,33	4,98		822

Tabell B.7 Deskriptiv analyse av uavhengige variabler i FFIs foreløpige datagrunnlag.

B.7 Korrelasjoner med individuelle egenskaper

I kapittel 5 er analysen av korrelasjoner mellom treffsikkerheten og de individuelle egenskapene basert på 833 deltagere med en score på minst én av disposisjonell eller innsatsrelatert variabel. For å undersøke robustheten til denne analysen er de samme analysene blitt gjennomført basert på bare de 199 deltagerne som er registrert med en score på alle variabler.

Tabell B.8 gir en deskriptiv analyse av snittscorene til de 199 deltagerne som oppfylte dette strengeste inklusjonskriteriet, tilsvarende tabellen basert på alle 833 deltagerne (se tabell B.7). Snittscorene til de to deltagerutvalgene er svært like. Unntakene er at de 199 deltagerne i snitt svarte på flere spørsmål (192 mot 146) og brukte flere unike sannsynlighetsestimater (40 mot 33). Ellers er scorene på kognitive evner, kunnskapsnivå og tenkemåter tilnærmet identiske, og internkonsistensen på samme nivå.

	Snitt	**SD**	**Min**	**Max**	**α**	**n**
Std. Brier-score	–0,06	0,23	–0,44	1,07	0,90	199
Shipley-2 Block Patterns (0–26)	18,18	4,96	1,00	26,00	0,88	199
CRT orig. (0–3)	2,55	0,74	0,00	3,00	0,49	199
CRT utv. (0–18)	15,18	3,09	4,00	18,00	0,82	199
Tallforst. (0–4)	2,82	1,25	0,00	4,00	0,66	199
Politisk kunnsk. (0–50)	35,21	7,04	8,00	48,00	0,83	199
Aktiv fordomsfri tenkning (1–7)	6,16	0,61	3,86	7,00	0,71	199
Kognitiv lukking (1–7)	3,98	0,84	2,07	6,20	0,83	199
Rev–pinnsvin påstand (1–7)	2,67	1,35	1,00	7,00		199
Motivasjon – være blant de beste (1–7)	5,00	1,55	1,00	7,00		199
Kognitiv motivasjon (1–7)	5,18	0,73	2,72	6,67	0,85	199
Antall unike estimater (første uke)	40,33	22,26	11,00	114,00		199
Brier forståelse (0–5)	1,09	1,40	0,00	5,00	0,72	199
Antall spørsmål besvart	192,09	42,64	57,00	240,00		199
Tid per spørsmål (minutter)	1,54	0,65	0,50	4,12		199

Tabell B.8 Deskriptiv analyse av uavhengige variabler i FFIs foreløpige datagrunnlag, basert kun på de 199 deltagerne med scores registrert på alle variabler.

Tabell B.9 viser alle korrelasjoner mellom deltagernes gjennomsnittlige standardiserte Brier-scores og hver uavhengige variabel, basert på deltagerutvalgene på hhv. 833 og 199 deltagere og ved bruk av både Pearsons *r* og Spearmans r_s som korrelasjonsmål. Verdiene oppgitt i den første kolonnene med resultater er de samme som er oppsummert i tabell 5.10.

Ved alle 833 deltagere forblir de fleste sammenhengene signifikante på 0.0001-nivå ved bruk av Spearmans r_s, som ikke forutsetter en bestemt fordeling av verdiene i motsetning til Pearsons *r*. Unntakene er aktiv fordomsfri tenkning, der p-verdien øker fra under 0.001 til under 0.01, og Brier-score forståelse, der p-verdien øker fra under 0.01 til under 0.05, mens korrelasjonen med kognitiv motivasjon nå blir signifikant ved at p-verdien faller fra 0.068 til under 0.01.

Det viktigste er at nesten alle uavhengige variabler som korrelerte med treffsikkerheten ved det største deltagerutvalget med alle 833 deltagere fortsatt korrelerer på minst 0.01-nivå, uansett korrelasjonsmål, når utvalget avgrenses til kun 199 deltagere. Unntakene er den originale CRT-testen, som i utgangspunktet hadde lav pålitelighet, og aktiv fordomsfri tenkning, der p-verdien nå faller utenfor akseptabelt nivå, og antall spørsmål besvart, der det nå ikke lenger er noen signifikant korrelasjon med treffsikkerheten.

