Indhold

1 Forord **2**

2 Indledning **3**

3 Konstruktion **3**

3.1 Supply-tests 3

3.2 Selection-tests 4

3.3 Hvad med usikkerheden? 5

4 Hvordan laver man en god multiple-choice test? **6**

5 Multiple-choice — den simple model **8**

5.1 Testresultatet som en stokastisk variabel 9

5.2 Middelværdi for stokastiske variable 10

5.3 Binomialfordelingen 10

5.4 Binomialfordelingen og test — et eksempel 12

6 En udbygget statistisk model **12**

6.1 Den statistiske usikkerhed i testsituationen 13

6.2 Statistisk inferens — en kort introduktion 14

6.3 Statistisk inferens i testsituationen 16

6.4 Modelforudsætninger og anvendelighed 17

6.5 Gæt i test 18

7 Konklusion og projektforslag **19**

8 Litteraturliste **20**

1 Forord

Denne rapport er udarbejdet i forbindelse med P0-projektperioden, 3/9-2001 - 21/9-2001 på den teknisk-naturvidenskabelige basisuddannelse på Aalborg Universitet. Formålet har været at udarbejde en detaljeret kortlægning af et problem som oplæg til den næste projektperiode, for hvilken temaet er "virkelighed og modeller". De i rapporten anvendte litteraturhenvisninger henviser til forfatterens efternavn — se litteraturlisten.

Når vi i rapporten skriver "vi", menes der gruppen som helhed.

Mange tak til vores vejleder Kim Emil Andersen for velvillig assistance!

Rapporten er udarbejdet af nedenstående:

Anders Gorst-Rasmussen

Lars Hornbæk Jensen

Dorte Klerke

Dennis Kraack

Charlotte Kramer

Anja Laursen

2 Indledning

Tests af individet får større og større relevans i det moderne samfund, hvor *kvailifikationer* er nøgleordet. Både testtager og testgiver er ofte dybt afhængige af testens nøjagtighed; tag for eksempel optagelsestesten til drømmeuddannelsen — eller arbejdsgiveren, der søger højt kvalificeret arbejdskraft.

Disse aspekter gør det af stor betydning, at en test giver et så klart og tydeligt billede af testtageren som muligt. Men hvordan sikres dette? Dette er det overordnede spørgsmål i denne rapport. Vi vil indledende kort introducere forskellige testtyper og konstruktionsusikkerheden — dernæst vil vi fokusere på udviklingen af en diskret statistisk model for den særlige multiple-choice test.

God fornøjelse med læsningen!

3 Konstruktion

De formelle oplysninger (testyperne) i dette afsnit er baseret på Gronlund, kap. 1, 8 og 9. Såfremt vi skal overveje en tests nøjagtighed og de usikkerhedsaspekter, der altid vil være til stede, er det naturligt først at overveje selve konstruktionen af testen. Det er kun rimeligt at forvente, at denne fase har stor betydning — et dårligt formuleret spørgsmål, eller et spørgsmål helt uden for sammenhængen siger jo eksempelvis intet om en testtagers evner. "Test" er imidlertid et temmelig bredt begreb — der findes færdighedstests, personlighedstests, holdningstests o.s.v. og den formelle testform varierer i lige så høj grad. Den let overskuelige multiple-choice test er én type — et højabstrakt stilemne er en anden. Den store variation gør det nødvendigt først at sætte sig ind i de grundlæggende forhold, der karakteriserer de enkelte testformer, hvis vi ønsker at overveje usikkerhedsaspekterne.
Altså, hvilken testform er optimal til et bestemt testformål — og hvilken usikkerhed er forbundet med konstruktionen af testen udfra en bestemt testmodel. Dette er nogle af de spørgsmål, vi vil tage stilling til i det følgende.

3.1 Supply-tests

Supplytesten er en særlig testform, hvor testtageren selv formulerer sig enten mundtligt eller skriftligt og således har en større frihed m.h.t. indholdet i testsvaret. Blandt typiske eksempler på supplytesten er:

- **Essay**
  Der gives et mere eller mindre bredt emne, hvorefter testtageren har en større frihed til at formulere sit svar. Dette kan betyde, at det fagligt testede område kan svinge ud og ind af det relevante område.

- **Problemløsning** Her stilles testtageren overfor et konkret problem og skal selv formulere en korrekt og hensigtsmæssig løsning.
- **Udfyldning med korte svar**
  Der gives en sætning eller påstand, hvor testtageren skal indsætte et ord eller en kortere sætning.

Den gennemgående frie form betyder typisk, at testgiveren vil have mindre kontrol over det svar, som bliver returneret. Dette vil i mange situationer selvfølgelig være uhensigtsmæssigt. Hvis man ønsker at måle bestemte færdigheder, er et omfattende svar til et simpelt spørgsmål helt unødvendigt og til mere skade end gavn. Antag f.eks. at et universitet tror, de har en dårlig underviser på kurset i avanceret analyse; dette vil de undersøge ved at teste de studerende i de helt grundlæggende begreber. I denne situation ville det være helt omsonst at spilde tid og energi på at udforme og rette en omfattende supplytest.
Supplytestens fortræffelighed og store anvendelighed kommer dog for alvor til sin ret i meget komplekse testsituationer. En jobsamtale, eksempelvis, kan vi opfatte som en avanceret form for supplytest. Her ønsker man dels at konstatere visse færdigheder ved ansøgeren, dels at få et overblik over holdninger, personlighed og vanskeligt "målelige" psykologiske færdigheder. Også i undervisningssituationer i abstrakte fag er det en fordel, hvis testtageren får mulighed for at vise sikker færden på mange planer — for eksempel stilskrivning.

3.2 Selection-tests

Selectionformen er et glimrende og normalt *ressourcebesparende* alternativ til supplytesten, og i talrige situationer være mindst ligeså sigende. Kort beskrevet er dette en form, hvor testtageren enten krydser felter af, som han eller hun mener indeholder det korrekte svar (multiple-choice), eller kombinerer forskellige elementer for at danne sande udsagn. Der findes utallige variationer, men for færdighedstesten alene er de mest anvendte

- **Sandt/falsk**
- **Sammensæt to der hører sammen** Et problem eller sætning bliver opstillet, og testtageren skal så finde den sætning, som indeholder svaret.
- **Multiple-choice** Her bliver et problem eller spørgsmål præsenteret, hvor testdeltageren så skal vælge netop ét rigtigt svar blandt flere forskellige svarmuligheder [1].

