NTNU

Norwegian University of
Science and Technology

A theoretical and empirical assessment of probabilistic multiple choice tests

Torunn Kval Bakken

Master of Science
Submission date: June 2017
Supervisor: Jarle Tufto, IMF

Norwegian University of Science and Technology
Department of Mathematical Sciences

A theoretical and empirical assessment of probabilistic multiple choice tests

NTNU

Torunn Kval Bakken

June 2017

Preface

This thesis concludes my master's degree in natural science with teacher education at the Norwegian University of Science and Technology (NTNU), with specialisation in mathematics and physics. The work on the thesis has been carried out during my ninth and tenth semester at the Department of Mathematical Sciences, from September 2016 to June 2017.

I would like to thank my supervisor Jarle Tufto for all his help and feedback during this process. I am very grateful for Truls Midthun who is always there for me when I need him the most. Last but certainly not least; my parents who always believe in me.

Torunn Kval Bakken

Trondheim, June 2017

Abstract

In this thesis, the probabilistic multiple choice test is analysed empirically and theoretically. It is suggested as an alternative to the traditional multiple choice test. The probabilistic multiple choice test has a long history. However, there are no known published research papers on the subject based on test results from Norwegian students.

We will compare the theoretical performance of the traditional and probabilistic multiple choice test. In addition, we will analyse their performance as estimators of level of knowledge. To estimate the level of knowledge, we want to be sure that the students estimate their abilities accurately. We will therefore analyse what may influence students to inaccurately estimate their abilities in a probabilistic multiple choice test. We call it overconfidence if the students overestimate their abilities, and conversely underconfidence if the students underestimate their abilities. Furthermore, we will take a closer look at score functions that could be suitable for the probabilistic multiple choice test.

This thesis is a quantitative research study of the probabilistic multiple choice test. The empirical research is done by a test administered to a group of students enrolled for the subject TMA4240 Statistics at NTNU, because of their knowledge of probability and statistics. Since the test was voluntary, an incentive to take the test was given in the form of a possibility to win one of two gift cards. The data provides a basis for analysis and inference on the probabilistic multiple choice test and the participant's overconfidence. The Dirichlet distribution is used to model the theoretical properties of the test. In addition, it is used to analyse the score functions that we evaluate the student's performance with.

The results show that the probabilistic multiple choice test with a logarithmic score function is an unbiased estimator of the level of knowledge of a participant. The participant's ability to correctly estimate their own level of confidence is influenced by their sex, the requirement of obtaining a minimum score, feedback and the score function their score is calculated by. We find that a good test for both female and male participants has a logarithmic score function and gives feedback during the test.

In the field of education, the probabilistic multiple choice method has the potential of redefining the use of multiple choice tests. First of all because it provides an accurate quantification of the student's level of confidence, and second of all by making the student's knowledge transparent to the educator.

Sammendrag

I denne oppgaven vil den probabilistiske flervalgsprøven analyseres empirisk og teoretisk. Den er foreslått som et alternativ til den tradisjonelle flervalgsprøven. Den probabilistiske flervalgsprøven har en lang historie. Det er likevel ingen kjente publiserte artikler om emnet basert på testresultater fra norske studenter.

Vi vil sammenlikne den teoretiske ytelsen av den tradisjonelle og probabilistiske flervalgsprøven. I tillegg vil vi analysere deres ytelse som estimatorer av kunnskapsnivå. For å estimere kunnskapsnivået vil vi være sikre på at studentene estimerer deres egen evne nøyaktig. Vi vil derfor analysere hva som kan påvirke studenter til å estimere sine evner unøyaktig under en probabilistisk flervalgsprøve. Vi kaller det overkonfidens hvis studentene overestimerer sine evner, og tilsvarende underkonfidens hvis studentene underestimerer sine evner. Videre vil vi ta en nærmere titt på score-funksjoner som kan være egnet for den probabilistiske flervalgsprøven.

Denne masteroppgaven er en kvantitativ undersøkelse av den probabilistiske flervalgsprøven. Den empiriske undersøkelsen er gjort ved å gi en prøve til en gruppe studenter som tar faget TMA4240 Statistikk på NTNU, på grunn av deres kunnskap om sannsynlighet og statistikk. Ettersom prøven var frivillig, ble et insentiv for å ta prøven gitt i form av en mulighet for å vinne et av to gavekort. Dataene danner et grunnlag for analyse og inferens om den probabilistiske flervalgsprøven, og deltakerens overkonfidens. Dirichlet-fordelingen er brukt til å modellere de teoretiske egenskapene ved prøven. I tillegg blir den brukt til å analysere score-funksjonene som vi evaluerer studentenes prestasjon med.

Resultatene viser at den probabilistiske flervalgsprøven med logaritmisk score-funksjon er en forventningsrett estimator av kunnskapsnivået til en deltaker. Deltakerens evne til å korrekt estimere deres eget kunnskapsnivå er påvirket av deres kjønn, kravet om å oppnå en minimum score, tilbakemelding og score-funksjonen deres score er regnet ut med. Vi finner at en god prøve for både kvinnelige og mannlige deltakere har en logaritmisk score-funksjon og gir tilbakemelding under prøven.

Innenfor utdanning har den probabilistiske flervalgsprøven potensialet til å redefinere bruken av flervalgsprøver. For det første fordi den gir en nøyaktig kvantifisering av elevens konfidensnivå og for det andre fordi den gjør elevens kunnskap synlig for læreren.

Table of Contents

Chapter 1

Introduction

In this thesis we will take a closer look at what is called the probabilistic multiple choice test. This is an interesting topic to investigate because multiple choice tests are frequently used in the Norwegian education system. Multiple choice tasks are even included in the national tests given to 5th, 8th and 9th graders. According to the Norwegian Directorate for Education and Training, the test results are used by the teachers to review their students development, and in guiding their own work. The municipalities and schools use the results from national tests as a foundation for further developing the quality of learning. Even researchers can access the results to use in their studies (Utdanningsdirektoratet, 2016).

Given that multiple choice tasks and tests are common and often used, it is important that they work in a satisfying manner. One of the main objectives of these tests is to give information about how well the students are performing and where the students need further guidance.

The Kansas silent reading test is acknowledged as the first multiple choice test used in a school, which was a reading test administered to selected children attending schools in Kansas. The exercises from this test were aimed to meet three qualifications. First of all, the interpretation of the exercise must be unique. That is to say, the exercises should be well defined and without ambiguous wording. Secondly, the answers must be right or wrong and nothing in between. Finally, they were to test the ability to obtain meaning from written material (Kelly, 1916).

In the traditional multiple choice test, the optimal decision for the participant is to choose the alternative that he/she finds most likely to be correct. In the case of complete ignorance, the participant might feel encouraged to guess which alternative is the correct one. Unless the participant can eliminate at least one alternative, the probability of successfully guessing the correct answer is then $1/m$, where m is the number of alternatives. Guessing is a serious flaw in the test design and should be avoided in order to properly evaluate the level of knowledge. The assessor has no way of knowing if a correct answer comes from a participant who understands the material, or from a participant who has been lucky and guessed the correct alternative. Penalty for incorrect answers or no penalty for leaving a question blank are some of the solutions to discourage participants from guessing

(Espinosa and Gardeazabal, 2010).

Some authors (e.g. Bernardo (1998), Ben-Simon, Budescu, and Nevo (1997)) have proposed that probabilistic multiple choice tests could be the answer to some of the shortcomings of traditional multiple choice tests. In probabilistic multiple choice tests the participants report a level of confidence for each alternative. The participant can then report a complete lack of knowledge by reporting $(1/m, 1/m, \ldots, 1/m)$. A participant with perfect knowledge can report a distribution as for example $(0, 1, 0, \ldots, 0)$, indicating that alternative (b) is the correct alternative. The reported level of confidence on each alternative can take any real value on an $(m-1)$-simplex, in contrast to binary true/false alternatives in traditional multiple choice tests (Bernardo, 1998).

The reported level of confidence is not necessarily a correct estimate of the participant's true knowledge of a topic. The participant may overestimate their abilities, thus reporting a level of confidence much higher than they should. Conversely, their estimate may also be too conservative, thus making the reported level of confidence too low. We will respectively call this behaviour overconfidence and underconfidence. This is an important part of the probabilistic multiple choice test, because as mentioned previously, the tests are supposed to provide information about the student. If the information given by the participant is not correct, the information is less valuable and more difficult to interpret directly.

The probabilistic multiple choice test has been implemented at the University of Stavanger by Bratvold (Unpublished). This has provided an interesting basis for the work carried out in this thesis.

We will attempt to analyse the performance of the probabilistic multiple choice test compared to the traditional multiple choice test. An important part of the probabilistic multiple choice test is the score function used to calculate the obtained score. We want to find a score function that will provide the best estimate of the participant's knowledge. We will also try to find a statistical model of what might influence the participants to incorrectly estimate their level of confidence. The combined analyses may be used to propose a probabilistic multiple choice test that performs well.

The outline of the thesis is as follows: In Chapter 2 Bayesian decision theory is provided as a framework for further analysis. In Chapter 3 the method for the test and data are presented. In Chapter 4 the method and results of the theoretical analysis is presented. In Chapter 5 a method for analysing the empirical data is introduced and in Chapter 6 the results of this analysis is presented. In Chapter 7 we discuss our results and suggestions for further work. All data analysis is done with R (R Core Team, 2016).

Chapter 2

Score functions

2.1 Bayesian decision theory

Any situation where choices are to be made among alternative courses with uncertain consequences are decision problems (Bernardo and Smith, 1994, p. 16-19). For a participant, each question in a probabilistic multiple choice test is therefore a decision problem. First, some general elements of the decision problem must be defined:

- A set of events, E
- A set of consequences, C
- A set of options/acts, A
- $\leq$ is a preference order, taking the form of a binary relation between the elements of A

In this thesis we want to analyse the decision problem quantitatively. In order to do so, we assume that the participants act rational when faced with a decision problem. Rationality is a principle of what is called quantitative coherence. By this, we mean that a preference order must be quantitatively precise and based on logical forms of behaviour (Bernardo and Smith, 1994, p. 23).

A prescription of what constitutes coherent behaviour can be made, but this does not imply that participants automatically *behave* coherent. It is merely a framework for analysing the decision problem. The three axioms, as stated by Bernardo and Smith (1994, p. 23-26), that prescribes rational behaviour is

- Axiom 1: comparability of consequences and dichotomised options
- Axiom 2: transitivity of preferences
- Axiom 3: consistency of preferences.

Axiom 1 states that the participant must be able to distinguish between the consequences in the decision problem at hand. This means that there are at least two consequences, for example c_1 and c_2, such that one of them is preferred over the other. The same applies for the dichotomised options. Thus, for two events with corresponding consequences, there exist at least two options where one of them is preferred over the other (Bernardo and Smith, 1994, p. 23-24).

Axiom 2 simply states that if option 1 is not preferred over option 2, and option 2 is not preferred over option 3, then option 1 is obviously not preferred over option 3 (Bernardo and Smith, 1994, p. 24-25).

Axiom 3 states that a preference pattern in consequences is not affected by knowing more about the uncertain events. Also, a preference pattern in consequences, and a corresponding preference pattern in options, will ultimately decide which event is evaluated to be most probable. Meaning that if an individual prefers to win rather than lose, and a choice of A is more preferable than choosing B, then the individual would evaluate B as more likely. Lastly, the axiom states that if two situations are such that the outcome of the first is not preferred over the second, then the second situation is preferable overall (Bernardo and Smith, 1994, p. 25-26).

In order to evaluate decision problems quantitatively we make an assumption of the existence of standard events (Bernardo and Smith, 1994, p. 29-30). A standard event can be compared with the use of standard units of measurement and the quantification is the numerical value of that unit. A person is weighed in kilograms and it is quantified by a numerical value, e.g. 60 kg. For the decision problem, a standard event is for example that we estimate an event as equally likely as a coin flip, and the quantification of this is 0.5.

Thus, a rational participant can state their degree of beliefs for a set of events as a probability distribution. According to Bernardo and Smith (1994, p. 33-35), any probability in this distribution is then a personal degree of belief. It is a numerical value of the personal uncertainty relation between events, and will for the rest of this thesis be referred to as level of confidence. We use the notation "level of confidence" because, during a test, the participant will evaluate the alternatives by his/her confidence that they are correct.

By applying the principle of quantitative coherency, a utility can be defined for the set of consequences. The utility is a function that maps the consequence of a decision problem to a numerical value. Assuming the utility gain is positive, a rational participant will have a preference pattern that maximises the expected value of the utility (Bernardo and Smith, 1994, p. 70-71). For the probabilistic multiple choice test, let $(\delta_1, \delta_2, \ldots, \delta_m)$ be the set of alternatives for a question, where m is the number of alternatives. Let $\boldsymbol{r} = \{(r_1, \ldots, r_m), r_i \geq 0, \sum_i r_i = 1\}$ be the individual's reported probability distribution over the set of possible answers. These are the decision variables in a probabilistic multiple choice test. For now, there is no reason to assume that the probability distribution, $\boldsymbol{r}$, accurately describes the true level of confidence of the participant. Therefore, we assume that $\boldsymbol{p}$ is the participant's honest probability distribution, where p_i is the probability the participant perceives alternative i to be correct. Each question has its own

probability distribution. The expected value of utility for the test is the expected score,

$$\bar{u}(\boldsymbol{r}) = \sum_{i=1}^{m} u(\boldsymbol{r}, \delta_i) p_i,$$

where $u(\boldsymbol{r}, \delta_i)$ is the score awarded to a participant who marks the probability distribution $\boldsymbol{r}$ when the correct answer is δ_i (Bernardo, 1998, p. 4-5).

Ordinarily, the utility function is concave for monetary values, but is approximately linear for small amounts of money. For the test in this thesis, the monetary gain is kept small, resulting in an approximately linear utility function in the score function.

2.2 Different score functions

In order to encourage honesty, a score function should have a maximum expected value if and only if the participant sets $\boldsymbol{r} = \boldsymbol{p}$. A score function that satisfies this property is called a proper score function. Another property of the score function, that is preferable in pure inference situations, is that the score function is local. Pure inference problems are situations where we are only concerned with the truth. The local score function is therefore purely a function of the probability assigned to the correct alternative. The score function can provide a basis to quantify the participant's level of knowledge in a multiple choice test (Bernardo and Smith, 1994, p. 70-72).