		FFI (833 deltagere)		**FFI (199 deltagere)**	
		Pearson	**Spearman**	**Pearson**	**Spearman**
Kognitive evner	Shipley-2 Block Patterns (0–26)	$r = –0.07$, $t(374) = –1.29$, $p = 0.197$	$r_s = –0.08$, $p = 0.103$	$r = –0.05$, $t(197) = –0.65$, $p = 0.515$	$r_s = –0.05$, $p = 0.515$
	CRT original (0–3)	**$r = –0.18$, $t(393) = –3.66$, $p < 0.001$**	**$r_s = –0.19$, $p < 0.001$**	$r = –0.16$, $t(197) = –2.27$, $p < 0.05$	$r_s = –0.16$, $p < 0.05$
	CRT utvidet (0–18)	**$r = –0.23$, $t(393) = –4.66$, $p < 0.0001$**	**$r_s = –0.25$, $p < 0.0001$**	**$r = –0.23$, $t(197) = –3.29$, $p < 0.01$**	**$r_s = –0.25$, $p < 0.001$**
	Tallforståelse (0–4)	**$r = –0.21$, $t(393) = –4.33$, $p < 0.0001$**	**$r_s = –0.23$, $p < 0.0001$**	**$r = –0.21$, $t(197) = –3.06$, $p < 0.01$**	**$r_s = –0.21$, $p < 0.01$**
Kunnskapsnivå	Politisk kunnskapsnivå (0–50)	**$r = –0.20$, $t(524) = –4.58$, $p < 0.0001$**	**$r_s = –0.21$, $p < 0.0001$**	**$r = –0.21$, $t(197) = –2.99$, $p < 0.01$**	**$r_s = –0.24$, $p < 0.001$**

Tenkemåter	Aktiv fordomsfri tenkning (1–7)	**$r = -0.17$, $t(475) = -3.86$, $p < 0.001$**	**$r_s = -0.15$, $p < 0.01$**	$r = -0.18$, $t(197) = -2.53$, $p < 0.05$	$r_s = -0.9$, $p = 0.19$
	Kognitiv lukking (1–7)	$r = -0.03$, $t(475) = -0.76$, $p = 0.45$	$r_s = 0.01$, $p = 0.781$	$r = -0.03$, $t(197) = -0.4$, $p = 0.692$	$r_s = 0.01$, $p = 0.911$
	Rev vs. pinnsvin – enkeltpåstand (1–7)	$r = -0.07$, $t(475) = -1.55$, $p = 0.123$	$r_s = -0.07$, $p = 0.106$	$r = -0.05$, $t(197) = -0.68$, $p = 0.499$	$r_s = -0.04$, $p = 0.597$
	Motivasjon – være blant de beste (1–7)	**$r = -0.20$, $t(475) = -4.51$, $p < 0.0001$**	**$r_s = -0.25$, $p < 0.0001$**	**$r = -0.2$, $t(197) = -2.88$, $p < 0.01$**	**$r_s = -0.26$, $p < 0.001$**
	Kognitiv motivasjon (1–7)	$r = -0.1$, $t(330) = -1.83$, $p = 0.068$	**$r_s = -0.14$, $p < 0.01$**	$r = -0.1$, $t(197) = -1.48$, $p = 0.142$	$r_s = -0.15$, $p < 0.05$
Oppgavespes. ferdigheter	Antall unike estimater (bare første uken)	**$r = -0.34$, $t(831) = -10.54$, $p < 0.0001$**	**$r_s = -0.38$, $p < 0.0001$**	**$r = -0.29$, $t(197) = -4.22$, $p < 0.0001$**	**$r_s = -0.31$, $p < 0.0001$**
	Brier-score forståelse (0–5)	**$r = -0.15$, $t(475) = -3.36$, $p < 0.001$**	$r_s = -0.11$, $p < 0.05$	**$r = -0.19$, $t(197) = -2.79$, $p < 0.01$**	**$r_s = -0.19$, $p < 0.01$**
Innsats	Antall spørsmål besvart	**$r = -0.20$, $t(831) = -6.02$, $p < 0.0001$**	**$r_s = -0.15$, $p < 0.0001$**	$r = -0.02$, $t(197) = -0.3$, $p = 0.764$	$r_s = -0.07$, $p = 0.313$
	Tid brukt per spørsmål	**$r = -0.35$, $t(820) = -10.73$, $p < 0.0001$**	**$r_s = -0.36$, $p < 0.0001$**	**$r = -0.39$, $t(197) = -6.02$, $p < 0.0001$**	**$r_s = -0,45$, $p < 0.0001$**

Tabell B.9 Korrelasjoner mellom deltagernes standardiserte Brier-scores og individuelle egenskaper, basert på ulike utvalg. Fet skrift betyr signifikant på 0.01-nivå.