Selection-testen må umiddelbart betragtes som en mere sikker måling af testtageren fremfor eksempelvis sandt/falsk-testen — forudsat at der er betydelig flere svarmuligheder. Testformen er mere struktureret end supply-formen og giver testgiveren et mere

[1] I f.eks. holdningstests er det dog også muligt at opstille flere "rigtige" svarmuligheder; denne særtype vil vi dog ikke behandle nærmere.

klart, direkte overblik over testtagernes umiddelbare (og mere eller mindre utilslørede) færdigheder indenfor testområdet.

Dens force er imidlertid også dens svaghed — selectiontesten er oftest begrænset til undersøgelse af *færdigheder* indenfor *et enkelt testområde*. Generelt må man betragte det som meget vanskeligt at konstruere en sikker selectiontest, der opererer på flere niveauer (færdigheder, holdninger o.s.v.), fordi det må forventes, at de psykologiske omstændigheder fra individ til individ gør en absolut, objektiv test gennemgående unøjagtig.

Selectiontesten er især populær på højere læreanstalter i mange engelsktalende lande; og det er selvfølgelig ikke ubegrundet. Testtypen giver f.eks. underviseren mulighed for at stille et langt bredere spektrum af spørgsmål, hvorved faren ved at støde på et uventet spørgsmål bliver formindsket betydeligt. I en traditionel supply-eksamenssituation (som vi kender dem her i landet), kan et enkelt uventet blandt de relativt få spørgsmål have stor betydning for testens udfald.

3.3 Hvad med usikkerheden?

Vi har nu kort gjort rede for visse afgørende karakteristika ved hhv. selection- og supplytesten. Den afgørende forskel ligger naturligvis i grundideen i den enkelte test — en supplytest vil i sin vurdering typisk være overvejende subjektiv og veje mange forskellige færdigheder, specifikke formuleringer o.s.v. mod testtagerens konkrete færdighed. En sådan vurdering kan med nutidens metoder kun vanskeligt eller slet ikke systematiseres, men må foretages af eksperter — undervisere ell. lignende. Bedømmelsen ved denne type test afhænger derfor i høj grad af ekspertens vurdering, vægtningen af de adskillige subjektive forhold — men det er vigtigt at understrege, at *testtagerens sikkerhed typisk er ganske god under denne testform*, eftersom bedømmelsesfasen også tager højde for gråzoner; altså ikke bare sandt-falsk. Testforløbet er her en *fleksibel* proces, hvor fejltrin og usikkerheder på et punkt relativt let kan korrigeres på et senere punkt. F.eks. kan eksperten, der bedømmer en besvarelse nemt korrigere for et dårligt spørgsmål ved blot at medtage dette i sine overvejelser.
De samme muligheder er ikke nært så let tilgængelige i selectiontestens bedømmelsesovervejelser. Denne er jo (for færdighedstesten) i sin udformning af streng objektiv karakter, og den systematisk-logiske opbygning indebærer en systematisk og derved objektiv vurdering (et svar er enten rigtigt eller forkert — der er ikke noget midt imellem!). Processen er her i sine rene form infleksibel og absolut — man kan ganske vist korrigere for f.eks. dårlige spørgsmål, men det er en noget mere omfattende proces end for supplytesten. Den logiske, absolutte opbygning har flere konsekvenser, hvoraf mange berører usikkerhedsproblemet. En række af ulemperne er:

- Konstruktionsfasen af testen kan være en alvorlig trussel mod testens sikkerhed. Medmindre man ofrer mange kræfter på at undersøge testens pålidelighed, kan man p.g.a. den absolutte opbygning f.eks. risikere, at et dårligt formuleret spørgsmål (eller slet og ret et dårligt spørgsmål), kan følge testtageren helt frem til bedømmelsen; og evt. være medvirkende til en ufortjent, dårligere vurdering; simpelthen fordi der

kun findes sandt-falsk. Som nævnt er dette af mindre betydning i supplytesten.

- Besvarelsen behøver kun i et vist omfang at være videnbaseret. Gæt er en fortræffelig og brugbar metode for testtageren, der støder på et uventet spørgsmål — dette er ligeledes en alvorlig trussel mod sikkerheden.
- Der må forventes at være en mærkbar statistisk usikkerhed.

  Eksempler på fordele kan bl.a. være:

- Automatiseret bedømmelse (f.eks. med computer/scanner) er muligt med store ressourcebesparelser til følge.
- Selectiontestens absolutte karakter, hvor der kun skelnes mellem ekstremerne, gør det nemmere og mere nærliggende at anvende matematiske metoder til vurdering af dels konstruktionsusikkerheden [2], dels den førnævnte statistiske usikkerhed for både testtager og testgiver. Sidstnævnte vil vi se nærmere på i afsnit 6, hvor vi bl.a. tager gætproblemet — og mulig korrektion herfor — op til overvejelse.

Den første af de nævnte fordele giver mange interessante perspektiver, der i første omgang overskygger ulemperne. Det ressourcebesparende aspekt er ikke begrænset til automatisk bedømmelse, men også m.h.t. automatisk fordeling af tests til testtagere. Her tænkes på mulighederne for tests via internettet — antag f.eks. at et stort firma har en række meget eftertragtede elevpladser, som de forventer tusindvis af internationale ansøgninger til. En indledende sortering v.h.a. en færdighedstest i f.eks. handel og økonomi vil være meget nærliggende — den kunne eksempelvis gennemføres som onlinetest af multiple-choice typen, og firmaet ville have sparet mange hundrede timer hos konsulentfirmaer. Vi har teknologien til at tage sådanne metoder i brug, og de mange rationaliseringer for tiden vil måske bane vejen for denne metode. Det er langtfra usandsynligt, at det er en sådan fremtid vi står overfor.

Vi har derfor fundet det relevant at behandle selectiontesten mere indgående. Dels fordi de kontekstuelle perspektiver ligger mere åbne, dels fordi det systematiske aspekt i usikkerheden bestemt er interessant og formentlig vil kunne føre til testforbedringer ad matematisk vej. Mere om dette fra afsnit 5. For at begrænse os, har vi valgt den mest repræsentative (og udbredte), men også en af de simpleste selectiontyper — *multiple-choice testen.*

4 Hvordan laver man en god multiple-choice test?

Afsnittet er baseret på Gronlund, kap. 3.
Vi har allerede konstateret, at konstruktionsfasen må anses for værende af stor betydning for selectiontestens sikkerhed. Som oplæg til den videre behandling har vi derfor fundet

[2] *Psykometrien* beskæftiger sig med dette — bl.a. *korrelationsanalyse* m.m. Jf. Crocker for en grundig behandling af dette.

det passende at overveje — hvad skal der til, helt basalt, for at usikkerhedsmomentet fra konstruktionen af en multiple-choice test minimeres?