Let the row vector $\mathbf{d} = (d_1, \ldots, d_m)$ be a vector of indicator variables indicating which answer alternative is correct. If alternative i is correct, then $d_i = 1$ and $d_j = 0$, for $i \neq j$. The stochastic variable $\boldsymbol{d} = (d_1, \ldots, d_m)$ is multinomially distributed with $\mathbf{d} \sim \text{Mult}(1, \mathbf{p})$. From known relations of the multinomial distribution, $E(d_i) = p_i$, $E(d_i^k) = p_i$, where $k = 1, 2, \ldots$ and $E(d_i d_j) = 0$ when $i \neq j$ (Bernardo and Smith, 1994, p. 433).

2.2.1 Simple score function

Neyhart and Abrassart (1984, p.74) suggest a simple score function with range $[0, 1]$. In this case, the score obtained for each question is equal to the probability reported by the participant for the correct alternative. With the notation introduced above, the simple score function is

$$s(\boldsymbol{r}, \boldsymbol{d}) = \sum_{i=1}^{m} r_i d_i. \tag{2.1}$$

We can now find the simple conditional expected score,

$$E(s(\boldsymbol{r}, \boldsymbol{d}) | \boldsymbol{p}) = E\Big(\sum_{i=1}^{m} r_i d_i \Big) = \sum_{i=1}^{m} E(r_i d_i) = \sum_{i=1}^{m} r_i p_i.$$

We consider the expectation conditional on $\boldsymbol{p}$ as we will later model $\boldsymbol{p}$ as a random variable in Chapter 3.

Notice how $E(s(\boldsymbol{r}, \boldsymbol{d})|\boldsymbol{p})$ varies linearly with $r_1, \ldots, r_m$. Thus, it follows that in order to maximise the expected score, the participant should set $r_i = 1$ for the alternative with the largest p_i, and all other $r_i = 0$. We can therefore conclude that the simple score function is not proper.

2.2.2 Quadratic score function

The quadratic score function with range $[0, 1]$ is according to Winkler and Murphy (1968, p. 754) defined as

$$Q(\mathbf{r}, \mathbf{d}) = 1 - \sum_i (r_i - d_i)^2. \tag{2.2}$$

We can find that the quadratic conditional expected score is

$$\begin{aligned} E(Q(\boldsymbol{r}, \boldsymbol{d})|\boldsymbol{p}) &= E\left(1 - \sum_{i=1}^{m} (r_i - d_i)^2 | \boldsymbol{p}\right) \\ &= 1 - \sum_{i=1}^{m} E(r_i^2 - 2r_i d_i + d_i^2 | \boldsymbol{p}) \\ &= 1 - \sum_{i=1}^{m} E(r_i^2|p) + 2\sum_{i=1}^{m} E(r_i d_i|\boldsymbol{p}) - \sum_{i=1}^{m} E(d_i^2|\boldsymbol{p}) \\ &= 1 - \sum_{i=1}^{m} r_i^2 + 2\sum_{i=1}^{m} r_i p_i - \sum_{i=1}^{m} p_i \\ &= \sum_{i=1}^{m} p_i^2 - \sum_{i=1}^{m} (r_i - p_i)^2. \end{aligned} \tag{2.3}$$

Note that the participant, in order to maximise his/her expected subjective score, must set $\boldsymbol{r}$ equal to $\boldsymbol{p}$. Thus the quadratic score function is proper.

2.2.3 Spherical score function

The spherical score function $S(\mathbf{r}, \mathbf{d})$ with range $[0, 1]$ is according to Winkler and Murphy (1968, p. 754) defined as

$$S(\boldsymbol{r}, \boldsymbol{d}) = \frac{\sum_{i=1}^{m} r_i d_i}{\left(\sum_i r_i^2\right)^{1/2}} \tag{2.4}$$

Hence, the spherical conditional expected score is

$$E(S(\boldsymbol{r},\boldsymbol{d})|\boldsymbol{p}) = \frac{\sum_{i=1}^{m} p_i r_i}{\left(\sum_{i=}^{m} r_i^2\right)^{1/2}}.$$

From Cauchy-Schwarz's inequality (Casella and Berger, 2002, p. 187),

$$\sum_{i=1}^{m} p_i r_i \leq \left(\sum_{i=1}^{m} r_i^2\right)^{1/2} \left(\sum_{i=1}^{m} p_i^2\right)^{1/2}$$

thus,

$$E(S(\boldsymbol{r},\boldsymbol{d})|\boldsymbol{p})) \leq \left(\sum_{i=1}^{m} p_i^2\right)^{1/2}$$

with equality holding if and only if $r_i = kp_i$ for all i, where k is a constant. The participant's expected score is maximised if equality holds. Since $\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} r_i = 1$, k equals one and the spherical score function is proper.

2.2.4 Logarithmic score function

The logarithmic score function with range $[-\infty, 0]$ is according to Winkler and Murphy (1968, p. 754-755) defined as

$$L(\boldsymbol{r},\boldsymbol{d}) = \sum_{i=1}^{m} d_i \ln(r_i). \tag{2.5}$$

The logarithmic conditional expected score is

$$E(L(\boldsymbol{r},\boldsymbol{d})|\boldsymbol{p}) = \sum_{i=1}^{m} p_i \ln r_i \tag{2.6}$$

Maximising $E(L(\boldsymbol{r},\boldsymbol{d})|\boldsymbol{r}))$ in (2.6) is equivalent to maximising

$$E(L(\boldsymbol{r},\boldsymbol{d})|\boldsymbol{p}) - \lambda\left(\sum_{i=1}^{m} r_i - 1\right) = \sum_{i=1}^{m} p_i \ln r_i - \lambda\left(\sum_{i=1}^{m} r_i - 1\right), \tag{2.7}$$

since$\sum_{i=1}^{m} r_i = 1$ and λ is a Lagrange multiplier. Differentiating (2.7) with respect to r_i and setting the result equal to zero yields

$$r_i = \frac{1}{\lambda} p_i.$$

Since $\sum_{i=1}^{m} r_i = \sum_{i=1}^{m} p_i = 1$, λ equals one, the optimal decision is to set $\mathbf{r} = \mathbf{p}$, and the logarithmic score function is proper.

2.2.5 Summary

We see directly from (2.1), (2.2), (2.4) and (2.5) that only the simple and the logarithmic score functions are local.

The different score functions can be made more comparable by linear transformations such that a score equal to 0 corresponds to complete ignorance and score equal to 1 corresponds to perfect knowledge. The score functions that are transformed to the mutual range are the quadratic and logarithmic score function, which are the two score functions used in the empirical study in this thesis. In order to find the proper linear transformation of the logarithmic and quadratic score functions we multiply by a constant and add a constant:

$$Q(\mathbf{p}, \mathbf{d}) = c_{q0} + c_{q1} \sum_{i=1}^{m} (p_i - d_i)^2$$

$$L(\mathbf{p}, \mathbf{d}) = c_{l0} + c_{l1} \sum_{i=1}^{m} (d_i \log(p_i)).$$

We want the score function to have value 0 when $\mathbf{p} = (1/m, \ldots, 1/m)$, and value 1 when $p_i = 1$ for alternative i, where $d_i = 1$. Therefore,

$$Q(\mathbf{p}, \mathbf{d}) = c_{q0} + c_{q1} \sum_{i=1}^{m} (1/m - d_i)^2 = 0$$
$$Q(\mathbf{p}, \mathbf{d}) = c_{q0} + c_{q1}((m-1)(0-0)^2 + (1-1)^2) = 1,$$

implying that

$$c_{q0} = 1$$
$$c_{q1} = -\frac{m}{m-1},$$

and

$$L(\mathbf{p}, \mathbf{d}) = c_{l0} + c_{l1} \sum_{i=1}^{m} (d_i \log(1/m)) = 0$$
$$L(\mathbf{p}, \mathbf{d}) = c_{l0} + c_{l1} \log(1) = 1,$$

implying that

$$c_{l0} = 1$$
$$c_{l1} = -\frac{1}{\log(1/m)} = \frac{1}{\log(m)}.$$

In Table 2.1 the score functions are listed with their respective properties.

Score function	Function	Range	Proper	Local
Simple	$s(\mathbf{r}, \mathbf{d}) = \sum_{i=1}^{m} r_i d_i$	$[0, 1]$	No	Yes
Quadratic	$Q(\mathbf{p}, \mathbf{d}) = \left[1 - \frac{m}{m-1} \sum_{i=1}^{m} (p_i - d_i)^2\right]$	$[-\frac{m+1}{m-1}, 1]$	Yes	No
Spherical	$S(\mathbf{p}, \mathbf{d}) = \frac{\sum_{i=1}^{m} p_i d_i}{\left(\sum_i p_i^2\right)^{1/2}}$	$[0, 1]$	Yes	No
Logarithmic	$L(\mathbf{p}, \mathbf{d}) = 1 + \frac{1}{\log(m)} \sum_{i=1}^{m} (d_i \log(p_i))$	$[-\infty, 1]$	Yes	Yes

Table 2.1: Summary table of the four different score functions

Chapter 3

Theoretical analysis

We will theoretically analyse some properties of probabilistic and traditional multiple choice tests. We assume that the probabilities reported by the participants are unbiased, i.e. the different answer alternatives turns out to be correct with probabilities equal to the probabilities reported by the participants.

The probabilities are the participant's level of confidence. The score functions can then be viewed as estimators of the participant's level of knowledge. The level of knowledge a participant has for different questions can vary in many ways. In this thesis we have chosen to model it by the Dirichlet distribution, as it has some of the properties we see in the test regime.

A random vector $\boldsymbol{x}$ is Dirichlet distributed, $\mathrm{Dir}(\boldsymbol{\alpha})$, for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, $\alpha_i > 0$, if its sample space is $x_i > 0$, $x_m = 1 - \sum_{i=1}^{m-1} x_i$, and its density is

$$f(x_1, \ldots, x_{m-1}) = \frac{\Gamma\Big(\sum_{i=1}^{m} \alpha_i\Big)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \prod_{i=1}^{m} x_i^{\alpha_i - 1}$$

(Bernardo and Smith, 1994, p. 134). First of all, the random vector $\boldsymbol{x}$ is the open $(m-1)$-dimensional simplex and corresponds to the probability distribution, $\boldsymbol{p}$, given by the participants for each question. Secondly, the concentration parameters $\alpha_1, \ldots, \alpha_m$ are used as concentration of knowledge for each question, where we assume $\alpha_1 = \ldots = \alpha_m = \alpha$.

The distribution is used to model how the participant's knowledge is distributed between different questions. For $\alpha \to \infty$ the probability density will be concentrated in the point $(p_1, \ldots, p_m) = (1/m, \ldots, 1/m)$. When the p_i's are distributed like this, between different randomly selected questions, we have a participant with no knowledge. Conversely, in the limit when $\alpha \to 0$ the probability density is concentrated in each of the m corners of the $(m-1)$-simplex, i.e. the points $(1, 0, \ldots, 0, 0), (0, 1, 0, \ldots, 0), \ldots, (0, 0, \ldots, 0, 1)$ with probability $1/m$ in each corner. This corresponds to a person with perfect knowledge. When $\alpha \approx 1$, this will correspond to something in between these two extrema. In this framework we assume that the concentration of knowledge for each participant is constant

throughout the test. A different assumption could be that a participant has perfect knowledge about some questions and no knowledge for others. This would lead to a different relationship between expectation and variance. This relationship, however, is not investigated any further in this thesis. The Dirichlet model is an attempt to model this variation somewhat realistically.

3.1 Derived expressions for subjective expected score and variance

A theoretical analysis of the traditional multiple choice test (MCT) and of the score functions suggested for the probabilistic MCT, can provide a basis to determine which is most effective. Key elements in this analysis are expected score and the variance of this score.

We have already found the conditional expected score for both the quadratic and the logarithmic score function. In order to find the expected score, and the variance of score analytically, we need a distribution that can be used for a probabilistic multiple choice test. From now on we assume that $\boldsymbol{p} \sim \text{Dir}(\alpha, m)$. We can find the unconditional expected score and variance by the law of total expectation and total variance for any given score function U (Kendall et al., 1991, p. 66). Thus,

$$E_p(U(\boldsymbol{p}, \boldsymbol{d})) = E_p(E(U(\boldsymbol{p}, \boldsymbol{d})|\boldsymbol{p}))$$

is the expected score for score function U, and

$$\text{Var}_p(U(\boldsymbol{p}, \boldsymbol{d}) = \text{Var}_p(E(U(\boldsymbol{p}, \boldsymbol{d})|\boldsymbol{p})) + E_p(\text{Var}(U(\boldsymbol{p}, \boldsymbol{d})|\boldsymbol{p}))$$

is the variance of the score for score function U, where

$$\text{Var}(U(\boldsymbol{p}, \boldsymbol{d})|\boldsymbol{p}) = E(U(\boldsymbol{p}, \boldsymbol{d})^2|\boldsymbol{p}) - E(U(\boldsymbol{p}, \boldsymbol{d})|\boldsymbol{p})^2.$$

For the traditional multiple choice test a participant will choose alternative i with the largest probability p_i from a vector $\boldsymbol{p}$. Under the assumption made in the beginning of the chapter, the probability of guessing the correct alternative is thus $\max(p_1, \ldots, p_m)$ conditional on $\mathbf{p}$. The unconditional probability is then given by the law of total probability (Kendall et al., 1991, p. 288-289) by integration over the vector $\boldsymbol{p}$ of the Dirichlet distribution,

$$\int \cdots \int \max(p_1, \ldots, p_m) f(p_1, \ldots, p_{m-1}) dp_1 \cdots dp_{m-1}.$$

To the best of my knowledge there is no closed form solution for this integral. Therefore, we will find the expectation and variance numerically, by the use of the binomial distribution and Monte Carlo integration. The probability is estimated by sampling from the Dirichlet distribution, we find the maximum value of each sample and take the average of them. We then have an estimated value for the probability that the correct answer is chosen. The score is then binomially distributed with

this estimated probability and number of questions as parameters. For R-code, see Appendix D.2.