B.8 Prediksjonsspesifikke tenkemåter

Tabell B.10 viser antall og andel deltagere som krysset av for hver av de prediksjonsspesifikke tenkemåtene. Til forskjell fra tabell 4.2 i kapittel 4, som var basert på 381 deltagere som har svart på minst 20 % av alle 240 spørsmål som har blitt stilt i turneringen som helhet, er denne basert på de 348 deltagerne som har svart på minst 20 % av de 150 spørsmålene som er avgjort så langt. Andelene som brukte de forskjellige tenkemåtene er imidlertid tilnærmet helt lik.

	Antall	**Andel**
Baserte meg på mageføelelsen min. (intuisjon)	233	67,0 %
Tok utgangspunkt i en teori eller generell oppfatning jeg hadde av fenomenet fra før, og brukte denne til å vurdere hva som ville skje i dette tilfellet. (deduktiv resonnering)	164	47,1 %
Tok utgangspunkt i det aktuelle spørsmålet, og tenkte gjennom hva ulike teorier ville sagt om hva som ville skje. (induktiv resonnering)	74	21,3 %
Lette etter informasjon fra flere forskjellige kilder. (aktiv fordomsfri tenkning)	45	12,9 %
Baserte meg på det første som slo meg som mest sannsynlig. (kognitiv lukking)	133	38,2 %
Baserte meg på hvordan utfallene har fordelt seg ved tilfeller av samme type hendelse. (referanseklasser)	141	40,5 %
Baserte meg på tekstene og ev. figurer som fulgte med spørsmålet, og vurderte utfallene ut fra dette. (ankring)	229	65,8 %
Estimerte hvor ofte hendelsen eller utfallet hadde skjedd før. (grunnfrekvens)	142	40,8 %
Baserte meg på snittet av flere, forskjellige estimater av utfallet. (wisdom of the crowd)	21	6,0 %
Baserte meg på et lignende, historisk tilfelle som jeg kjente utfallet av. (bruk av én historisk analogi)	82	23,6 %
Baserte meg på flere lignende, historiske tilfeller med forskjellige utfall. (bruk av flere historiske analogier)	71	20,4 %
Tok utgangspunkt i dagens situasjon/nivå, og justerte min prediksjon deretter. (ankring)	227	65,2 %
Baserte meg på den siste utviklingen som hadde skjedd i saken, da spørsmålet ble stilt. (tilgjengelighetsheuristikk)	108	31,0 %
Fordelte prosentene slik at jeg fikk best mulig score hvis jeg traff, men samtidig unngikk å få en veldig dårlig score hvis jeg bommet. (optimalisering av Brier-score)	92	26,4 %
Baserte meg på en fremskrivning av den samme utviklingen som frem til nå. (ekstrapolasjon)	114	32,8 %

Tenkte på hva som gjorde at jeg bommet/traff på tidligere spørsmål. (post-mortem analyse)	41	11,8 %
Tok hensyn til uforutsigbare, overraskende hendelser som kunne påvirke utfallet. (sorte svaner)	61	17,5 %
Annet.	12	3,4 %

Tabell B.10 Antall og andel deltagere som brukte ulike prediksjonsspesifikke tenkemåter.

B.9 Norske superforecastere vs. resten

Tabell B.11 sammenligner verdiene til superforecasterne, de nest beste og alle andre deltagere på alle variablene som ble brukt i de bivariate analysene. Her oppgis den gjennomsnittlige scoren, standardavviket (SD), maksimums- og minimumsverdiene og antallet deltagere (n) som verdiene er baserte på.