Konfrontationen mellem den subjektive konstruktionsfase og den objektive testnatur bør så vidt muligt minimeres — for det er her, usikkerheden lettest kan opstå. D.v.s præcision, homogenitet og systematik er nøgleordene, når en god test skal konstrueres, og billedet af testtageren skal være så korrekt som muligt. En perfekt test er desværre en temmelig hypotetisk størrelse, men nedenfor følger en uddybning af nøgleordene; det minimum af overvejelser, man bør gøre sig under konstruktionsfasen:

- Overskuelighed. Sørg for at placere teksten i selve spørgsmålet frem for i svarmulighederne, for derved at gøre sidstnævnte lettest overskuelige.
- Entydighed. Sørg for at der er et, entydigt korrekt svar. Færdighedstesten af selectiontesten bygger jo netop på, at der er ét og kun ét rigtigt svar.
- Sproglig homogenitet. Svarmulighedernes længde skal ikke variere for meget fra hinanden — dette kunne fejlagtigt henlede testtagerens opmærksomhed på en bestemt svarmulighed.
- Systematik og præcision. Spørgsmålene bør være klart og præcist formulerede for at undgå utilsigtet vildledning af testtageren.
- Homogenitet i svarmuligheder. De forkerte svarmuligheder skal være fristende for den eller de personer som mangler den nødvendige viden for at svare rigtigt på det pågældende spørgsmål.
- Tilfældighed. Placeringen af det rigtige svar bør varieres på en tilfældig måde, f.eks. ved at trække lod eller få en maskine/computer til at vælge rækkefølgen af svarene.
- Uafhængighed. Forskellige spørgsmål bør være enkeltstående, uafhængige af svaret på tidligere spørgsmål — i modsat fald kan enkelte, uventede spørgsmål blive af stor betydning for den endelig score.

Eksempel

Hvad er årsagen til, at vi har nat og dag?

A: Jorden drejer rundt om sin akse.
B: Jorden drejer rundt om solen.
C: Skyer skygger for solens lys.
D: Jorden bevæger sig ind og ud af solens skygge.
E: Solen drejer rundt om jorden.

Dette er et typisk spørgsmål i en multiple-choice test, hvor der er 5 svarmuligheder — men kun et rigtigt svar. Formuleringen er klar og overskuelig, svarmulighederne har

gennemgående alle en rimelig appel; homogeniteten er tydelig. Alligevel er spørgsmålet næppe perfekt — man kunne f.eks. argumentere, at svar C er så urealistisk, at det for en testtager overhovedet ingen appel vil have!

Det skal understreges, at de nævnte konstruktionsovervejelser kun er minimumskrav! Den optimale konstruktion af en test er en langvarig og træg proces, der behandles nærmere i *psykometrien*. Her vil ovenstående overvejelser kun være en lille del af konstruktionsfasen — dernæst vil man afprøve spørgsmålene, analysere resultater, rette i testen, afprøve den igen o.s.v. [3]
Fra konstruktionsproblemet vil vi nu bevæge os over til helt andet (og betydelig mere interessant) aspekt i forbindelse med multiple-choice. Det er allerede nævnt kort, at testtypens absolutte og logiske natur må gøre den ganske velegnet til matematiske analyser. Det er vores håb, at det er muligt at konstruere en så omfattende matematisk model, at man i det mindste kan tage højde for visse af usikkerhedsaspekterne. Herunder figurerer naturligvis den *statistiske usikkerhed*, men bestemt også et vigtigt spørgsmål som problemet med testtagere, der gætter svarene.
I denne rapport vil vi kun overveje de meget grundlæggende dele af en sådan model og desforuden opstille en række af de relevante problemstillinger, man kunne ønske at belyse med en fuldt udviklet model. Som en introduktion til dele af den bagvedliggende teori vil vi tage udgangspunkt i en særlig simpel model — testbesvarelsen ved gætteri alene!

5 Multiple-choice — den simple model

Som nævnt i afsnit 3.3 er selectiontestens logisk-systematiske opbygning fortræffelig at analysere med statistiske metoder. Statistiskken kan hjælpe os med at forstå den stokastiske, d.v.s. tilfældige natur, der altid vil være ved en konkret testbesvarelse — i det følgende vil vi derfor introducere grundlæggende statistiske begreber, der skal danne basis for en nærmere analyse af denne tilfældighed, den statistiske usikkerhed. Vi forudsætter, at læseren har kendskab til de mest grundlæggende begreber indenfor sandsynlighedsregning og statistik, herunder definitionen statistiske eksperimenter, udfaldsrum, definitionen på sandsynlighed m.m. For uddybning af dette, jf. f.eks. Engelhardt kap. 1.
Som lovet vil vi i dette afsnit overveje et specielt, uhyre simpelt eksempel på anvendelse af statistiske begreber m.h.t. multiple-choice. Begreberne illustrerer imidlertid glimrende, hvordan grundlæggende statistiske ideer kan overføres på det mere generelle multiple-choice problem.
Antag, at vi giver en testtager en bestemt test — og vi beder vedkommende gætte helt tilfældigt. Bagefter giver vi ham et nyt eksemplar af testen, og han gætter igen o.s.v. Vi kan hurtigt indse, at det vil være temmelig tilfældigt, hvilken score testtageren opnår i en bestemt test; men samtidig er det heller ikke vanskeligt at konstatere, at der på *en eller anden måde må være en form for systematik i denne tilfældighed*. Hvis der eksempelvis er 4 svarmuligheder ved et spørgsmål, må sandsynligheden for et rigtigt svar være 25%.

[3] For nærmere behandling af disse spørgsmål, jf. f.eks. Crocker/Algina.

For at give en formel beskrivelse af denne generelle systematik, vil vi nu formelt indføre flere afgørende begreber.

5.1 Testresultatet som en stokastisk variabel

Grundideen bag begrebet *stokastisk variabel* er et af de vigtigste overhovedet i den grundlæggende statistik. Vi har valgt at starte med dette — og ikke med udfaldsrum ell. lign., da opfattelsen af testresultatet som en stokastisk variabel er specielt afgørende for den videre problembehandling. Den resterende del af afsnit 5 er baseret på Engelhardt, kap. 2 og 3.