For the adjusted range of the logarithmic and quadratic score function, with $n = 27$ questions, $m = 4$ alternatives, the expectation and variance of the total score are, respectively

$$E_p(Q(\boldsymbol{p}, \boldsymbol{d})) = -\frac{n}{3} + n\frac{4(\alpha+1)}{3(m\alpha+1)}$$

$$\mathrm{Var}_p(Q(\boldsymbol{p}, \boldsymbol{d})) = n\frac{16(\alpha+1)}{9}\left(- \frac{(2m+1)\alpha+3}{(m\alpha+1)^2} - \frac{3(m\alpha^2+(m+4)\alpha+6)}{(m\alpha+1)(m\alpha+2)(m\alpha+3)} + \frac{2((m+2)\alpha+6)}{(m\alpha+1)(m\alpha+2)}\right)$$

and

$$E_p(L(\boldsymbol{p}, \boldsymbol{d})) = n + n\frac{(\psi(\alpha+1) - \psi(m\alpha+1))}{\log(m)}$$

$$\mathrm{Var}_p(L(\boldsymbol{p}, \boldsymbol{d})) = n\frac{(\psi_1(\alpha+1) - \psi_1(m\alpha+1))}{\log(m)^2}.$$

The range of the traditional MCT is adjusted such that the expected score is 0 when a participant is ignorant and 1 when a participant has perfect knowledge. Thus, the expected score and variance is

$$E_p(T(\boldsymbol{p}, \boldsymbol{d})) = n\frac{(m\hat{p}-1)}{(m-1)}$$

$$\mathrm{Var}_p(T(\boldsymbol{p}, \boldsymbol{d})) = n\frac{m^2\hat{p}(1-\hat{p})}{(m-1)^2}.$$

See Appendix B for a more in-depth mathematical derivation of these expressions and Appendix D.2 for R-code.

3.2 Score functions as estimators of knowledge

A commonly used measure of information is Shannon's expected information. According to Bernardo and Smith (1994, p. 79-81), this expected information of a discrete distribution given by $\boldsymbol{p}$, is defined as

$$E(I(\boldsymbol{p})) = \sum_{i=1}^{m} E(\log p_i) = \sum_{i=1}^{m} p_i \log p_i. \tag{3.1}$$

Within the framework of decision theory, maximising the Shannon information is a particular instance of maximising expected utility (Bernardo and Smith, 1994, p. 81). This particular instance is the pure inference experiment, which is an experiment where we are only interested in the truth. The experiment in this case is the quiz, where we are only interested in the correct alternatives (the truth). In the context of multiple choice tests, the Shannon information is arguably a reasonable measure of the level of knowledge that a participant has about a particular question. Obtained scores based on different score functions can be viewed as estimates of the expected information and the score functions as such estimators can be compared by assessing their bias, variance and mean square error.

A good estimator is usually an unbiased estimator, however even though an estimator is unbiased or the bias is small, the variance could still be large, thus making the estimator unfit. We will use the mean square error as a measure of the tradeoff between bias and variance of an estimator. The mean square error (MSE) of an estimator W of a parameter θ is the function of θ defined by $E(W-\theta)^2$, i.e.

$$E(W-\theta)^2 = \operatorname{Var} W + (EW-\theta)^2 = \operatorname{Var} W + (\operatorname{Bias} W)^2$$

(Casella and Berger, 2002, p. 330).

3.3 Analysis

In Figure 3.1 the expected score and variance are plotted against each other. The traditional MCT has consistently a larger variance than the probabilistic MCT with logarithmic score function, and slightly smaller variance than the quadratic score function when the expected score is high ($\approx$ 18 or higher). As expected, the variance of the score for the traditional MCT is large when the participant is highly misinformed. We see this where the expected score is close to 0. Misinformed means that the participant is sure that the incorrect alternative is the correct one. This result is intuitively clear since the participant would consistently put a high probability for the wrong alternatives. The large peak at around $E_p(T(\boldsymbol{p}, \boldsymbol{d})) = n/m \approx 7$ is also as expected since the participant is highly unsure and will therefore guess the alternative, thus making the variance the largest.

The probabilistic MCT shows promising result, both for the quadratic and logarithmic score function. The logarithmic score function has, however, consistently smaller or equal variance to the quadratic score function. The variance is largest when the expected score is around $E_p = n/2 = 13.5$. Intuitively we would expect the variance to be small for both correctly informed and misinformed participants. The reason is that the participants would personally be quite sure and consistent about what they think is correct. The variance is intuitively at its largest for participants who do not know what is correct. The Dirichlet distribution appears to be a good fit for the properties of multiple choice tests.

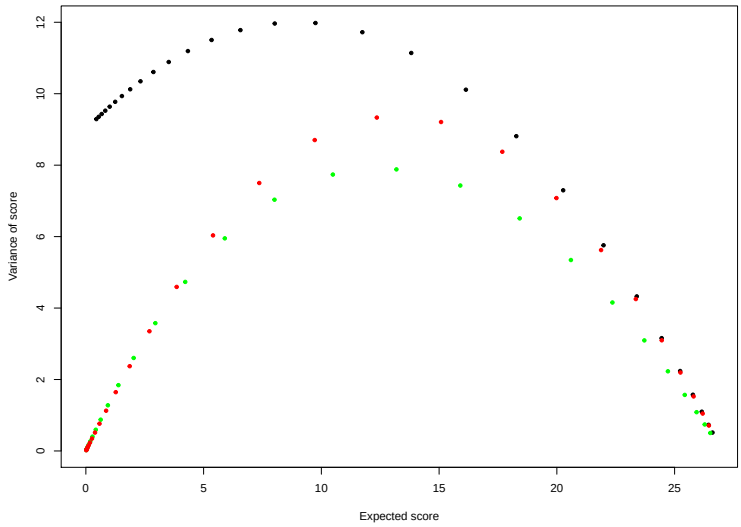

Figure 3.1: Plot of the expected score and variance for the probabilistic MCT when we use logarithmic (green) and quadratic (red) score function, and the traditional MCT (black). The variance of the traditional MCT is the largest of the three. The probabilistic MCT with quadratic score function has slightly larger variance than with logarithmic score function.

The logarithmic score function is an unbiased estimator of the information because the expected score from the logarithmic score function is the same as the information, $E(L(\boldsymbol{p}, \boldsymbol{d})|\boldsymbol{p}) = \sum_{i=1}^{n} p_i \log p_i$. The quadratic score function and the traditional MCT are not unbiased estimators of the information. To evaluate the bias of the quadratic score function and the traditional MCT, a plot of the two different biases are shown in Figure 3.2 and Figure 3.3 respectively. From Figure 3.2, the quadratic score function appears to be nearly unbiased towards the far right and far left. In the middle of the plot we see some sign of the quadratic score function being a little biased.

From Figure 3.3, we see that the traditional MCT is biased, and therefore does not appear to be a good estimator of level of knowledge. The bias is seen in the non-linear relationship between the probabilistic MCT with logarithmic score function and the traditional MCT. To investigate further, the mean square error is plotted for the unbiased estimator (logarithmic) and the two biased estimators (quadratic and traditional). The mean square error of the logarithmic score function is obviously just the variance of the score function since it is an unbiased estimator of knowledge (information).

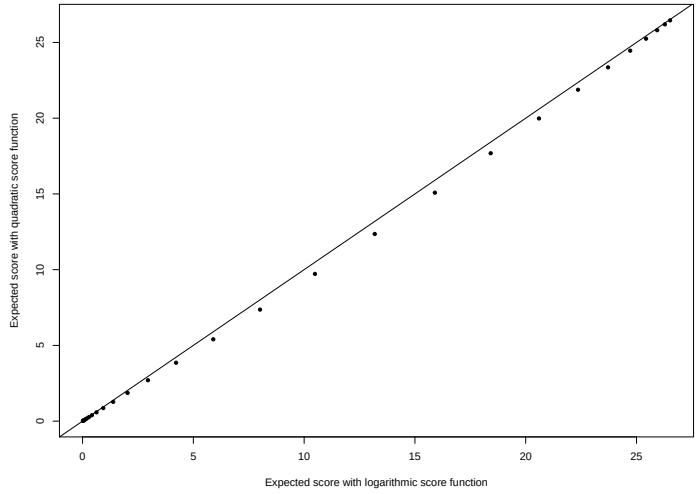

Figure 3.2: Plot of the bias between the expected score of the logarithmic and quadratic score function. The quadratic score function appears to be close to unbiased towards the far left and right. Some signs of a little bias in the middle.

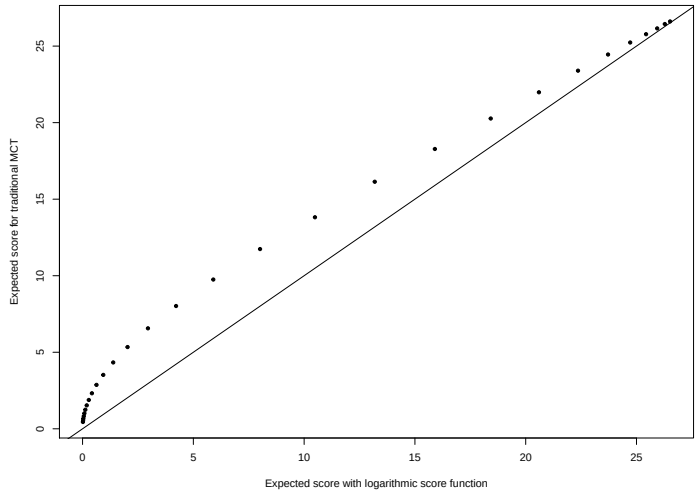

Figure 3.3: Plot of the bias between the expected score of the logarithmic score function and the traditional MCT. The traditonal MCT appears to be biased.

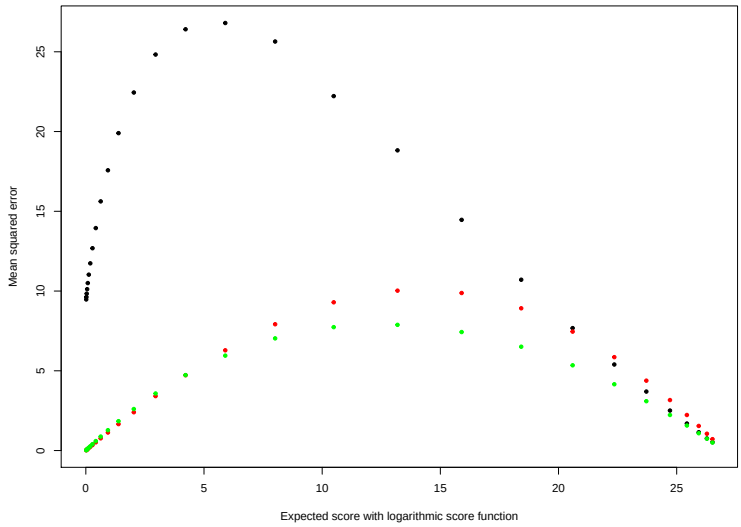

Figure 3.4: Plot of the mean square error of the estimators of information, namely the logarithmic score function (green), the traditional MCT (black) and the quadratic score function (red).

From Figure 3.4 it is clear that the traditional MCT is not only a biased estimator, but also has the largest MSE out of the three, except for high expected score. For high expected scores, there is only a small difference between the three estimators, where the logarithmic score function still outperforms the other two. The MSE of the traditional MCT is relatively large for low to average expected scores. When the expected score is ≈ 5, the MSE of the traditional MCT is at an all time high where the MSE is $4-5$ times larger than the MSE of the quadratic and logarithmic. Thus, for the traditional MCT to be as accurate as the probabilistic MCT, the test has to have $4-5$ times as many questions.

Chapter 4

Empirical study

4.1 Methods

4.1.1 Participants

The participants in this experiment were 89 students at NTNU, taking the class TMA4240 Statistics. The test was voluntary, but had a prize for two winners (gift card of value 500 NOK). Each participant were randomly chosen for quadratic/logarithmic score function, feedback/no feedback and minimum score/maximum score, see Appendix E. The two prizes were divided between the minimum score group and the maximum score group. The minimum score group had to get a "passing" grade, which was equivalent to getting a score of 10.8, and the winner was chosen at random from the participants who managed to get at least the minimum score. The winner in the maximum score group was chosen with a probability dependent on the obtained score, thus the higher score the higher probability of winning the prize. The participants were made aware of which score function they were scored by, if they had to get a minimum score and if they received feedback, before the test started. A web page (https://wiki.math.ntnu.no/probquiz) was created for the participants with a description of the two score functions, and some theoretical background for why they should report their level of confidence honestly.

The reason for picking participants taking a statistical course was to make sure that the participants taking the test understood the theoretical background for the score functions, and the probabilistic part of the multiple choice test. A certain level of knowledge about probability is required in order to take the test. The participant has to be able to grasp the concept of why reporting their honest level of confidence will maximise the expected score. In order to accurately estimate how much the participant knows, it is important that the probabilistic method is understood such that it does not influence the variance of the score (Poizner, Nicewander, and Gettys, 1978, p. 84).

4.1.2 Factors, levels and measurements of data

Factors are used to denote any treatment or therapy applied to the subjects being measured, or any relevant feature characteristic of those subjects. Different versions, extents, or aspects of a factor are referred to as levels. In this case there are different factors with different levels. Whenever measurements are made, they can be classified as either quantitative or qualitative measurements. Quantitative measurements are for example height in m or weight in kg. A qualitative measurement is for example to state your mood as happy or sad. Two measurements are said to be similar if their units are the same and dissimilar otherwise (Larsen and Marx, 2014, p. 449-452). The probability distribution marked by the participants for each question is therefore a quantitative measurement of similar units.

Prior to the experiment, four categories were chosen for further study:

- Sex with three groups/levels: male/female/not chosen
- Score function is binary: logarithmic/quadratic
- Feedback is binary: true/false
- Minimum score is binary: true/false

4.1.3 About the test

The test in this thesis was based on a wide range of knowledge areas. This was a conscious choice to make sure the participants would be able to answer regardless of their interests. Therefore, the questions were based on different areas of common quiz-related questions. The reason for providing the test in a quiz-based form was to make sure that as many participants as possible would take the test, since it was not possible to make the test a mandatory part of the class. Therefore, a middle ground was met by making the test short and with a small monetary reward for two winners. The test was short because the participants could be less likely to finish the test if it was long and demanding. The small monetary reward was to provide some incentive to do the test at all.