	Superforecastere						Nest beste deltagere						Alle andre deltagere					
	Snitt	**SD**	**Min**	**Max**	**α**	**n**	**Snitt**	**SD**	**Min**	**Max**	**α**	**n**	**Snitt**	**SD**	**Min**	**Max**	**α**	**n**
Shipley-2 Block Patterns (0–26)	19,19	4,74	9,00	26,00	0,89	36	19,10	4,64	7	26	0,88	39	17,30	5,00	1	26	0,88	301
CRT orig. (0–3)	2,84	0,37	2,00	3,00	0,09	43	2,65	0,72	0	3	0,59	37	2,37	0,84	0	3	0,50	315
CRT utv. (0–18)	16,42	2,20	8,00	18,00	0,77	43	16,03	2,46	10	18	0,78	37	14,66	3,07	4	18	0,79	315
Tallforst. (0–4)	3,37	0,85	1,00	4,00	0,52	43	3,24	0,89	1	4	0,42	37	2,59	1,25	0	4	0,61	315
Politisk kunnskap - FFI (0–50)	38,27	5,82	20,00	48,00	0,80	51	37,24	5,17	14	45	0,73	45	34,88	6,73	8	50	0,81	430
Aktiv fordoms-fri tenkning (1–7)	6,32	0,51	4,86	7,00	0,72	46	6,12	0,51	4	7	0,5	39	6,13	0,59	3	7	0,69	392
Kognitiv lukking (1–7)	3,77	0,80	2,13	5,13	0,83	46	3,89	0,76	2,4	5,93	0,83	39	3,89	0,81	1,60	6,20	0,82	392
Rev–pinnsvin – påstand (1–7)	2,63	1,18	1,00	6,00		46	2,56	1,35	1	6		39	2,83	1,48	1	7		392
Motivasjon – være blant de beste (1–7)	5,78	1,07	2,00	7,00		46	5,26	1,52	1	7		39	4,86	1,58	1	7		392
Kognitiv motivasjon (1–7)	5,43	0,57	0,57	6,56	0,76	46	5,35	0,66	3	6,33	0,82	30	5,14	0,71	2,72	6,72	0,85	256

	Superforecastere						Nest beste deltagere						Alle andre deltagere					
	Snitt	**SD**	**Min**	**Max**	**α**	**n**	**Snitt**	**SD**	**Min**	**Max**	**α**	**n**	**Snitt**	**SD**	**Min**	**Max**	**α**	**n**
Antall unike estimater - Første uke	51,07	25,16	21,00	112		60	40,25	21,18	15	108		60	30,86	16,69	2	114		713
Brier-score forståelse (0–5)	1,91	1,86	0,00	5	0,83	46	0,74	1,16	0	5	0,66	39	0,79	1,05	0	5	0,52	392
Antall spørsmål besvart	166,72	65,91	40	240		60	160	62,67	38	239		60	143,3	61,04	30	240		713
Tid brukt per spørsmål (minutter)	1,83	0,69	0,65	4,12		58	1,65	0,79	0,5	4,98		60	1,32	0,53	0,33	3,77		704

Tabell B.11 Variabelverdier i FFIs og GJPs turneringer.

Referanser

‘Edge Master Class 2015: A Short Course in Superforecasting’, *Edge*, 17. aug.–21. sep. 2015. https://www.edge.org/conversation/philip_tetlock-edge-master-class-2015-a-short-course-in-superforecasting-class-i. Besøkt 6. des. 2021.

‘How to Be Less Terrible at Predicting the Future’, *Freakonomics*, 14. jan. 2016. http://freakonomics.com/podcast/how-to-be-less-terrible-at-predicting-the-future-a-new-freakonomics-radio-podcast/. Besøkt 6. des. 2021.

‘Measuring Accuracy in Prediction Markets and Opinion Poll/Pools’, *Cultivate Labs*. https://www.cultivatelabs.com/posts/measuring-accuracy-in-prediction-markets-and-opinion-poll-pools. Besøkt 6. des. 2021.

‘Setter mål om 30 prosent kvinner’, *Forsvarets Forum*, 4. juni 2019. https://forsvaretsforum.no/setter-mal-om-30-prosent-kvinner/106481. Besøkt 6. des. 2021.

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L. og Mellers, B. (2017), ‘Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls’, *Management Science*, 63:3, ss. 587–900.

Atanasov, P., Witkowsi, J., Ungar, L., Mellers, B. og Tetlock, P. (2020), ‘Small steps to accuracy: Incremental belief updaters are better forecasters’, *Organizational Behavior and Human Decision Process*, 160, ss. 19–35.