Definition 1 (Diskret stokastisk variabel) *Lad U være et udfaldsrum for et statistisk eksperiment. En diskret* stokastisk variabel *er en funktion $X : U \to D \subseteq N_0$. Lad $u \in U$ være et udfald — værdien af $X(u) = x$ kaldes en* realisation *af den stokastiske variabel X.* [4]

Vi kan generelt opfatte antallet af rigtige i en test som en realisation af en stokastisk variabel — i gætsituationen er dette indlysende. Hvis man krydser svar af helt tilfældigt, vil det være lige så tilfældigt, hvilken score man ender med. Men også i situationen, hvor man tænker sig om, kan vi opfatte scoren som en realisation af en stokastisk variabel — den nøjagtige score er jo altid ukendt. Man kan måske sige noget om, hvor man regner med at den vil ligge, men testspecifikke omstændigheder vil altid give en vis variation. Måske læser man ikke et eller flere spørgsmål ordentligt og svarer dermed forkert. Måske gætter man ved nogle spørgsmål — og så er vi igen tilbage ved den simple model.
Vi vil nu definere tæthedsfunktionen for en stokastisk variabel og begrebet *stokastisk uafhængighed*, der har stor betydning for vores senere resultater.

Definition 2 (Tæthedsfunktion) *Ved tæthedsfunktionen f for en (diskret) stokastisk variabel X forstås en funktion, der for enhver realisation $x \in Vm(X)$ knytter sandsynligheden $P(X = x)$* [5]*. D.v.s. at*

$$f(x) = P(X = x), x \in Vm(X).$$

Definition 3 (Stokastisk uafhængighed) *De stokastiske variable $X_1, X_2, \ldots, X_n$ siges at være* uafhængige*, hvis*

$$f(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} f(x_i)$$

[4] Betragter man sæt af stokastiske variable, taler man om en *stokastisk vektor*
[5] Tæthedsfunktionen kaldes også sandsynlighedsfunktionen eller fordelingsfunktionen

5.2 Middelværdi for stokastiske variable

Vi får gentagne gange brug for en særlig funktion af stokastiske variable — middelværdien. Typisk introducerer man sammen med middelværdien også en anden funktion, *variansen* — denne kræver dog en række dybere overvejelser, hvorfor vi helt har udeladt den. Interesserede henvises til Engelhardt pp. 73-75.

Definition 4 (Middelværdi for stokastisk variabel) *Lad X være en diskret stokastisk variabel med tæthedsfunktionen $f(x)$. Middelværdien $E(X)$ for X defineres som summen*

$$E(X) = \sum_x xf(x)$$

Middelværdien for den stokastiske variabel X er blot en generaliseret udgave af gennemsnittet for et sæt målinger $\frac{1}{n}\sum_{i=1}^{n} x_i$, hvor "tæthedsfunktionen" er en konstant $f(x) = \frac{1}{n}$. Der gælder adskillige regneregler for middelværdien, men ingen af dem er særligt overraskende. Vi får senere brug for det faktum at middelværdien er en *linær transformation* — d.v.s. hvis $X_1, X_2, \ldots X_n$ er stokastiske variable, hvor n-vektoren $\mathbf{x}$ er en realisation af disse, så er

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum\left(\left(\sum_{i=1}^{n} x_i\right) f(\mathbf{x})\right) = \sum_{i=1}^{n} E(X_i)$$

5.3 Binomialfordelingen

Vi har nu de fleste af de nødvendige teoretiske værktøjer til at beskrive den lettere urealistiske testsituation, hvor testtageren blot gætter tilfældigt ved samtlige spørgsmål.

Problematikken ligger naturligvis i at bestemme en tæthedsfunktion for den stokastiske variabel givet i testresultatet. Hvor stor er sandsynligheden for f.eks. at få 37 rigtige i en bestemt test ved gæt alene. Dette spørgsmål kan vi efter denne sektion ganske let besvare.
Multiple-choice testens logiske opbygning gør her for alvor arbejdet lettere for os. Vi kan nemlig tillade os at nøjes med at overveje besvarelsen af *et enkelt spørgsmål ad gangen.* Mulighederne er som bekendt enten rigtigt eller forkert, d.v.s. vi har to udfald, og dermed et særligt simpelt udfaldsrum.[6]

Besvarelsen af et spørgsmål kan vi betragte som en diskret stokastisk variabel — hvor udfaldet *rigtig* tildeles funktionsværdien 1 og udfaldet *forkert* værdien 0. Hvis der er n mulige svar, og $p = \frac{1}{n}$, så er $P(X = 1) = p$, mens $P(X = 0) = 1 - p$. Dermed har vi alle funktionsværdier for tæthedsfunktionen og kan hurtigt fastlægge en forskrift. Vi ser, at den kan skrives som

$$f(x) = \begin{cases} p^x(1-p)^{1-x} & for \quad x = 0, 1 \\ 0 & ellers \end{cases}$$

[6] I særlige situationer, som vi *ikke* vil overveje, kan man skelne mellem mere end blot rigtigt og forkert. Derfor kan metoderne, vi her gennemgår, *ikke* anvendes!

Dette er et konkret eksempel på den såkaldte *Bernoullifordeling* jf. Engelhardt p. 91.

Definition 5 (Bernoullifordelingen) *Lad et eksperiment have netop to udfald,* succes *og* fiasko*, hvor* $P(succes) = p \Rightarrow P(fiasko) = 1 - p$.
Et sådant eksperiment kaldes et Bernoulli-forsøg *og kan beskrives ved den særlige* Bernoulli stokastiske variabel X

$$X(e) = \begin{cases} 0 & hvis \quad e = fiasko \\ 1 & ellers \end{cases}$$

Tæthedsfunktionen for X*, Bernoullifordelingen, kan da skrives* [7]

$$f(x) = \begin{cases} p^x(1-p)^{1-x} & for \quad x = 0, 1 \\ 0 & ellers \end{cases}$$