Upon taking the test, each alternative and question were permuted in a random order such that the participants could not easily compare their tests. In addition, the random permutation was important in order to analyse possible trends in overconfidence over time. If the overconfidence becomes smaller during the test, it is easier to see the effect if the questions and alternatives are given to the participants in a random order.

Every question and alternative used for the test that was performed can be found in Appendix G. For simplicity, alternative (a) is the correct alternative for every question. This will not matter while the test is carried out since the order of the alternatives and questions are randomly permuted. The layout of the online test, from now on referred to as the quiz, is shown in Figure 4.1.

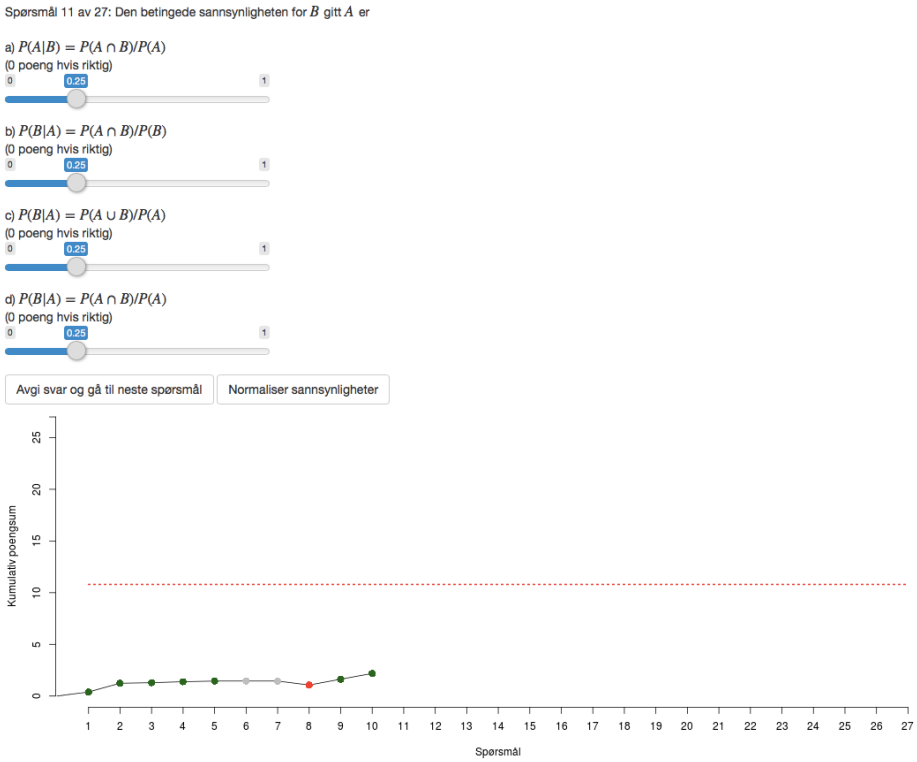

Figure 4.1: Screenshot of the web app where the quiz was issued with participant being scored by the logarithmic score function, receiving feedback and obtaining at least a minimum score of 10.8 (see `https://jtufto.shinyapps.io/multiple-choice/`).

4.2 Data

Because the experiment involved students volunteering to take a somewhat time consuming test with only a modest expected benefit, some participants may not report their own subjective probabilities in a truthful way. Another possibility is cheating, that is, students finding the correct answers to different questions via various online resources while taking the test. As an attempt to exclude such cases from the data analysis, outliers were identified as in the following paragraphs.

Information about a participant is limited, and they had the possibility of being completely anonymous. As a consequence of this, it was optional for the participants to state their gender. Each participant could therefore choose to set their sex to be "not chosen" or "female"/"male".

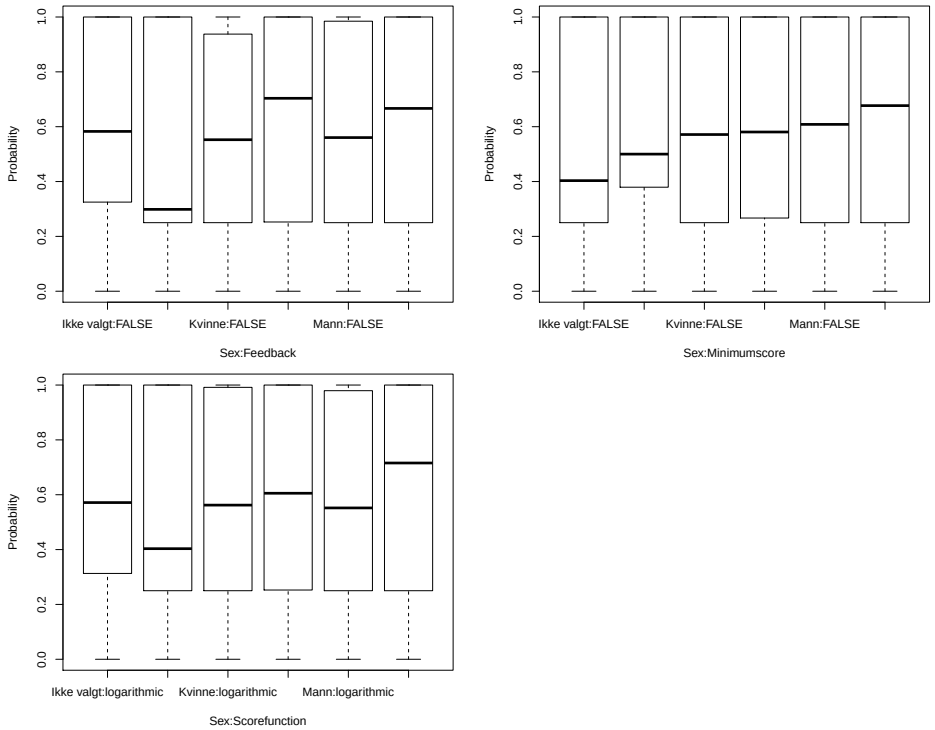

Figure 4.2: Box-plot of probabilities with covariates Sex: Feedback, Sex:Minimum score and Sex: Score function. In the upper left corner we have six boxes for the interaction covariate Sex: Feedback. For the first box we have sex "not chosen" and no feedback, second box we have sex "not chosen" and feedback, third we have sex "female" and no feedback, fourth we have sex "female" and feedback, fifth we have sex "male" and no feedback and sixth we have sex "male" and feedback. We have the same order of the boxes for the boxplots in the upper right corner and lower left corner.

From Figure 4.2 we can see that the sex "not chosen" display behaviour that is opposite of the other sexes. Male and female participants have an increase in

probability value when they get feedback or their score function is quadratic. Participants with unknown sex display a decrease in probability for the same covariates, notice especially how it plummets when feedback is TRUE. This type of behaviour seems counter-intuitive as participants with unknown sex, naturally must be either male/female or transgender, i.e. their behaviour should be somewhat similar. The group of participants with unknown sex consisted of only 10 individuals in comparison to 50 men and 29 women. Therefore, the participants which we do not know the sex of has been omitted from further analysis.

From inspecting the responses, some of them seemed to be out of place. Especially response $3, 48, 59, 68$ and 69 were peculiar. These six subjects spent very little time on the test, ($< 6\,\text{min}, < 3\,\text{min}, < 5\,\text{min}, < 9\,\text{min}, < 7\,\text{min}$), in comparison to the average $\approx 18\,\text{min}$. In addition, participant 3 answered every question perfectly in less than 6 minutes which is definitely strange. By spending less than nine minutes on the test, the subject spends, on average, less than 20 seconds on each question. This hardly leaves any time left to evaluate the answers. Therefore, the participants mentioned are omitted from further analysis.

From Figure F.1 in the appendix, the probabilities assigned to each alternative for every question and participant is plotted against covariates. The covariate sex does not appear to be significant on its own. Covariates feedback, minimum score and score function does however appear to have some effect on the probabilities.

From Figure F.2 in the appendix, the interaction between sex and the score function appears to have an effect, but the effect is more or less the same for men and women. The most noticeable effect appears to come from the interactions between covariates. Notice how for example women who get feedback have an increased probability value.

From Figure F.3 in the appendix, the interaction between the score function and the minimum score appears to have an effect. Notice especially for participants with quadratic score function and requirement of a minimum score. The interaction effect minimum score and feedback appears to have an influence on the probability values when both are set to FALSE. The interaction between feedback and the score function shows that the effect is most noticeable for change in the score function and less effect for change in the feedback covariate.

Chapter 5

Method for statistical analysis of the empirical data

In this chapter the statistical method for analysing the empirical data is presented. We will introduce a new statistical model as a suggestion for modelling the level of confidence provided by the participants.

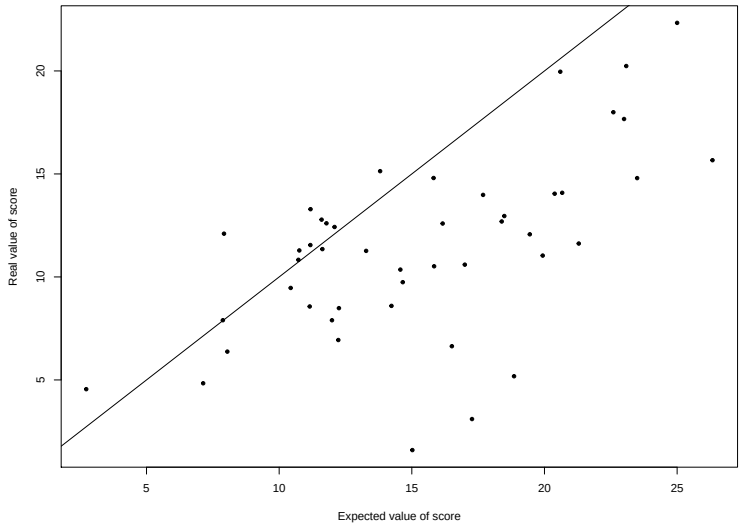

Figure 5.1: Obtained score vs expected value of score for the quadratic score function.

In Figure 5.1 the obtained score on the y-axis and the expected value of score on the x-axis are plotted for the participant's score by the quadratic score function. The straight line serves as a reference for when participants can accurately estimate

their level of confidence. When the obtained score is the same as the expected score, the participant has not overestimated nor underestimated his/her own level of confidence. Points above the straight line indicates underconfident behaviour, where the participant achieves a higher score than he/she expected. Overconfident behaviour is indicated by points below the straight line, where the obtained score is lower than the participant expected.

A model of the overconfidence can be proposed to further investigate how a probabilistic MCT performs in practice. Linear regression could possibly be used in the analysis of the quadratic score function, but the logarithmic score function can not be modelled properly using that method. The problem arises because the range of the logarithmic score function is $[-\infty, 1]$, where $-\infty$ of course is not a finite limit. It would be naive to assume that all participants in a test can correctly asses their probabilities such that none of them will put probability 0 to an alternative that turns out to be correct. Also, the assumption of normally distributed residuals with homoscedastic variance of an ordinary linear regression model may not hold (Fahrmeir, Kneib, Lang, and Marx, 2013, p. 75).

5.1 Statistical model

Below, an alternative method to model the probability distribution is suggested. The goal is to estimate the probability that each alternative is correct as a function of the subjective probabilities reported by each participant, using the known correct alternatives as the data. Let r_{ijk} denote the reported probability of participant i on question j for alternative k. We consider a model where the probability that a given alternative k, on question j, for participant i turns out to be correct is given by

$$r'_{ijk}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{r_{ijk}^{(e^{\boldsymbol{x}_i\boldsymbol{\alpha}})} + e^{\boldsymbol{y}_i\boldsymbol{\beta}}}{\sum_{k=1}^{m} \left(r_{ijk}^{(e^{\boldsymbol{x}_i\boldsymbol{\alpha}})} + e^{\boldsymbol{y}_i\boldsymbol{\beta}} \right)}$$

Here, $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are vectors of numerical covariates and dummy variables encoding categorical variables. They are the i-th row-vectors of the model matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$. To see how these model matrices are created and used in R code for fitting the model, see Appendix D.3.

The two functions $\boldsymbol{x}_i\boldsymbol{\alpha}$ and $\boldsymbol{y}_i\boldsymbol{\beta}$ are linear predictors, as defined by McCullagh and Nelder (1989, p. 56-60). In this case, the column vector of coefficients are $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The explanatory variables are the row vectors of covariates, $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ for participant i, where $i = 1, \ldots, 74$. The linear predictors predict how much the probabilities given by the participant should be adjusted for under-/overconfidence.

Similar to the linear regression analysis, the aim of modelling the probability distribution is to estimate the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This can be done by maximising the multinomial likelihood

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i,j} {r'_{ij1}}^{d_{j1}} {r'_{ij2}}^{d_{j2}} \ldots {r'_{ij4}}^{d_{j4}},$$

or equivalently the log likelihood

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i,j} \sum_{k} d_{jk} \ln r'_{ijk}(\boldsymbol{\alpha}, \boldsymbol{\beta}),$$

where $d_j = (d_{j1}, \ldots, d_{jm})$ are indicator variables indicating which alternative is correct and incorrect for question j.

Note that for $\boldsymbol{x}_i\boldsymbol{\alpha} = 0$ and $\boldsymbol{y}_i\boldsymbol{\beta} = -\infty$, the modified probabilities r'_{ijk} are equal to the probabilities reported by the participant. This corresponds to a participant that has inferred his/her subjective probabilities correctly, such that he/she is neither over- nor underconfident.

Negative $\boldsymbol{x}_i\boldsymbol{\alpha}$ deflates high reported probabilities r_{ijk} to smaller values and inflates low non-zero reported probabilities, with the modified predicted probabilities tending towards identical values as $\boldsymbol{x}_i\boldsymbol{\alpha} \to -\infty$. This models a form of overconfidence, where the probability that the correct alternative is k is closer to $1/m$ than expected from the probabilities reported by the participant.

Conversely, positive $\boldsymbol{x}_i\boldsymbol{\alpha}$ inflates high reported probabilities r_{ijk} and deflates low non-zero reported probabilities. As $\boldsymbol{x}_i\boldsymbol{\alpha} \to \infty$, the modified predicted probabilities will have one probability approaching 1, and the others approaching zero. This models a form of underconfidence, where the probability that the correct alternative is k is closer to 1 than expected from the probabilities reported by the participant.