Baron, J. Scott, S. Fincher, K. og Metz, S. E. (2015), ‘Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)?’, *Journal of Applied Research in Memory and Cognition*, 4:3, ss. 265–284.

Beadle, A. W. (2017), ‘Er du bedre til å forutsi fremtiden enn ekspertene?’, *forskning.no*, 30. sept. 2017. https://forskning.no/krig-og-fred-fremtidsforskning-kronikk/kronikk-er-du-bedre-til-a-forutsi-fremtiden-enn-ekspertene/1161870. Besøkt 6. des. 2021.

Beadle, A. W. (2018), ‘FFIs prediksjonsturnering – idé- og metodebeskrivelse’, *FFI-rapport 18/00108* (Kjeller: FFI).

Beadle, A. W. (2021), ‘FFIs prediksjonsturnering – spørsmålskatalog’, *FFI-rapport 21/00736* (Kjeller: FFI).

Beadle, A. W. (2021), ‘Hvordan forbedre treffsikkerheten til prediksjoner av internasjonal politikk? – en litteraturgjennomgang’, *FFI-rapport 21/00735* (Kjeller: FFI).

Beadle, A. W. (2021), ‘Tilleggsdokumentasjon til foreløpige resultater fra FFIs prediksjonsturnering’, *FFI-notat 22/00133*(Kjeller: FFI).

Beadle, A. W. og Diesen, S. (2015), 'Globale trender mot 2040 – implikasjoner for Forsvarets rolle og relevans', *FFI-rapport 2015/01452* (Kjeller: FFI).

Beadle, A. W., Diesen, S., Nyhamar, T. og Bostad, E. K. (2019), 'Globale trender mot 2040 – et oppdatert fremtidsbilde', *FFI-rapport 19/00045* (Kjeller: FFI).

Bors, D. A. og Stokes, T. L. (1998), 'Raven's Advanced Progressive Matrices: Norms for First-Year University Students and the Development of a Short Form', *Educational and Psychological Measurement*, 58:3, ss. 382–398.

Boudreau, C. og Lupia, A. (2011), 'Political Knowledge', i Druckman, J. N., Green, D. P., Kuklinski, J. H. og Lupia, A., red., *Cambridge Handbook of Experimental Political Science*, ss. 171–183.

Brier, G. W. (1950), 'Verification of Forecasts Expressed in Terms of Probability', *Monthly Weather Review*, 78:1.

Bukkvoll, T., Glærum, S., Johansen, I., Diesen, S. og Lia, B. (2014), 'En gjennomgang av FFIs scenariogrunnlag for Forsvarets langtidsplanlegging', *FFI-rapport 2014/01154* (Kjeller: FFI). Begrenset.

Bæk, S. (2019), 'Forsvarets tidligere langtidsplaner – hvor godt har de sikkerhetspolitiske beskrivelsene truffet?', *FFI-notat 19/01609* (Kjeller: FFI).

Cacioppo, J. T. og Berntson, G. G. (1994), 'Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates', *Psychological Bulletin*, 115:3, ss. 401–423.

Cacioppo, J. T. og Petty, R. E. (1982), 'The Need for Cognition', *Journal of Personality and Social Psychology*, 42:1, ss. 116–131.

Cacioppo, J. T. og Petty, R. E. (1984), 'The Efficient Assessment of Need for Cognition', *Journal of Personality Assessment*, 48:3, ss. 306–307.

Carlsen, B., Müftüoglu, I. B. og Riese, H. (2014), 'Forskning i media: Forskere om motivasjon og erfaringer fra medieintervjuer', *Norsk medietidsskrift*, 21:3, ss. 188–208.

Cassidy, M. F. og Buede, D. M. (2009), 'Does the accuracy of expert judgment comply with common sense: caveat emptor', *Management Decision*, 47:3, ss. 454–469.

Chang, W., Atanasov, P., Patil, S., Mellers, B. og Tetlock, P. (2017), 'Accountability and adaptive performance under uncertainty: A long-term view', *Judgment and decision making*, 12:6, ss. 610–626.

Chang, W., Chen, E., Mellers, B. og Tetlock, P. (2016), 'Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments', *Judgment and Decision Making*, 11:5, ss. 509–526.

Cokely, E. T., Galesic, M., Schulz, E. og Ghazal, S. (2012), 'Measuring Risk Literacy: The Berlin Numeracy Test', *Judgment and Decision Making*, 7:1, ss. 25–47.