Vi kan nu opfatte en testbesvarelse af en test på n spørgsmål som en *sekvens af Bernoulliforsøg*, $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ — hvad er da fordelingen for $X = \sum_{i=1}^{n} X_i$, testscoren? Da der antages stokastisk uafhængighed mellem besvarelserne (et af konstruktionskravene til en god multiple-choice test, jf. afsnit 3.3) — d.v.s. et testsvar afhænger ikke af svaret på nogle af de øvrige — er sandsynligheden for en *bestemt* sekvens med k rigtige svar og $n - k$ forkerte givet ved $\prod^{n} P(X = 1) \prod^{n-k} P(X = 0) = p^n(1-p)^{n-k}$.
Vi ønsker imidlertid sandsynligheden for *enhver* sekvens med n rigtige og $n - k$ forkerte — og må derfor overveje antallet af mulige kombinationer — altså antal måder, hvorpå vi kan kombinere n rigtige og $n - k$ forkerte uden hensyn til rækkefølgen. Det er let at indse, at der må være $K(n, n-k) = K(n, k)$ forskellige sådanne sekvenser [8]— altså er sandsynlighedsfunktionen for summen X af Bernoulli stokastiske variable givet ved

$$P(X = n) = K(n, k)p^x(1-p)^{n-k}$$

Denne særlige tæthedsfunktion kaldes *binomialfordelingen* — er en stokastisk variabel X binomialfordelt med parametre n og p, skriver man typisk X $b(n; p)$. V.h.a. denne har vi nu lavet en konkret model for vores simplificerede test — det vil vise sig, at hvis vi tegner tæthedsfunktionen for en binomialfordelt stokastisk variabel, vil i grove træk få en "klokkeformet" kurve omkring middelværdien symmetrisk omkring middelvrædien. Vi kan bl.a. beregne, hvor stor sandsynligheden er for et vist antal rigtige, $P(X = k)$ og med en smule mere teori kunne vi også gøre rede for sandsynligheden for mindst at have et vist antal rigtige, $P(X \geq k)$. Der er et væld af muligheder, som vi dog kun vil komme ganske kort ind på med et mindre eksempel — derefter vil vi bevæge os videre til en mere realistisk model.

Sætning 1 (Middelværdi for binomialfordelingen) *Hvis* X *er en diskret stokastisk variabel fordelt b(n,p) med værdimængde* $x_1, x_2, \ldots, x_n$*, så gælder der for middelværdien* μ*, at*

$$\mu = E(X) = np$$

[7] Egentlig burde dette være en selvstændig sætning, men resultatet er så indlysende, at det er undtagelsesvist er medtaget i definitionen.

[8] For en nærmere uddybning af permutationer og kombinationer jf. Engelhardt pp. 34-35

Følgende bevis er baseret på Antonius p. 237.

Bevis 1 *Vi betragter igen den stokastiske variabel X som en sum af Bernoulli stokastiske variable. For den i'te Bernoulli stokastiske variabel er middelværdien givet ved $E(X_i) = 1 \cdot p + 0 \cdot (1-p) = p$. Da vi ved, at middelværdien er en lineær transformation fås, at $E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = np$.*

5.4 Binomialfordelingen og test — et eksempel

Vi vil nu give et konkret eksempel på, hvordan vi kan bruge den opstillede binomialfordeling til at belyse vores specielle test med konsekvent gætning.
Antag, at en multiple-choice test består af 100 spørgsmål, hver med 5 forskellige svarmuligheder. Så er sandsynligheden for at svare rigtigt (succes) ved gæt alene givet ved $p = \frac{1}{5} = 0,2$, mens sandsynligheden for fiasko bliver $1 - \frac{1}{5} = \frac{4}{5} = 0,8$. Den stokastiske variabel, der tæller scoren i testen er altså binomialfordelt med parametre $(100; 0,2)$, i.e.

$$P(X = k) = K(100, k)p^{0,2}(1-p)^{0,8}$$

Dette giver en middelværdi på $\mu = E(X) = np = 100 \cdot 0,2 = 20$ — altså 20 rigtige svar. Det er faktisk ganske mange point, man kan score uden nogen form for tankevirksomhed — og det er et af de alvorligste problemer, vi må overveje i en mere udbygget model (jf. afsnit 6)

Vores gennemgang har indtil videre været baseret på, at testtageren gættede ud i den blå luft. Det er naturligvis dybt urealistisk, men i næste afsnit vil vi forsøge at udbygge binomialmodellen, så den kan anvendes i praktiske, realistiske situationer til at beskrive den statistiske usikkerhed ved tests. Husk på, at ethvert testresultat er en stokastisk variabel, der i princippet kan give alle mulige testscores, blot med større og mindre sandsynlighed. Det er et alvorligt problem, som blandt andet kan kompenseres for v.h.a. den såkaldte *statistiske inferens*.

6 En udbygget statistisk model

Indtil videre har vi kigget på konstruktionen af multiple-choice tests og sågar med den grundlæggende statistik fundet en simpel model for en testsituation. I den virkelige verden er det dog næppe megen nytte til, hvis man konsekvent antager, at testtageren gætter. Derfor vil følgende afsnit forsøge at anvende binomialmodellen og grundbegreber fra den mere avancerede statistik til at opbygge et muligt fundament for en brugbar, diskret model — en model, der skal kunne tage højde for i hvert fald de mest nødvendige og relevante aspekter ved test af individet. Hvorvidt det overhovedet er muligt at lave en tilstrækkelig god model — og hvilke metoder man med fordel kan anvende — kan vi på ingen måde tage endelig stilling til i denne rapport, men det er bestemt et interessant emne for P1!
Vi vil indledende opstille følgende krav

- Modellen skal kunne beskrive testsituationen for individet udfra vedkommendes viden.
- Skal kunne tage stilling til visse usikkerhedsaspekter.
- Skal være så omfattende som muligt inden rimelighedens grænser.