The additional term $e^{\boldsymbol{y}_i\boldsymbol{\beta}}$ models overconfidence of a different form. This term is necessary in order to make alternatives that have been assigned a probability of zero, possible outcomes. This models how for example some participants may be more likely than others to overconfidently set some probabilities to zero. When the term $e^{\boldsymbol{y}_i\boldsymbol{\beta}}$ is positive ($\boldsymbol{y}_i\boldsymbol{\beta} > -\infty$), reported zero-probabilities are replaced by a small value. Depending on $\boldsymbol{\beta}$, the predicted probability that an alternative is correct, given that a zero-probability is reported, is equal to some small number which is estimated from the empirical data. This number may again depend on the covariates in $\boldsymbol{y}_i$.

5.1.1 Probability integral transform residuals

We will use the probability integral transform residuals to analyse the fit of the proposed model (Gamerman and Lopes, 2006, p. 13). These residuals are defined as the probability that the score is lower than what the participant actually obtained, thus, the residual is $u = P(\text{score} < s)$. If the model of overconfidence fits the data set well, these residuals should follow an approximately continuous uniform distribution. We know from linear regression that for a model that fits well to the data, the residuals should be approximately normally distributed. They are only approximately normally distributed because the residuals are the observed differences from the estimated model and the response in the data set (Fahrmeir et al., 2013, p. 79). We can use the same logic for the statistical model in this thesis as well, where the residuals are the probability of the simulated score being lower than the observed score. This is only one of many different types of residuals that we could have defined.

First of all, simulations are needed in order to estimate the probability because the score function can take 27^4 different discrete values. Secondly, there is only one test for each participant, so for this thesis there is not enough data to find the distribution based on the empirical data.

The correct alternative is sampled at random without replacement, with probability from the probability distribution given by the participant, for each question. From the sample of correct alternatives and the given probabilities, the score is calculated using the score function that the participant was scored by during the test. From this, the values of u and the distribution of the score can be found for each participant. In addition, the expected score from the distribution of the score predicted by the statistical model can be found. See Appendix D.4 for R-code of the statistical model and the probability integral transform residuals.

5.2 Find and evaluate the model

The stepwise selection is a combination of forward selection and backward elimination (Fahrmeir et al., 2013, p. 151). The statistical model suggested in this thesis is based on two model matrices. The stepwise selection appears to be the most intuitive way of finding an estimated model to fit the data. For each step of finding the model, the covariates must be evaluated for both linear predictors in the model, i.e. the $\boldsymbol{X}$ and $\boldsymbol{Y}$ model matrix. Here, the model matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ are used to estimate the parameter values of the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. One sub-model is to remove $e^{\boldsymbol{y}_i\boldsymbol{\beta}}$ entirely from the model. However, the data has probability zero under this sub-model, so this can be rejected immediately in favour of a different model where it is included.

Likelihood ratio tests as defined by Casella and Berger (2002, p. 375) is used in order to evaluate if the null hypothesis should be rejected for the alternative model. This is done by first finding the maximum likelihood for each model and use the likelihood ratio. The likelihood ratio is defined as

$$\lambda(x) = \frac{\sup\limits_{\Theta_0} L(\theta|x)}{\sup\limits_{\Theta} L(\theta|x)}.$$

The asymptotic distribution of the likelihood ratio test is,

$$-2\log\lambda(x) \sim \chi^2_\nu$$

where the degrees of freedom ν is equal to the difference in number of parameters in the H_1 and H_0 hypothesis. The null hypothesis is rejected if and only if $-2\log\lambda(x) \geq \chi^2_{\nu,\alpha}$ (Casella and Berger, 2002, p. 490).

The Akaike information criterion (AIC) is defined as

$$\text{AIC} = -2 \cdot l(\hat{\boldsymbol{\theta}}) + 2k,$$

where $l(\hat{\boldsymbol{\theta}})$ is the maximum value of the log-likelihood and k is the number of free parameters to be estimated (Fahrmeir et al., 2013, p. 148). Smaller values of the AIC represent a better model fit.

There is no exact method of detecting an outlier. However, an outlier is an observation that does not follow the model fitted to the data. One way of detecting outliers is therefore to look for large residuals (Fahrmeir et al., 2013, p. 160).

Chapter 6

Statistical analysis of empirical data

In this chapter we will use the methods developed and described in Chapter 5 to analyse the empirical data.

6.1 Quiz data

The average of every participant's probabilities per problem and for the correct alternative can be seen in Figure 6.1. Because alternative (a) is the correct one, it is to be expected that the probability reported for this alternative is the largest for each question. Notice, however, that this is not the case for question 7. Further, question 20, 22, 26 and 27 are also problematic (see Appendix E.1 and Figure 6.2). Here, we see that multiple alternatives has nearly the same average reported level of confidence. The participants may have found it difficult to single out one alternative that they have a larger level of confidence is correct, compared to the other alternatives. We see this as questions where more than one alternative has large reported probabilities. For question 26, we can see that all four alternatives has almost the same average probability. Notice in particular question 27, where the average level of confidence for one of the wrong alternatives is larger than for the correct alternative.

To analyse what it would look like if this was a traditional MCT, the alternative with the highest marked probability for each question can be replaced by the value of 1 and the other alternatives set to 0. The average probability assigned to the correct alternative plotted against the fraction of participants who successfully assigned the highest probability on the correct alternative is shown in Figure 6.3.

We can see from Figure 6.3 that there is a difference between the traditional and probabilistic multiple choice test. The largest difference between them is for question 6 and 24 (highlighted in red), which are 0.82 vs 0.66 and 0.81 vs 0.65 respectively. Thus, for a traditional MCT, a teacher may assume that more than 80% of the students in a class understand problem 6 and 24 well enough. However,

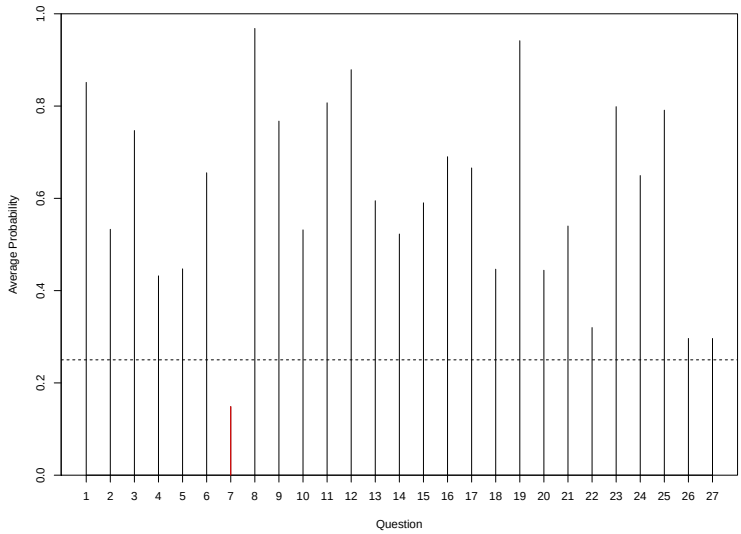

Figure 6.1: Plot of the average probability assigned to the correct alternative for every problem in the quiz. The dotted line represents a probability of 0.25, i.e. $(1/m)$ where $m = 4$ in this quiz.

for the probabilistic MCT, the average student has 0.66 and 0.65 level of confidence that this alternative is correct. Even though the difference is not as large as this for every question, it is crucial for a teacher to have convincing evidence of understanding among the students, in order to move forward.

Entropy is a widely accepted measure of uncertainty, which is the negative value of Shannon's information that was introduced in Section 3.2 (Bernardo and Smith, 1994, p. 79). A plot of the entropy for each question is shown in Figure 6.4. Clearly, the average participant is relatively certain, particularly question 8, 12 and 19 (shown in red), which have the highest average probability assigned to them. The largest entropy can be seen for question 22 and 26 (shown in blue), indicating confusion for the average participant. This is also seen in the average probability assigned to the correct alternative which is close to $1/m$. Thus, the average participant has been able to correctly identify when he/she is confused and thus assign a uniform distribution over the alternatives.

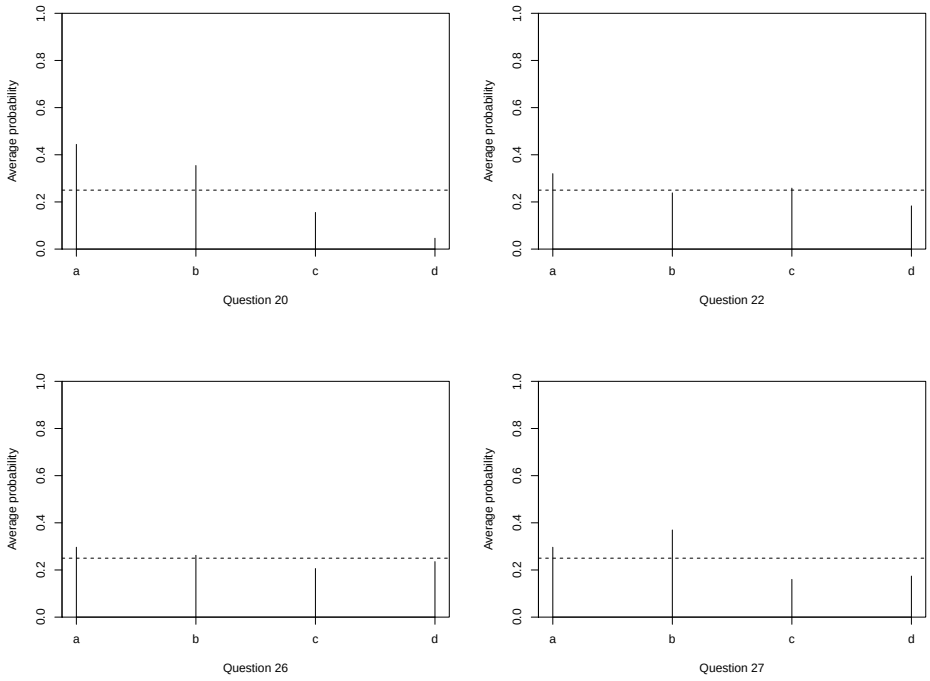

Figure 6.2: Plot of the average probability assigned to each alternative for question 20, 22, 26 and 27.The dotted line represents a probability of 0.25, i.e. $(1/m)$ where $m = 4$ in this quiz.

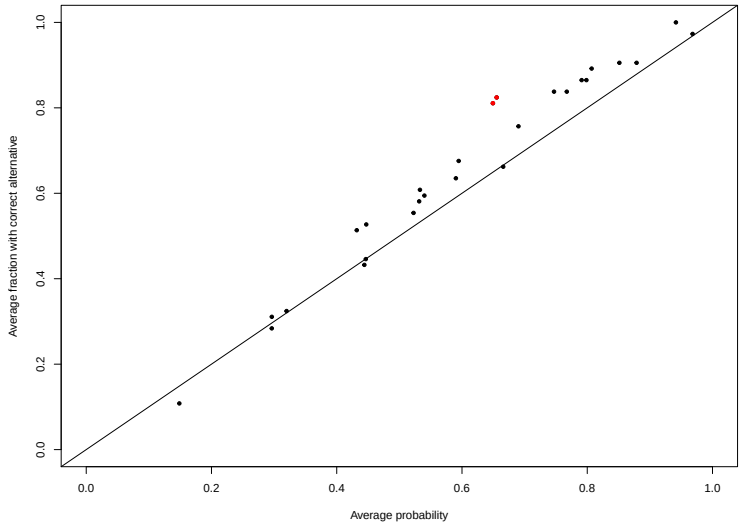

Figure 6.3: Average probability assigned to the correct alternative in a probabilistic MCT plotted against the fraction of participants who would mark the correct alternative in a traditional MCT. Question 6 and 24 (highlighted in red)

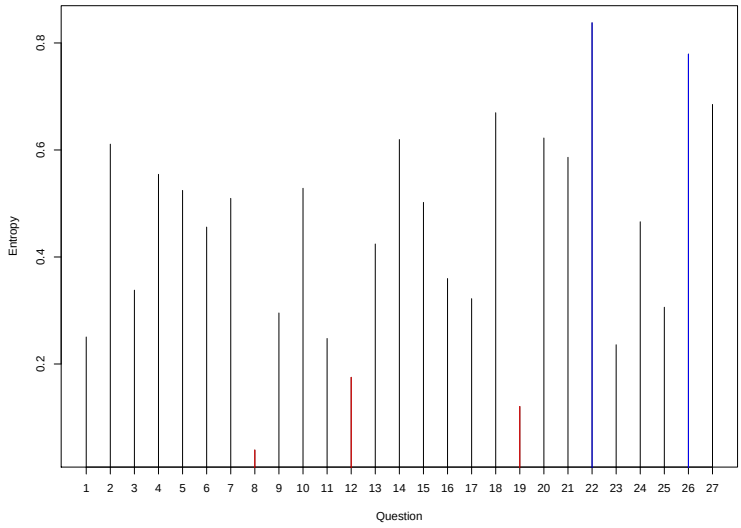

Figure 6.4: Plot of the average entropy for each question.

6.2 Model selection

A model without covariates is fitted to the data in order to estimate the parameters and log-likelihood for the null hypothesis. The results are shown in Table 6.1. The log-likelihood is $l_0 = -1693.287$ for H_0 and $\text{AIC}_0 = 3390.574$. The parameter estimates $\alpha_0 = -0.093$ and $\beta_0 = -3.212$ indicate that without any explanatory covariates, the probabilities marked by the participants are too extreme. Thus, from Section 5.1, large probabilities are deflated and small probabilities are inflated. For example, if the reported probabilities on a particular question are 0.6, 0.3, 0.1 and 0, then the predicted probabilities based on this model become proportional to

$$\begin{aligned} 0.6^{e^{-.093}} + e^{-3.212} &= 0.6^{0.911} + 0.04 = 0.66 \\ 0.3^{e^{-.093}} + e^{-3.212} &= 0.3^{0.911} + 0.04 = 0.37 \\ 0.1^{e^{-.093}} + e^{-3.212} &= 0.1^{0.911} + 0.04 = 0.16 \\ 0^{e^{-.093}} + e^{-3.212} &= 0 + 0.04 = 0.04, \end{aligned}$$

respectively, and equal to 0.540, 0.300, 0.130 and 0.032 after normalisation.

Table 6.1: Estimated parameters for the H_0 model.