Dweck, C. (2006), *Mindset: The new psychology of success* (New York: Random House).

Ekspertgruppen for Forsvaret av Norge (2015), *Et felles løft* (Forsvarsdepartementet).

Ericsson, K. A. (2014), 'Why expert performance is special and cannot be extrapolated from studies of performance in the general population: A response to criticisms', *Intelligence*, 45, ss. 81–103.

Ericsson, K. A., Krampe, R. T., og Tesch-Romer, C. (1993), 'The role of deliberate practice in the acquisition of expert performance', *Psychological Review*, 100:3, ss. 363–406.

Etterretningstjenesten (2021), *Fokus 2021*.

Frederick, S. (2005), 'Cognitive Reflection and Decision Making', *Journal of Economic Perspectives*, 19:4, ss. 25–42.

Friedman, J. A. (2019), *War and Chance: Assessing Uncertainty in International Politics* (Oxford University Press).

Frisk, E. (2021), 'Cronbachs alfa', *Statistisk ordbok*. https://www.statistiskordbok.se/ord/cronbachs-alfa/. Besøkt 6. des. 2021.

Haran, U., Ritov, I. og Mellers, B. A. (2013), 'The role of actively open-minded thinking in information acquisition, accuracy, and calibration', *Judgment and Decision Making*, 8:3, ss. 188–201.

Helland-Riise, F, og Martinussen, M. (2017), 'Måleegenskaper ved de norske versjonene av Ravens matriser [Standard Progressive Matrices (SPM)/Coloured Progressive Matrices (CPM)]', *PsykTestBarn*, 2:2.

Johansen, I. (2006), 'Scenarioklasser i Forsvarsstudie 2007: En morfologisk analyse av sikkerhetspolitiske utfordringer mot Norge', *FFI-rapport 2006/02664* (Kjeller: FFI).

Johnson, D. D. P. (2004), *Overconfidence and war: The havoc and glory of positive illusions* (Cambridge, MA: Harvard University Press).

Kahneman, D. (2013), *Tenke, fort og langsomt* (Oslo: Pax Forlag).

Kahneman, D. og Klein, G. (2009), 'Conditions for Intuitive Expertise: A Failure to Disagree', *American Psychologist*, 64:6, ss. 515–526.

Kahneman, D. og Tversky, A. (1977), 'Intuitive prediction: Biases and corrective procedures', *Technical Report PTR- 1042-77-6* (Virginia: DARPA).

Karlsen, J. E. og Øverland, E. F. (2010), *Carpe Futurum* (Oslo: Cappelen Damm).

Krosnick, J. A. og Presser, S. (2010), 'Question and Questionnaire Design', i Marsden, P. V. og Wright, J. D., red., *Handbook of Survey Research*, 2. utg. (Bingley: Emerald Group Publishing Limited), ss. 263–313.

Kruglanski, A. W. og Webster, D. M. (1996), 'Motivated closing of the mind: "Seizing" and "freezing."', *Psychological Review*, 103:2, ss. 263–283.

Lipkus, I. M., Samsa, G. og Rimer, B. K. (2001), 'General Performance on a Numeracy Scale among Highly Educated Samples', *Medical Decision Making*, 21:1, ss. 37–44.

Lysne, V. og Olsen, T. (2017), 'Konfidensintervaller – hva kan de fortelle deg?', *Norsk tidsskrift for ernæring*, 1–2017.

Mary, I., Duranczyk, J. og Stottlemyer, S. L. (2013), 'Confidence Interval: Assumptions and Conditions', *OpenStax CNX*, 21. aug. 2013. https://cnx.org/contents/KnmPEWac@2/Confidence-Interval-Assumptions-and-Conditions. Besøkt 6. des. 2021.

McBride, M. F., Fidler, F. og Burgman, M. A. (2012), 'Evaluating the accuracy and calibration of expert predictions under uncertainty: predicting the outcomes of ecological research', *Diversity and Distributions*, 18:8, ss. 782–794.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E. og Tetlock, P. (2015), 'The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics', *Journal of Experiment Psychology: Applied*, 21:1, ss. 1106–1115.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L. og Tetlock, P. (2015), 'Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions', *Perspectives on Psychological Science*, 10:3, ss. 267–281.