6.1 Den statistiske usikkerhed i testsituationen

Indtil videre har vi mestendels diskuteret den usikkerhed, der er forbundet med selve testkonstruktionen. Det kan heller ikke fremhæves nok — det er denne, der er afgørende. Man behøver eksempelvis ikke meget kritisk sans for at kunne konstatere, at testresultater fra en test i lineær algebra, der kun har spørgsmål om lineære ligningssystemer, på ingen måde kan betragtes som et mål for elevens viden om lineær algebra. Hvis der overhovedet skal være nogen mening med at anvende tests til at sætte tal på en persons formåen, må man forlange en vis konsistens og relevans i forhold til testsituationen. Med andre ord må vi forlange, at testspørgsmålene er gode repræsentanter for hele testområdet. I statistikken er dette et generelt problem — eksempelvis forudsætter pålidelige meningsmålinger jo også, at undersøgelserne er udført blandt et repræsentativt udvalg af befolkningen.
Vi vil derfor opstille følgende definition for den konkrete talværdi for en persons *nøjagtige formåen*. Definitionen skal alene ses i relation til den ønskede model — som en simpel forenkling, på ingen måde nogen psykologisk kendsgerning

Definition 6 *Lad M, $|M| = N$, være en passende stor mængde af* repræsentative *testspørgsmål indenfor et givet testområde. Hvis en person kan besvare n af de N spørgsmål, definerer vi personens formåen i dette testområde som forholdet $\frac{n}{N}$.*

Definitionen er naturligvis ikke entydig og afhænger helt af testområdet. F.eks. kunne en person hævde, at vedkommende havde fundet 10000 spørgsmål, der repræsenterede et kursus i statistik — en anden person ville måske hævde, at der skulle mindst 20000 spørgsmål til, før området var tilfredsstillende dækket. Denne mangel på entydighed er selvfølgelig problematisk [9] men i første omgang ligegyldig, da vi ikke på nogen måde vil forsøge at anvende den konkret. Den er alene en hypotetisk størrelse, der skal tjene som en hjælp til at forstå, hvad vi rent matematisk mener med en persons formåen.
Princippet i den nøjagtige bestemmelse af en persons formåen kan imidlertid give os et overordnet billede af, hvordan man mest praktisk kan konstruere og opfatte en realistisk test — nemlig som *et tilfældigt udvalgt antal spørgsmål fra den større mængde spørgsmål, der dækker hele testområdet.*[10]

[9] Tilnærmet entydighed kunne muligvis opnås ved anvendelse af kontinuerte modeller, jf. afsnit 6.4.

[10] Dette *er* faktisk en anvendt metode — man konstruerer store testbanker med tusindvis af spørgsmål; og en underviser ell. lignende kan så sammensætte en test udfra mere eller mindre tilfældigt valgte spørgsmål.

I statistikken kaldes sidstnævnte *populationen*, mens førstnævnte er en *stikprøve*, der giver anledning til *observationer* af besvarelsen (rigtig eller forkert). Såfremt vi opfatter en test på denne måde (og det kræver unægtelig visse forudsætninger, hvis relevans og realisme bestemt kan diskuteres), har vi en lang række matematiske værktøjer til rådighed i den såkaldte *statistiske inferens*.

6.2 Statistisk inferens — en kort introduktion

Statistisk inferens er groft sagt en videreudvikling af den *deskriptive statistik*. Sidst nævnte er velkendt — her præsenteres man for en enkeltstående mængde af data, der skal organiseres og præsenteres. Oplysningerne i Danmarks Statistik er et fortræffeligt eksempel på konkret anvendelse af deskriptiv statistik.

I statistisk inferens bygges der videre på dette. Her får man givet en mængde data (en stikprøve) som en tilfældig del af en større helhed (populationen) og fortolker og analyserer disse i relation til helheden. Spørgsmålet er altså; hvor meget kan vi sige om populationen ud fra informationerne i en stikprøve — hvor meget kan vi stole på vores deskriptive statistik?

Overført på testsituationen lyder spørgsmålet derfor — hvad siger en enkelt test om, hvordan en testtager ville klare sig i mange testsituationer? Dette er en direkte omformulering af usikkerhedsspørgsmålet for testtageren. Hvor statistisk sikker kan vedkommende være på, at han eller hun får en så rimelig bedømmelse af sine evner som muligt ved blot en enkelt test?
Vi vil nu kort indføre de mest basale af de begreber, der er nødvendige for at forstå karakteren af den statistiske problemstilling i forbindelse med selection-testusikkerheden. På baggrund af den matematiske introduktion vil vi fremhæve nogle af de afgørende forhold, der kunne danne baggrund for en mere omfattende P1-projektanalyse. Afsnittet er i overvejende grad baseret på Engelhardt, kap. 9.

Princippet i stikprøveudtagning bygger som nævnt på, at man gerne vil lære noget om populationens ukendte fordeling; eller simplere, en eller flere ukendte parametre for populationens kendte fordeling. De ukendte parametre befinder sig i *parameterrummet*, Ω, der indeholder alle mulige værdier af parametrene. Ønsker vi således for en binomialfordelt population at bestemme sandsynlighedsparameteren p, ved vi på forhånd, at $0 \leq p \leq 1$, d.v.s. $\Omega = \{p|0 \leq p \leq 1\}$.

Vi vil nu definere stikprøveudtagning som en metode til at undersøge den konkrete værdi af en eller flere ukendte parametre

Definition 7 (Stikprøve) *En stikprøve på n elementer er en mængde af n stokastiske variable $(X_1, X_2, \ldots, X_n)$ udtaget fra en population med fordeling $f(x|\theta)$ (hvor θ er en ukendt parameter), således at det for tæthedsfunktionen gælder*

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i).$$

En stikprøve er altså defineret ved en udtagning af elementer *uafhængigt* af hinanden (stokastisk uafhængighed som i afsnit 5.1) [11], og det er af stor betydning, hvis vi skal kunne anvende stikprøven til at inferere om populationen. Naturligvis ville det være optimalt, hvis stikprøven bedst muligt repræsenterede populationen; problemet er bare, at der altid er et eller flere ukendte forhold ved populationen, der umuliggør dette (ellers var der jo ingen grund til stikprøveudtagning). Tilfældig udtagning af elementer vil være den naturlige metode og må derfor forventes at kunne indeholde de mest korrekte informationer om populationen.
Ud fra stikprøven må vi nu forvente at kunne drage i hvert fald visse slutninger om den ukendte parameter θ for populationens fordeling. Antag at vi måler på en bestemt egenskab for en population og at den ukendte parameter er middelværdien for denne egenskab, μ — så er det rimeligt at antage, at stikprøvens middelværdi $\overline{X}$ kan anvendes til at *estimere* μ. Stikprøvens middelværdi er et særligt eksempel på en generel klasse af funktioner af stokastiske variable, *statistikker*; en såkaldt *estimator*.

Definition 8 (Statistik og estimator) *En* statistik *er en funktion* $\tau(\boldsymbol{X})$ *af den stokastiske n-vektor* $\boldsymbol{X} = (X_1, X_2, \ldots X_n)$*, således at* τ ikke *afhænger af ukendte variable. Lad* $\boldsymbol{X}$ *have en fordeling med den ukendte parameter* $\theta \in \Omega$*. En statistik, der anvendes til at estimere værdien af* θ *(d.v.s. afbilder ind i parameterrummet* Ω*), kaldes en* punktestimator*, mens funktionsværdien* $\tau(\boldsymbol{x})$ *af en realisation af* $\boldsymbol{X}$ *kaldes et* punktestimat.