Parameter	Covariate	Estimate	Standard error
α_0	Intercept	-0.093	0.092
β_0	Intercept	-3.212	0.112

An alternative model was found by considering different models with degrees of freedom $(2, \ldots, 10)$. Starting with 2 parameters (the H_0 model), parameters were added successively for both $\boldsymbol{X}$ and $\boldsymbol{Y}$. For every subset, the model with the lowest AIC was found. Ultimately, the AIC became larger for more than 9 parameters. Therefore, the model with 9 parameters with the lowest AIC is the one that is suggested as the alternative H_1 model. The only covariates that were considered in this method were sex, feedback, minimum score and score function.

The two linear predictors in symbolic notation (McCullagh and Nelder, 1989, p. 56-60) of the suggested alternative model is

$$\begin{aligned} \boldsymbol{x_i\alpha} &= \text{minimum score} + \text{sex} + \text{feedback} + \text{sex} : \text{feedback} \\ \boldsymbol{y_i\beta} &= \text{score function} + \text{feedback} + \text{score function} : \text{feedback}. \end{aligned} \tag{6.1}$$

With a log-likelihood of $l_1 = -1680.886$ and 9 degrees of freedom, the result of the likelihood ratio test (LRT) is

$$-2\log\lambda(x) = -2\log\frac{L_0(\alpha, \beta|x)}{L_1(\alpha, \beta|x)} = -2 \cdot (l_0 - l_1) = 24.802,$$

where $\chi^2_{(9-2),0.05} = 14.067 < -2\log\lambda(x)$. Thus, the H_0 model should be rejected at $\alpha = 0.05$ level of significance. From Table 6.3 and Table 6.4, the AIC of the H_1 model is 3379.772 and has the lowest AIC of the nearest model suggestions. The

Table 6.2: Estimated parameters for the alternative H_1 model with lowest AIC

Parameter	Covariate	Estimate	Standard error
α_0	Intercept	0.283	0.211
α_1	Minimum score TRUE	0.383	0.159
α_2	Sex Male	−0.768	0.231
α_3	Feedback TRUE	−0.080	0.273
α_4	Sex Male:Feedback TRUE	0.513	0.322
β_0	Intercept	−4.473	0.522
β_1	Scorefunction quadratic	1.491	0.559
β_2	Feedback TRUE	1.545	0.560
β_3	Scorefunction quadratic: Feedback TRUE	−1.910	0.615

nearest models are the models where one covariate (1 degree of freedom) is either added to or dropped from the model.

The LRT is used to test whether a null hypothesis should be rejected. The model in the null hypothesis is always the model with the least amount of covariates. Thus, from Table 6.3 we can conclude that minimum score and the interaction covariate of score function and feedback is significant. However, we do not find evidence to reject the null hypothesis that sex and feedback should be excluded. None of the alternative hypothesis of adding a covariate resulted in rejection of the null hypothesis. We can therefore, in this case, conclude to keep the suggested model in Table 6.2.

An interesting covariate is the interaction between feedback and question position. This was evaluated after the alternative H_1 model was found and is therefore an added covariate to the existing model. The question position is a variable that tells us in what order the questions were presented to the participant. Feedback was given to some participants. It is of interest to see whether the feedback covariate shows signs of influence on over- or underconfidence in the participants, i.e. if the feedback influences the students to adjust their expectations during the course of the test. In Table 6.5 we can see the results of adding the interaction covariate of feedback and question position. There is no clear evidence that the interaction between feedback and question position is a better fitted model of the probabilities than the suggested H_1 model.

Table 6.3: Drop one covariate from the suggested H_1 model. For likelihood ratio test: $\chi^2_{1,0.05} = 3.841$.

Linear predictor	Covariate dropped	LRT	ΔAIC
$\boldsymbol{x_i\alpha}$	Minimum score	$6.008 > \chi^2_{1,0.05}$	4.008
	Sex : Feedback	$2.436 < \chi^2_{1,0.05}$	0.436
$\boldsymbol{y_i\beta}$	Score function : Feedback	$12.732 > \chi^2_{1,0.05}$	10.732

From Table 6.2 we see that the standard error is larger than the estimated parameter value for the covariate feedback. This might suggest that it is not significant. Further testing by dropping covariates feedback and interaction be-

Table 6.4: Add one covariate to the suggested H_1 model. For likelihood ratio test: $\chi^2_{1,0.05} = 3.841$.

Linear predictor	Covariate added	LRT	ΔAIC
$\boldsymbol{x_i\alpha}$	Score function	$0.710 < \chi^2_{1,0.05}$	1.290
	Sex : Minimum score	$0.200 < \chi^2_{1,0.05}$	1.800
	Minimum score : Feedback	$0.694 < \chi^2_{1,0.05}$	1.306
$\boldsymbol{y_i\beta}$	Sex	$1.312 < \chi^2_{1,0.05}$	0.688
	Minimum score	$0.058 < \chi^2_{1,0.05}$	1.942

Table 6.5: Add the interaction between the covariates question position and feedback. For likelihood ratio test: $\chi^2_{2,0.05} = 5.991$.

Linear predictor	Covariate added	LRT	ΔAIC
$\boldsymbol{x_i\alpha}$	Question position + Question position : Feedback	$1.401 < \chi^2_{2,0.05}$	2.599
$\boldsymbol{y_i\beta}$	Question position + Question position : Feedback	$2.130 < \chi^2_{2,0.05}$	1.870

tween sex and feedback resulted in a log-likelihood of -1683.779. The LRT gives $\chi^2 = 5.786 > \chi^2_{2,0.05}$. Based on the LRT we would conclude not to reject this model vs the suggested model in Table 6.2. However, we will use AIC as the model selection criterion in this thesis and proceed with the suggested model in Table 6.2.

From the method described in Section 5.1, for the linear predictor $\boldsymbol{x_i\alpha}$, we know that negative parameter values will increase the adjustment for overconfidence, and positive parameter values will decrease the adjustment. Negative predicted values will adjust the probability distribution towards the discrete uniform distribution. Let us again assume that the reported probabilities on a particular question are 0.6, 0.3, 0.1 and 0, but this time it is reported by two people with different sex. The male participant will have predicted probabilities proportional to

$$\begin{aligned} 0.6^{e^{0.283-0.768}} + e^{-4.473} &= 0.742 \\ 0.3^{e^{0.283-0.768}} + e^{-4.473} &= 0.488 \\ 0.1^{e^{0.283-0.768}} + e^{-4.473} &= 0.254 \\ 0^{e^{0.283-0.768}} + e^{-4.473} &= 0.011, \end{aligned}$$

and equal to 0.496, 0.326, 0.170 and 0.007 after normalisation. The female participant will have predicted probabilities proportional to

$$\begin{aligned} 0.6^{e^{0.283}} + e^{-4.473} &= 0.519 \\ 0.3^{e^{0.283}} + e^{-4.473} &= 0.214 \\ 0.1^{e^{0.283}} + e^{-4.473} &= 0.058 \\ 0^{e^{0.283}} + e^{-4.473} &= 0.011, \end{aligned}$$

and equal to 0.647, 0.267, 0.072 and 0.013 after normalisation. Thus, given that all other covariates are kept at their reference level, we can see that the female participants are *underconfident* and male participants are *overconfident*.

We can use the results from Table 6.2 to discuss the other covariates in the linear predictor $\boldsymbol{x}_i\boldsymbol{\alpha}$. It is interesting to note that male participants are *underconfident* if they must achieve a minimum score and get feedback during the test. Female participants, however, are underconfident regardless of which level the other covariates are set to. Female participants appear to become more underconfident if they must achieve a minimum score. Male participants however, will become less overconfident if they must achieve a minimum score. Female participants are least underconfident and male participants are least overconfident if they receive feedback during the test.

We can see from Table 6.2 that the predicted value of the linear predictor $\boldsymbol{y}_i\boldsymbol{\beta}$ will always be negative. Therefore, a small positive constant is added for every participant. How large this added constant is, depends on the score function and if feedback is given to the participant. According to the model, the added constant will be largest if the participant received feedback during the test and was scored by the logarithmic score function. The added constant is almost as large if the participant did not get feedback, and was scored by the quadratic score function.

This suggests that participants who were scored by the logarithmic score function and received feedback during the test more often put zero probability on alternatives. Since this linear predictor models a type of overconfidence, it then suggests that these participants are overconfident. It is interesting to note that the smallest value of the added constant is found when the participant was scored by the logarithmic score function and did not get feedback during the test. This may suggest that a test that influences the participants to correctly estimate their level of confidence is a test without feedback, where their score is calculated by the logarithmic score function.

6.3 Residual analysis

From Figure 6.5 we can see the residuals of the H_0 model fitted to the uniform distribution, and it is clearly ill-fitted showing signs of large discrepancies. Thus, the residuals are not approximately uniformly distributed. From Figure 6.6 the residuals u of the suggested H_1 model is fitted to the uniform distribution. It seems highly reasonable that the residuals of the suggested model approximately follow the assumption of the continuous uniform distribution. The over- and underconfidence are satisfactory modeled by the H_1 model with the given covariates.

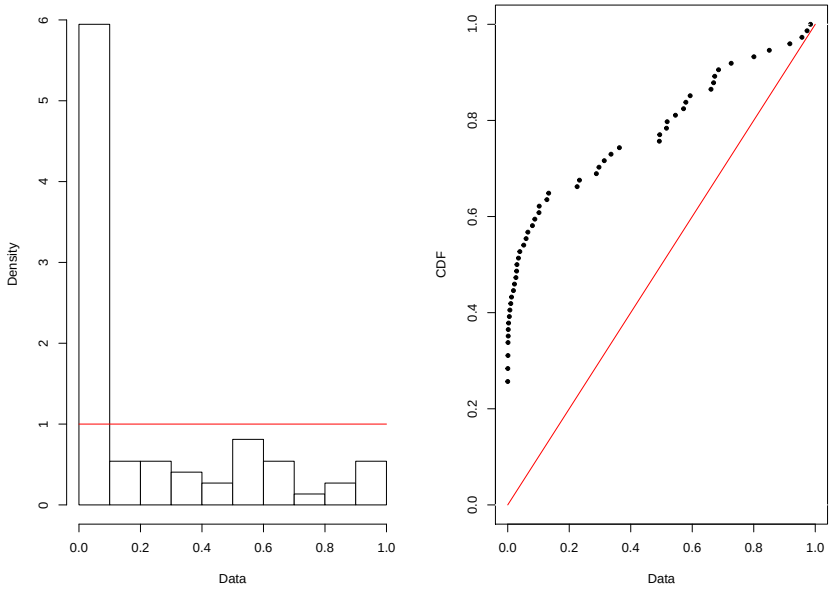

Figure 6.5: Empirical and theoretical probability density function (right) and cumulative distribution function (left) for the residuals of the H_0 model (black dots) and the continuous uniform distribution (red line).

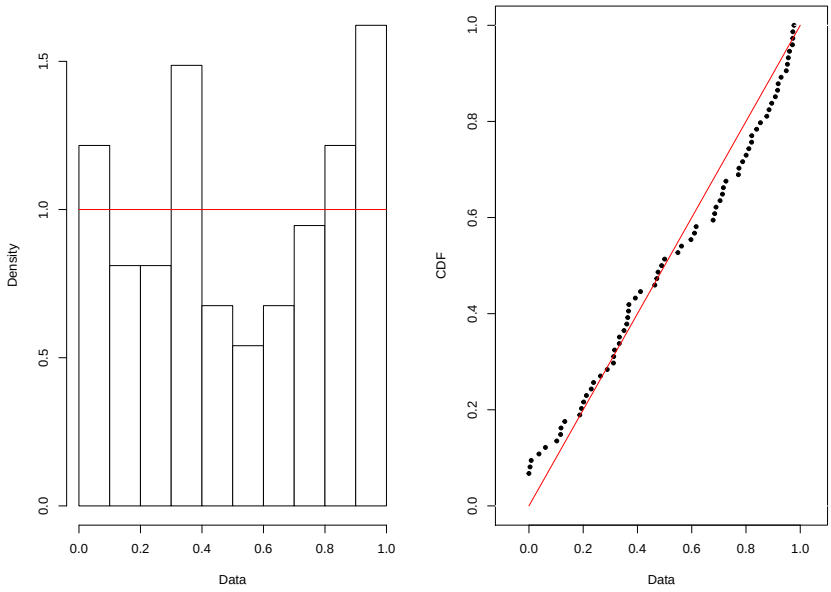

Figure 6.6: Empirical and theoretical probability density function (right) and cumulative distribution function (left) for the residuals of the H_1 model (black dots) and the continuous uniform distribution (red line).

6.4 Over-/underconfidence analysis

The probabilities can be modelled by using the covariates from Section 6.2 and samples are simulated with the method described in Section 5.1.1. The distribution of the score can be seen for participant 1 in Figure 6.7. The distribution of the score is shown at the bottom of Figure 6.7 are simulated from the probability distributions that are adjusted with the linear predictors for participant 1. As we can see, the distribution of the score is shifted towards a higher score. Participant 1 achieved an actual score of 9.8.

We want to investigate the participant's under- or overconfidence. The subjective expected score and the obtained score are found numerically by using the probability distribution provided by the participant. The statistical model presented previously is used to predict a distribution of the score. From the predicted distribution we can find an estimate of the score. This predicted estimate can then be plotted against the subjective expected score. This is shown in Figure 6.8 for the quadratic score function and in Figure 6.9 for the logarithmic score function. The subjective expected score vs the obtained score are shown in black and the subjective expected score vs the estimated score are shown in red for the score function used during the test. The corresponding values are shown in grey and pink for the participants that were scored by the opposite score function during the test.

Overconfident participants will have expected scores that are higher than what they actually obtained, i.e. points below the straight line. The straight line illustrates the "perfect" participant that correctly estimates their abilities such that the subjective expected score exactly matches the obtained score.

The statistical model predicts the distribution of the observed scores well. We see this as red and pink points which are surrounded by the black and grey points for the quadratic score function. Thus, the linear predictors model the under-/overconfident behaviour well. For the logarithmic score function we still have some issues with infinitely negative score when we estimate the score. It is therefore not as easy to see the effect of the statistical model in this plot.