Mellers, B., Tetlock, P. og Arkes, H. R. (2019), 'Forecasting tournaments, epistemic humility and attitude depolarization', *Cognition*, 188, ss. 19–26.

Merkle, E., Steyvers, M, Mellers, B. og Tetlock, P. (2016), 'Item Response Models of Probability Judgments: Application to a Geopolitical Forecasting Tournament', *Decision*, 3:1, ss. 1–19.

Merkle, E., Steyvers, M., Mellers, B. og Tetlock, P. (2017), 'A neglected dimension of good forecasting judgment: The questions we choose also matter', *International Journal of Forecasting*, 33:4, ss. 817–832.

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J. og Tenneye, E. R. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', *Management Science*, 63:11, ss. 3552–3565.

Muehlhauser, L. (2019), 'How Feasible Is Long-range Forecasting?', *Open Philanthropy*, 10. okt. 2019. https://www.openphilanthropy.org/blog/how-feasible-long-range-forecasting. Besøkt 6. des. 2021.

Neustadt, R. E. og May, E. R. (1986), *Thinking in Time: The Uses of History for Decision-Makers* (New York: The Free Press).

Niederle, M. og Vesterlund, L. (2011), 'Gender and Competition', *Annual Review of Economics*, 3:1, 601–630.

Nilstun, C. (2021), 'amatør', *Store norske leksikon*. https://snl.no/amatør. Besøkt 6. des. 2021.

Nygård, H. M. (2015), 'Prediksjon i Internasjonal politikk', *Internasjonal Politikk*, 73:4, ss. 467–487.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K. og Dickert, S. (2006), 'Numeracy and Decision Making', *Psychological Science*, 17:5, ss. 407–413.

Plomin, R., Shakeshaft, N. G., McMillan, A. og Trzaskowski, M. (2014), 'Nature, nurture, and expertise', *Intelligence*, 45, ss. 46–59.

Politiets sikkerhetstjeneste (2021), *Nasjonal trusselvurdering 2021*.

Pripp, A. H. (2018), 'Pearsons eller Spearmans korrelasjonskoeffisienter', *Tidsskriftet for Den norske legeforening*, 8. mai 2018.

Roets, A. og Hiel, A. V. (2011), 'Item selection and validation of a brief, 15-item version of the Need for Closure Scale', *Personality and Individual Differences*, 50:1, ss. 90–94.

Schrodt, P. A. (2013), 'Seven deadly sins of contemporary quantitative political analysis', *Journal of Peace Research*, 51:2, ss. 287–300.

Shipley, W. C., Gruber, C. P., Martin, T. A. og Klein, A. M. (2009), *Shipley-2 Manual* (Western Psychological Services).

Sirota, M. og Juanchich, M. (2018), 'Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the Cognitive Reflection Test', *Behavior Research Methods*, 50, 2511–2522.

Skovlund, E. (2017), 'Når bør man velge en ikke-parametrisk metode?', *Tidsskriftet for Den norske Legeforening*, 16. mai 2017.

Smeby, J.-C. (2021), 'ekspertise', *Store norske leksikon*. https://snl.no/ekspertise. Besøkt 6. des. 2021.

Strand, K. R., Gisnås, H. og Eggereide, B. (2019), 'Utvikling i sentrale HR-parametere i forsvarssektoren – et dypdykk hos Forsvarsbygg og Forsvarets forskningsinstitutt', *FFI-rapport 19/01599* (Kjeller: FFI).

Surowiecki, J. (2005), *The Wisdom of Crowds* (New York: Anchor Books).

Svartdal, F. (2021), 'korrelasjon – psykologi', *Store norske leksikon*. https://snl.no/korrelasjon_-_psykologi. Besøkt 6. des. 2021.

Taber, K. (2017), 'The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education', *Research in Science Education*, 48:1, ss. 1–24.

Taleb, N. N (2010), *The Black Swan* (New York: Random House).

Tetlock, P. (2005), *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton: Princeton University Press).

Tetlock, P. E. (1998), 'Close-call counterfactuals and belief-system defenses: I was not almost wrong but I was almost right', *Journal of Personality and Social Psychology*, 75:3, ss. 639–652.

Tetlock, P. E. (2017), *Expert Political Judgment: How Good Is It? How Can We Know?* (New Jersey: Princeton University Press).