Det afgørende er her, at punktestimatoren som en funktion af stokastiske variable selv er en stokastisk variabel. Det skal forstås således, at f.eks. en stikprøvemiddelværdi er en stokastisk variabel, der varierer fra stikprøve til stikprøve. Det har afgørende betydning i testproblemet, hvor usikkerheden for testtageren jo netop skyldes, at en enkelt test ikke er den endelige vurdering af vedkommendes formåen, men derimod en realisation af en stokastisk variabel (der har en ny fordeling), hvis rimelighed må vurderes.
Netop denne stokastiske natur gør, at punktestimater desværre har visse begrænsninger. Eksempelvis er et punktestimat $\overline{X}$ for middelværdien μ jo netop bare en realisation af en stokastisk variabel og kan derfor afvige betragteligt fra den sande værdi af μ (jf. evt. Engelhardt kap. 8). Denne forskel giver en fejl, hvis middelværdi er kendt som estimatorens *bias*.

Definition 9 (Bias) *Lad* $\tau(\boldsymbol{x}) = \hat{\theta}$ *være et punktestimat for parameteren* θ*. Estimatorens bias defineres som middelværdien af stikprøvefejlen* $\theta - \hat{\theta}$

$$bias = E(\theta - \hat{\theta})$$

Det følger, at hvis $E(\theta - \hat{\theta}) = 0$*, er estimatoren* unbiased.

Fejlen i punktestimater kan altså minimeres ved henholdsvis at

- Anvende unbiased estimatorer.

[11] De stokastiske variable er uafhængige og ens fordelt, *independent and identically distributed, i.i.d*

- Udtage mange stikprøver og anvende middelværdien af punktestimaterne som estimat.

Der findes metoder til at teste, hvorvidt estimatorer er unbiased eller ej, men det vil vi ikke komme nærmere ind på her. Det er blot vigtigt at bemærke, at punktestimater ikke er nogen endegyldig metode til estimationsproblemer — og specielt i testsituationen er det nok mere relevant at foretage såkaldte *intervalestimater*[12] (jf. Engelhardt kap. 11). Den teoretiske baggrund for disse er dog temmelig omfattende, men de kunne være interessante at studere i et fremtidigt projekt.

6.3 Statistisk inferens i testsituationen

Vi vil nu forsøge at bruge den nye teori på testproblemet. Vi definerede tidligere en persons formåen som den brøkdel testspørgsmål en testtager kan svare på ud af en stor, repræsentativ mængde spørgsmål indenfor testområdet. Situationen er derfor helt ækvivalent med den simplificerede model, vi anvendte for multiple-choice test i afsnit 5, baseret på gæt alene. Her repræsenterede parameteren p i binomialfordelingen sandsynligheden for, at man kunne gætte et svar til et givet spørgsmål. Nu er p et mål for, hvorvidt testtageren kan besvare spørgsmålet, altså et mål for testtagerens viden om området [13]. Vi bemærker, at i nærværende situation er p en relativ størrelse, der afgøres af testtageren. I gæt-situtationen var p en absolut størrelse, der afhang af testens udformning). Under alle omstændigheder er det indlysende, at antallet af rigtige svar i populationen af " potentielle besvarelser" (spørgsmål, der har den egenskab, at ved en "måling", d.v.s. besvarelse, opnås en realisation af en Bernoulli stokastisk variabel) i et bestemt testområde er binomialfordelt med den ukendte parameter p, testtagerens præcise formåen. Vi har altså et eksperiment, der består af

- En tilstrækkelig stor, repræsentativ population af spørgsmål.
- En bestemt testtager.
- En "måling" på poulationen relativt i forhold til testtageren, hvor antallet af rigtige svar i stikprøven, i.e. testen, er en sum af realisationer af Bernoulli stokastiske variable, altså en binomialfordelt stokastisk variabel.

Antag, at testen har n spørgsmål. Ud fra summen af realisationerne af de Bernoulli stokastiske variable (d.v.s. testscoren), der repræsenterer et testsvar, $s = \sum_{i=1}^{n} x_i$, kan vi da konstruere punktestimatet $\hat{p}$ givet ved

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{s}{n}$$

[12] Et sidste alternativ kunne være at anvende *hypotesetests*; en systematisk formulering af den statistiske inferens, groft sagt — jf. Engelhardt kap. 12

[13] Det diskuteres i afsnit 6.6, hvorvidt det er rimeligt at holde gæt- og videnssituationen så skarpt adskilt.

Bemærk, at vi for én test kun får et enkelt punktestimat. Dette vil sjældent være nok for at foretage et rimeligt punktestimat — medmindre testen indeholder urealistisk mange spørgsmål. Enten må man finde andre metoder — eller også må man gennemføre en række parallelle tests. D.v.s. man kunne give en række mindre tests over en kort periode — for hver af disse stikprøver kunne en værdi for $\hat{p}$ beregnes, og middelværdien $E(\hat{p})$ beregnes. P.g.a. tidsfaktoren/forsinkelsen, kan vi ikke forvente at dette vil være det samme, som at give hele testen på én gang.
Metoden giver naturligvis kun mening, såfremt $\hat{p}$ er en unbiased estimator, men i så fald burde den også give mere præcise estimater end rene punktestimater.
Dette kan vi dog ikke umiddelbart udtale os om. Skulle vi arbejde videre med modelkonstruktionen i P1, ville det være meget nærliggende at foretage en mere systematisk undersøgelse af de forskellige estimationsmetoder og deres anvendelighed netop i dette problem — og en undersøgelse af anvendeligheden af andre estimatortyper. Måske når man frem til den konklusion, at det i visse situationer, hvor testresultatet er af afgørende betydning for testtageren, kan være hensigtsmæssigt at bruge biased estimatorer (usikkerheden kommer testtageren til gode).
Det bør bemærkes, at det i ovenstående godt kan virke lidt løst, at vi taler om en binomialfordeling for antal rigtige blandt populationen, når vi overhovedet ingen anelse har om antalsparameteren n — og endda har afvist at diskutere den. Udfra en matematisk vinkel er det hensigtsmæssigt at lade $n \to \infty$ for det kunne med lidt omformulering føre til en mere entydig definition af en testtagers formåen. Vi ville få en kontinuert fordeling (en såkaldt *normalfordeling*), hvor p bliver erstattet med middelværdien μ o.s.v. Dette ville dog have krævet en meget omfattende introduktion til de adskillige formelle begreber, der ligger bag *kontinuerte stokastiske variable*.