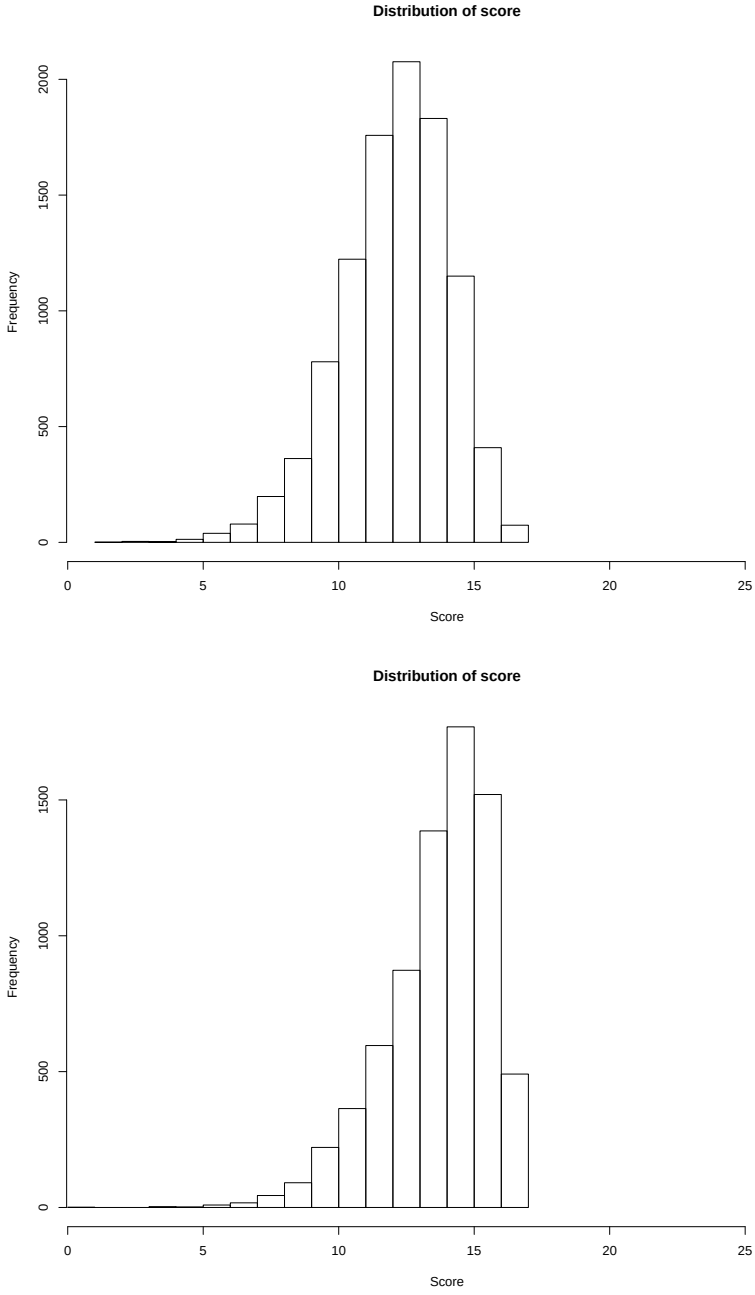

Figure 6.7: Distribution of the score for the probabilities given by participant 1 (top). Distribution of the score for the inflated/deflated probabilities given by participant 1 (bottom).

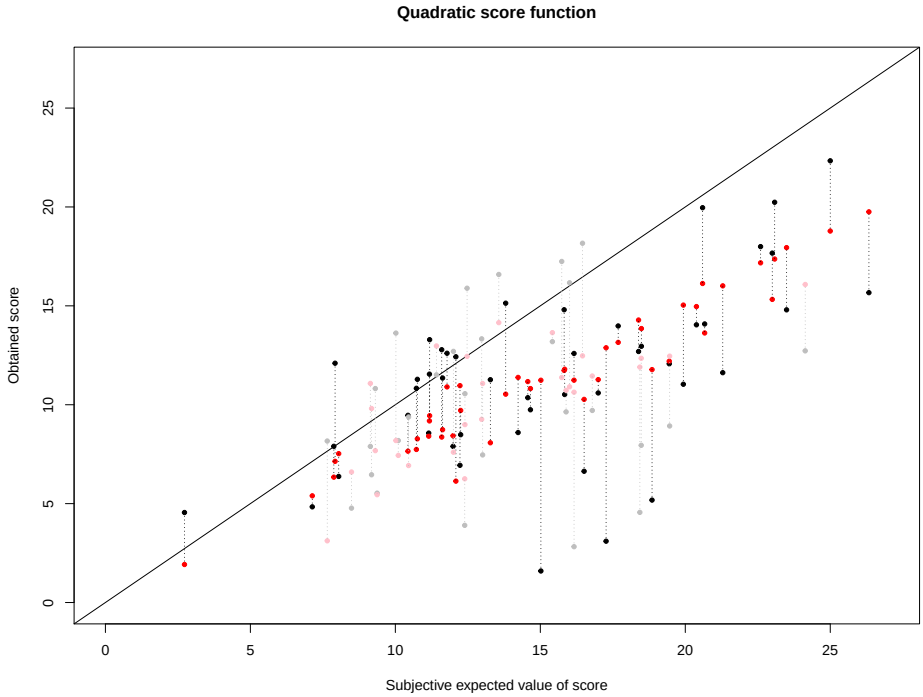

Figure 6.8: Expected score vs obtained score. Expected values based on reported subjective probabilities for participants scored by the quadratic score function are shown in black. Subjective expected value of score vs predicted expected value of score based on the statistical model are shown in red. Participants who were scored by the logarithmic score function during the test are shown respectively in grey and pink.

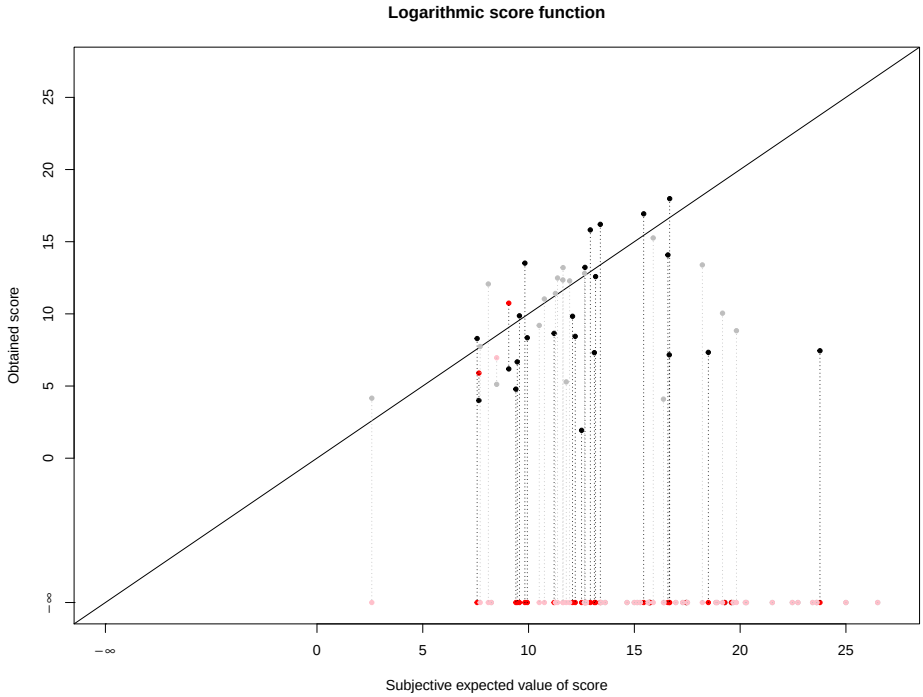

Figure 6.9: Expected score vs obtained score. Expected values based on reported subjective probabilities for participants scored by the logarithmic score function are shown in black. Subjective expected value of score vs predicted expected value of score based on the statistical model are shown in red. Participants who were scored by the quadratic score function during the test are shown respectively in grey and pink.

Chapter 7

Discussion

Based on the results from analysis of both the empirical and simulated data, a probabilistic multiple choice test appears to perform better than the traditional multiple choice test. First of all, based on the theoretical analysis in Section 3.3 the amount of data required is $4-5$ times larger for a traditional multiple choice test to be as accurate as the probabilistic multiple choice test (logarithmic score function). Secondly, with the probabilistic multiple choice test we have the ability to investigate the understanding of each individual separately, without taking a mean of the entire group. Finally, with the logarithmic score function and probabilistic multiple choice we have an unbiased estimator of the level of knowledge. In addition, this estimator has, partly for this reason, the lowest variance of all three estimators.

From the statistical model of the empirical data, the logarithmic score function and no feedback appears to have a positive effect on the type of overconfident behaviour that is modeleld by the linear predictor $\boldsymbol{y}_i\boldsymbol{\beta}$. The predicted constant that is added to the probabilities given by the participants, is smallest when the participants do not get feedback during the test and are scored by the logarithmic score function. Fischer (1982, p. 367) found that the logarithmic score function encourages better estimation, but underlines that the logarithmic score function may be worse at encouraging better estimation when using larger probabilities. We do not get the same result in this thesis for the logarithmic score function, since the score function's predicted effect is not dependent on the probability distribution. The score function is a covariate only in the linear predictor $\boldsymbol{y}_i\boldsymbol{\beta}$, and not in the linear predictor $\boldsymbol{x}_i\boldsymbol{\alpha}$ in the exponent of p. Thus, even if the probability becomes larger, the linear predictor that depends on the score function will stay the same.

We take a look at the type of overconfidence that is modelled by the linear predictor $\boldsymbol{x}_i\boldsymbol{\alpha}$ as well. From the results of this linear predictor, we find that the female/male participants will be least underconfident/overconfident if they receive feedback during the test. It is important that a test is fair for both genders, and we can therefore conclude that a well-constructed test may be a test that gives feedback during the test, scores are calculated by the logarithmic score function, and there is no demand of a minimum score.

A reservation against the probabilistic multiple choice tests, and in particular with logarithmic score function, is that it is computationally more demanding. However, with the access to computers we have today, this is hardly an obstacle to overcome. The test can easily be administered online as we have done in this thesis, or the answers can be transferred manually from paper to computer.

We can therefore conclude, based on the results found in this thesis, that the probabilistic multiple choice test appears to be a good alternative to the traditional multiple choice test. Furthermore, the logarithmic score function is an unbiased estimator of level of knowledge, in addition to affect the participants to more accurately estimate their level of confidence.

7.1 Improvements

To make sure that the participants actually understood the score functions and the reasoning behind why they should be honest, a test prior to the quiz should probably have been given to the participants. Since this was not done, it could have contributed to the variability in the score.

Another improvement suggested is to use curriculum-specific questions and not typical quiz-related questions. The participants could have taken the test more seriously and spent more time on the test to make sure their answers clearly depicted their level of knowledge. If it was curriculum-specific, it could have been used as a test to grade student's performance in a class, in stead of providing a monetary value as incentive to take the test.

We suspect that the participants could have restarted the app until they were picked to be scored by the quadratic score function. They may have wanted to avoid the logarithmic score function because they did not want to risk getting $-\infty$ as score, from which it is impossible to recover from. A score of $-\infty$ would ultimately exclude them from the chance of winning a gift card. That is probably why a much larger group of participants, to be exact 17 females and 31 males, were scored by the quadratic score function. In comparison, the group of participants that were scored by the logarithmic score function consisted of only 12 females and 19 males. Therefore, it may be best to administer a test with just the logarithmic score function. The participants could then be urged to not assign 0 to any alternative, as it is more or less impossible in any situation to be absolutely sure that something is wrong. This is a solution that Bratvold (Unpublished) has used for his students, which seems to be a good idea. Another solution to this problem, as suggested by Bickel (2010, p. 350), is that the students replace probabilities that are zero with a small number, for example 0.001. This way they would avoid getting a score which is impossible to recover from, while still displaying their level of confidence precisely.

Instead of replacing zero probabilities with some arbitrary small number such as 0.001, a potentially useful application of the statistical model (chapter 5) is to use it in such probability adjustments. In Figure 7.1 and 7.2 we can see what the obtained score would have been if the probability distributions had been adjusted according to the statistical model. The black points are the actual obtained scores

and the red points are the scores if we adjust the probability distributions according to the statistical model. As we can see, the obtained score would have been closer to the diagonal line in the plot. The diagonal line represents the score of participants who are neither over- nor underconfident. Notice in particular that none of the participants who were scored by the logarithmic score function would get an infinite negative score, if the probabilities would have been adjusted before calculating the obtained score.

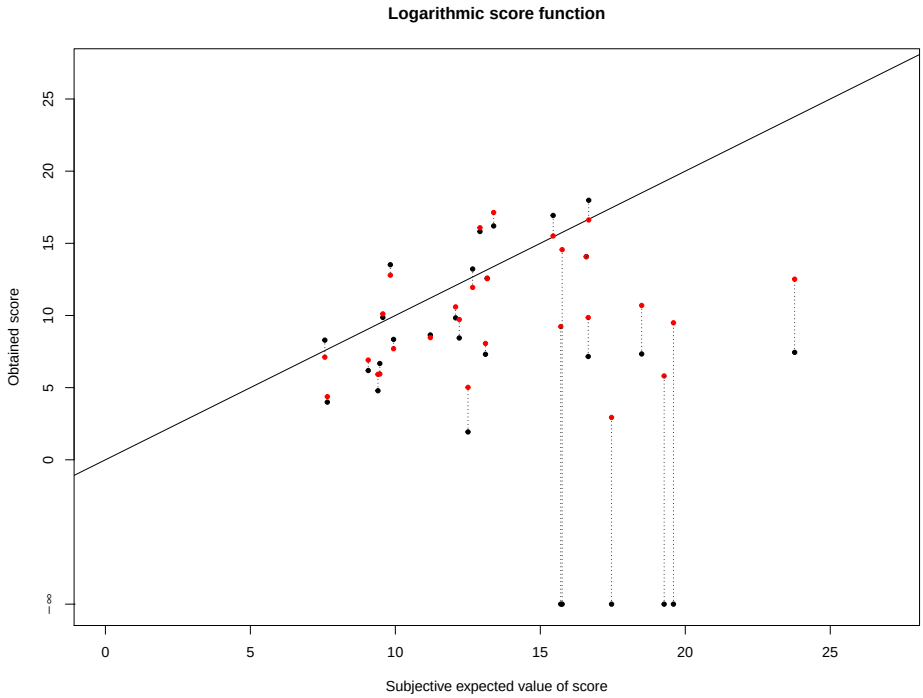

Figure 7.1: Subjective expected value of score vs obtained score. Black points represent the original obtained score calculated from the probability distribution provided by the participant. Red points represent the obtained score calculated from the adjusted probability distribution according to the statistical model.