Tetlock, P. E., Mellers, B. A. og Scobilic, J. P. (2017), 'Bringing probability judgments into policy debates via forecasting tournaments', *Science*, 355:6324, ss. 481–483.

Tetlock, P. og Gardner, D. (2015), *Superforecasting: The Art and Science of Prediction* (London: Random House Books).

Tetlock, P., Mellers, B., Rohrbaugh, N. og Chen, E. (2014), 'Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate', *Current Directions in Psychological Science*, 23:4, ss. 290–295.

Webster, D. M. og Kruglanski, A. W. (1994), 'Individual differences in need for cognitive closure', *Journal of Personality and Social Psychology*, 67:6, ss. 1049–1062.

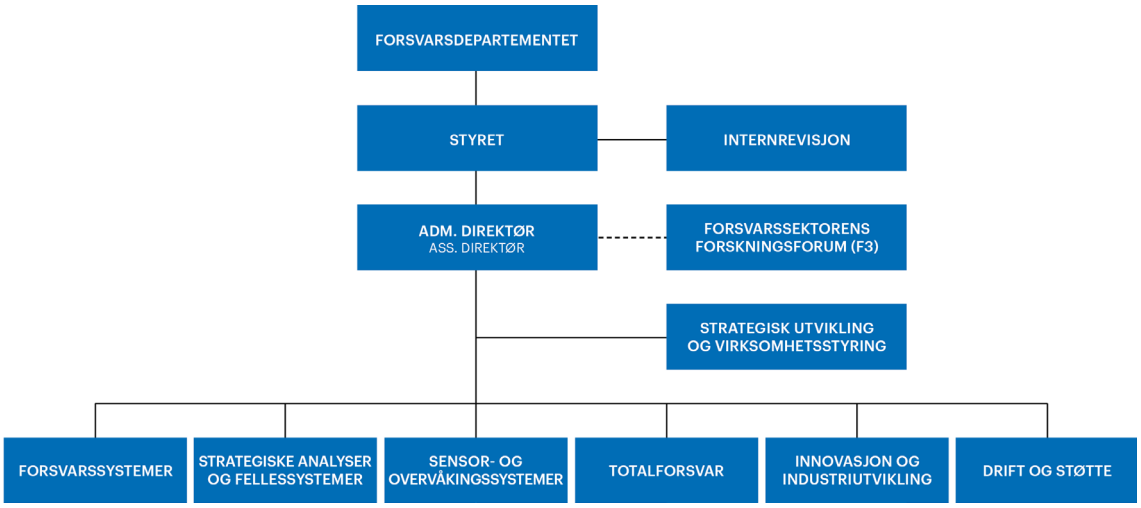
Åtland, K., Beadle, A., Diesen, S., Glærum, S., Mørkved, T., Nyhamar, T. og Stenersen, A. (2018), 'Gjennomgang av FFIs scenariogrunnlag for Forsvarets langtidsplanlegging, 2018', *FFI-rapport 18/00669* (Kjeller: FFI). Begrenset.

Om FFI
Forsvarets forskningsinstitutt ble etablert 11. april 1946. Instituttet er organisert som et forvaltningsorgan, med særskilte fullmakter underlagt Forsvarsdepartementet.

FFIs formål
Forsvarets forskningsinstitutt er Forsvarets sentrale forskningsinstitusjon og har som formål å drive forskning og utvikling for Forsvarets behov. Videre er FFI rådgiver overfor Forsvarets strategiske ledelse. Spesielt skal instituttet følge opp trekk ved vitenskapelig og militærteknisk utvikling som kan påvirke forutsetningene for sikkerhetspolitikken eller forsvarsplanleggingen.

FFIs visjon
FFI gjør kunnskap og ideer til et effektivt forsvar.

FFIs verdier
Skapende, drivende, vidsynt og ansvarlig.



Forsvarets forskningsinstitutt
Postboks 25
2027 Kjeller

Besøksadresse:
Instituttveien 20
2007 Kjeller

Telefon: 63 80 70 00
Telefaks: 63 80 71 15
Epost: post@ffi.no

Norwegian Defence Research Establishment (FFI)
P.O. Box 25
NO-2027 Kjeller

Office address:
Instituttveien 20
N-2007 Kjeller

Telephone: +47 63 80 70 00
Telefax: +47 63 80 71 15
Email: post@ffi.no