6.4 Modelforudsætninger og anvendelighed

Vi har nu givet et konkret, foreløbigt udkast til, hvordan man kunne udbygge gætmodellen ved anvendelse af diskrete statistiske modeller. Denne tilgang er langtfra endegyldigt korrekt — og der er i hvert fald en række punkter, der kan påpeges. I første omgang bør vi kort forsøge at vurdere, hvorvidt der er tale om en *god model*. Vi kan i hvert fald konstatere, at den bygger på en række forudsætninger, nogle mindre realistiske end andre:

1. Testspørgsmålene kan betragtes som en stikprøve fra en population af mange, repræsentative og tilstrækkeligt vanskelige spørgsmål indenfor testområdet. Dette vil kun sjældent være tilfældet — i undervisningssituationer vil man f.eks. have en tilbøjelighed til at favorisere spørgsmål, som man erfaringsmæssigt ved, at eleverne har svært ved. Derfor er kriterierne for stikprøveudtagning ikke opfyldt, fordi sandsynligheden for at et spørgsmål bliver udtaget afhænger af dets relative vanskelighed.

2. Vi må forlange, at definition 6 (eller en variation heraf) giver mening — kort sagt, vi antager at vi med god nøjagtighed kan sætte et konkret, mere eller mindre fast

tal på en persons formåen. Denne antagelse kan formentlig kun forsvares, hvis man samtidig antager gennemgående psykologisk stabilitet.

3. Slutteligt — men dette er ret afgørende — at testen repræsenterer en stikprøve, der er tilstrækkelig stor til, at vi kan vurdere p og andre parametre med en optimal nøjagtighed.

4. Vi har valgt, stik modsat gætmodellen, at antage, at testtageren overhovedet ikke gætter. Urealistisk? Mere om dette i afsnit 6.6.

Hvad fortæller modellen os så, hvis vi i første omgang accepterer forudsætningerne? Resultaterne er først og fremmest af kvalitativ art

- Vi er blevet opmærksomme på den statistiske usikkerhed i testvurderingen
- Testforudsætningerne må betragtes som afgørende for en evt. models nøjagtighed
- Meget nøjagtige bedømmelser (ud fra forudsætningerne) kræver enten mange, eller meget store tests.

Disse må betragtes som de vigtigste, men er blot nogle af de slutninger, man kunne drage.

6.5 Gæt i test

Som tidligere lovet vil vi nu kort vende tilbage til en særlig interessant potentiel mangel ved modellen, antagelsen, at testtageren alene svarer udfra viden. Netop i en multiple-choice test er det nok ret usandsynligt for de fleste testtagere — medmindre der er ekstremt mange svarmuligheder pr. spørgsmål, kommer man ikke udenom, at sandsynligheden for at gætte rigtigt er relativt stor. Med den binomiale model for konsekvent gæt gennem en multiple choice test, jf. afsnit 5, så vi, at det faktisk er foruroligende mange rigtige, man kan få ved ren tilfældighed.
Dette problem er en af de alvorligste mangler ved multiple-choice, og vanskeliggør en eksakt vurdering af en testtager betydeligt. Problemet er af så konkret art, at vi allerede nu kan foreslå en række løsningsmuligheder:

- Vi kan vælge at ignorere det, eller bedre endnu — antage, at hvert spørgsmål i en givet test har så tilstrækkeligt mange svarmuligheder, at gæt virkelig kan ignoreres uden mærkbart tab af præcision.

- Der kan korrigeres for gæt v.h.a. simple korrektionsformler — formler, der straffer for forkerte svar. Anvendelsen af sådanne gør gætstrategien temmelig nyttesløs for testtageren; det kan bedst betale sig blot at springe spørgsmål, der ikke kan besvares ud fra viden, over.

- At medtage overvejelserne i modellen; forbedre den rent matematisk.

Sidstnævnte mulighed er oplagt stof i et P1-projekt om testproblemet; og selve formuleringen af problemet er faktisk slet ikke vanskelig.

Vi vælger ganske enkelt at betragte den stokastiske variabel S, testtagerens score, som summen $S = X + G$, hvor X er den score, eleven har opnået ved viden, og G er scoren fra gæt. Problemet er så bare — hvilken fordeling vil X have; og hvordan kan man bestemme sandsynligheden for, at et realiseret udfald af den stokastiske variabel G er mindre end n for et givet heltal n? Eller hvilke metoder er mest rimelige, hvis man samtidig skal tage højde for, at visse måske overhovedet ikke gætter?

7 Konklusion og projektforslag

Testproblemet har vist sig at være en sand jungle af overvejelser af vidt forskellig karakter indenfor vidt forskellige fagområder. I et naturvidenskabeligt orienteret P1-projekt ville det derfor være mest hensigtsmæssigt at foretage den forenklende konklusion, at f.eks. konstruktionsproblemet kan ignorereres; eller i hvert fald er af mindre betydning. Fokus kunne da lægges på videreudviklingen af testmodellen fra afsnit 6. En mulig overskrift kunne da være: **En statistisk model for multiple-choice test**. Vi har allerede nævnt mange af de konkrete problemer, som kunne være relevante — kan vi f.eks. forbedre modellen ved anvendelse af kontinuerte fordelinger? Hvordan kan vi tage højde for gætproblemet på den mest hensigtsmæssige måde? Og er det muligt at tilføje finesser, der gør vurderinger mere rimelige for testtageren?
Spørgsmålene er mange, og emnet er meget åbent. En indgående forståelse for de dybere dele af problemstillingen vil vi formentlig først få efter et indgående studie af mere avancerede statistiske metoder.

8 Litteraturliste

1. Antonius, Søren m.fl:
   Matematik A 2. udg.
   Danmark 1996, ISBN: 87-7783-841-6

2. Crocker, Linda m. fl.:
   Introduction To Classical & Modern Test Theory
   USA 1986, ISBN: 0-03-061634-4

3. Engelhardt, Bain:
   Introduction to probability and mathematical statistics 2. ed.
   USA 1992, ISBN: 0-534-92930-3

4. Gronlund, Norman F:
   Constructing achievement tests
   USA 1977, ISBN: 0-13-169235-6