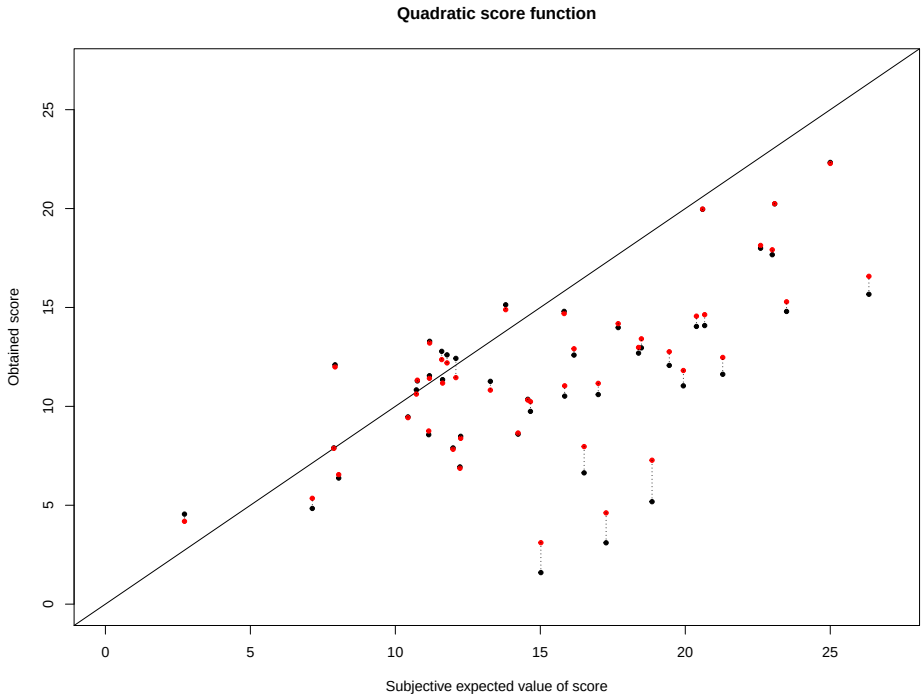

Figure 7.2: Subjective expected value of score vs obtained score. Black points represent the original obtained score calculated from the probability distribution provided by the participant. Red points represent the obtained score calculated from the adjusted probability distribution according to the statistical model.

7.2 Further work

In this thesis we used a multiple choice test in the form of a quiz, which was not curriculum-specific. For further research, it would be interesting to explore what the result would have been if the test had been issued in a class for a curriculum-specific test. Especially within national tests that are used specifically to get an overview of how well the Norwegian students are doing, it would be relevant to implement the probabilistic multiple choice in certain areas of the test.

Another area to investigate is how young the participants can be and still understand the theoretical background of the test such that their behaviour is rational. Intuitively, the younger the participants are when they learn how the test works, the easier it is to implement it smoothly when they get older.

Further, it would be interesting to examine the use of probabilistic multiple choice tests over an extended period of time. The ability to evaluate uncertain events takes practice and skill to be accurate. A research area could be to follow a group of students and measure their development in estimating their own knowledge. Does a probabilistic multiple choice test contribute to individuals being better at decision making in a situation with uncertainty?

The method of the probabilistic multiple choice tests can be transferred to other areas of expertise. As an example, medical diagnosis can be treated as a decision problem with uncertain outcomes and consequences of actions. The score function can in this sense be used as a score of certainty from the professional decision maker. These type of situations have been researched for weather forecasting by Winkler, Muñoz, Cervera, Bernardo, Blattenberger, Kadane, Lindley, Murphy, Oliver, and Ríos-Insua (1996), Winkler (1971), Winkler (1969) and Winkler and Murphy (1968). Johnstone (2007) has researched a method for scoring financial analysts according to their ability to predict uncertain events, much like weather forecasters. These examples show that the results that are found in this thesis are applicable to other areas than education.

Bibliography

Ben-Simon, A., Budescu, D. V., Nevo, B., March 1997. A comparative study of measures of partial knowledge in multiple-choice tests. Applied Psychological Measurement 21 (1), 65–88.

Bernardo, J. M., 1998. A decision analysis approach to multiple-choice examinations. In: Applied decision analysis. Springer, pp. 195–207.

Bernardo, J. M., Smith, A. F., 1994. Bayesian Theory. Wiley Series in Probability and Mathematical Statistics. Wiley.

Bickel, J. E., 2010. Scoring rules and decision analysis education. Decision Analysis 7 (4), 346 – 357.
URL http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=64499560&site=ehost-live

Bratvold, R. B., Unpublished. Strictly proper scoring rules in decision analysis education in petroleum engineering at uis.

Casella, G., Berger, R. L., 2002. Statistical Inference, 2nd Edition. Duxbury advanced series. Duxbury.

Espinosa, M., Gardeazabal, J., 2010. Optimal correction for guessing in multiple-choice tests. Journal of Mathematical Psychology 54, 415–425.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression: Models, Methods and Applications. Springer, Dordrecht.

Fischer, G. W., 1982. Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting. Organizational Behavior and Human Performance 29 (3), 352 – 369.
URL http://www.sciencedirect.com/science/article/pii/0030507382902501

Gamerman, D., Lopes, H. F., 2006. Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference, 2nd Edition. Texts in Statistical Science Series. Chapman Hall/CRC, Boca Raton.

Johnstone, D., 2007. Economic darwinism: Who has the best probabilities? Theory and Decision 62 (1), 47–96.
URL http://dx.doi.org/10.1007/s11238-006-9006-2

Kelly, F. J., 1916. The kansas silent reading tests. Journal of Educational Psychology 7 (2), 63–80.

Kendall, M., Stuart, A., Ord, J., 1991. Kendall's advanced theory of statistics, 5th Edition. Vol. 2. Edward Arnold, London.

Larsen, R. J., Marx, M. L., 2014. Introduction to Mathematical Statistics and Its Applications, 5th Edition. Pearson Higher Ed USA.

McCullagh, P., Nelder, J., 1989. Generalized linear models, 2nd Edition. Chapman and Hall, Ch. 3.4.

Neyhart, C. A., Abrassart, A., 1984. A recommended formula for scoring probabilistic multiple-choice tests. Journal of Accounting Education 2 (1), 71 – 81.
URL http://www.sciencedirect.com/science/article/pii/0748575184900265

Poizner, S. B., Nicewander, W. A., Gettys, C. F., 1978. Alternative response and scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response modes. Applied Psychological Measurement 2 (1), 83–96.

R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Viennea, Austria.

Utdanningsdirektoratet, 03 2016. Hva er nasjonale prøver?
URL http://www.udir.no/eksamen-og-prover/prover/om-nasjonale-prover/

Wikipedia, January 2017. Dirichlet distribution.
URL https://en.wikipedia.org/wiki/Dirichlet_distribution

Winkler, R. L., 1969. Scoring rules and the evaluation of probability assessors. Journal of the American Statistical Association 64 (327), 1073–1078.
URL http://www.jstor.org/stable/2283486

Winkler, R. L., 1971. Probabilistic prediction: Some experimental results. Journal of the American Statistical Association 66 (336), 675–685.
URL http://www.jstor.org/stable/2284212

Winkler, R. L., Muñoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., Lindley, D. V., Murphy, A. H., Oliver, R. M., Ríos-Insua, D., 1996. Scoring rules and the evaluation of probabilities. Test 5 (1), 1–60.
URL http://dx.doi.org/10.1007/BF02562681

Winkler, R. L., Murphy, A. H., 1968. "good" probability assessors. Journal of applied Meteorology 7 (5), 751–758.

Appendix A

Probability distributions

Here we present the probability distributions and their properties that are used in this thesis.

A.1 Dirichlet distribution

A random vector $\boldsymbol{x}$ is Dirichlet(α) distributed for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, $\alpha_i > 0$, if its sample space is $x_i > 0$, $x_m = 1 - \sum_{i=1}^{m-1} x_i$ and its density is

$$f(x_1, \ldots, x_{m-1}) = \frac{\Gamma\Big(\sum_{i=1}^{m} \alpha_i\Big)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \prod_{i=1}^{m} x_i^{\alpha_i - 1}$$

(Bernardo and Smith, 1994). Let $x_i = p_i$, $\sum_{i=1}^{m} p_i = 1$ and assume $\alpha_1 = \ldots = \alpha_m = \alpha$, then

$$f(p_1, p_2, \ldots, p_m) = p_1^{\alpha-1} \ldots p_m^{\alpha-1} \frac{\Gamma(m\alpha)}{\Gamma(\alpha)^m}.$$

For the rest of this section all the results are from Wikipedia (2017). The mean of p_i and $\ln p_i$ is

$$\begin{aligned} E(p_i) &= \frac{\alpha_i}{\sum_i \alpha_i} = \frac{1}{m} \\ E(\ln p_i) &= \psi(\alpha_i) - \psi\Big(\sum_i \alpha_i\Big) = \psi(\alpha) - \psi(m\alpha). \end{aligned}$$

The covariance is

$$\mathrm{Cov}[\log(p_i), \log(p_j)] = \psi_1(\alpha_i)\delta_{ij} - \psi_1(\alpha_0) = \psi_1(\alpha)\delta_{ij} - \psi_1(\alpha).$$

The different moments of the Dirichlet distribution can be found by the following

function

$$E\Big[\prod_{i=1}^{m} p_i^{\beta_i}\Big] = \frac{\Gamma\Big(\sum_{i=1}^{m}\alpha_i\Big)}{\Gamma\Big[\sum_{i=1}^{m}(\alpha_i+\beta_i)\Big]} = \frac{\Gamma\Big(m\alpha\Big)}{\Gamma\Big[m\alpha+\sum_{i=1}^{m}\beta_i\Big]}. \tag{A.1}$$

A.2 Beta and binomial distribution

The results presented in this section are found in Casella and Berger (2002). The beta family of distributions is a continuous family on (0,1) indexed by two parameters. The Beta(α,β) probability density function is

$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \quad 0<x<1, \quad \alpha>0, \quad \beta>0 \tag{A.2}$$

where $B(\alpha,\beta)$ denotes the beta function,

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1})dx.$$

The beta function is related to the gamma function through the following identity

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The moments of the beta distribution, for $n>-\alpha$, is

$$\begin{aligned} E(X^n) &= \frac{1}{B(\alpha,\beta)}\int_0^1 x^n x^{\alpha-1}(1-x)^{\beta-1}dx \\ &= \frac{1}{B(\alpha,\beta)}\int_0^1 x^{(\alpha+n)-1}(1-x)^{\beta-1}dx. \end{aligned}$$

We recognizee the intergrand as the kernel of a Beta$(\alpha+n,\beta)$ pdf; hence,

$$E(X^n) = \frac{\Gamma(\alpha+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+n)\Gamma(\alpha)}.$$

Using this and the relation between expected value and variance with n=1 and n=2, we can calculate the mean and variance of the Beta(α,β) distribution as

$$E(X) = \frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad \text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

The binomial distribution is defined as

$$P(Y=y|n,p) = \binom{n}{y}p^y(1-p)^{n-y}, \qquad y=0,1,2,\ldots,n.$$

Appendix B

Expected score and variance

Here, I use the Dirichlet distribution to model variation in the state of knowledge. Assuming the participants are honest, the probabilities used in the Dirichlet distribution are the same as the honest subjective probabilities, $p_1, \ldots, p_m$. The α is a level of knowledge the candidate has, which is assumed equal for all alternatives. Consider a test with m alternatives for question j, where $j \in \{1, \ldots, n\}$, then the distribution of $p_1, \ldots, p_m$ for question j is defined as

$$f(p_1, \ldots, p_m) = \frac{\Gamma\left(\sum_{i=1}^{m} \alpha_i\right)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \prod_{i=1}^{m} p_i^{\alpha_i - 1}, \qquad m = 4$$

The marginal distribution of p_i, for a particular alternative i, assuming that the distribution is symmetric such that all $\alpha_1 = \ldots = \alpha_m = \alpha$, is then

$$f_{p_i}(\boldsymbol{p}) = \frac{\Gamma(m\alpha)}{\Gamma(\alpha)\Gamma((m-1)\alpha)} p_i^{\alpha-1}(1-p_i)^{(m-1)\alpha-1} \tag{B.1}$$

Let $(V_1, V_2, \ldots, V_m)$ be the event that alternative $1, 2, \ldots, m$ is chosen as the answer. Assuming that the participants has derived their subjective probabilities by correct statistical inference, we then have that

$$P(V_i|p_1, \ldots, p_m) = I\{p_i = \max(p_1, \ldots, p_m)\}, \tag{B.2}$$

where $\mathrm{I}\{p_i = \max(p_1, \ldots, p_m)\}$ is an indicator function which has value 1 when p_i is the maximum, and value 0 otherwise. Equation (B.2) expresses that alternative i is chosen as the correct answer by a rational participant, with probability 1 only if p_i is the largest. The conditional probability that alternative i is correct, given the confidence p_i, is $P(R_i|p_i) = p_i$.

The probability of the event that the correct answer is chosen is

$$\begin{aligned} P(V_1 \cap R_1 \cup \ldots \cup V_m \cap R_m) &= \sum_{i=1}^{m} P(R_i \cap V_i) \\ &= mP(R_1 \cap V_1) \\ &= m \int \ldots \int P(V_1 \cap R_1 | p_1, \ldots, p_m) f(p_1, \ldots, p_m) dp_1 \ldots dp_{m-1} \\ &= m \int \ldots \int p_1 \, \mathrm{I}\{p_i = \max(p_1, \ldots, p_m)\} f(p_1, \ldots, p_m) dp_1 \ldots dp_{m-1}. \end{aligned} \tag{B.3}$$

Because of the indicator function, the integrand is zero within the area where $p_1 > p_2, \ldots, p_m$. This comes from the assumption that all participants act rational according to the axioms of rational behaviour. To the best of my knowledge there is no closed form solution for this integral, except in the case of $\alpha = 1$ and $m = 2$, where Equation (B.3) simplifies to 3/4. We will therefore find the expectation and variance numerically by the use of the binomial distribution and Monte Carlo integration. We estimate the probability by sampling from the Dirichlet distribution, find the maximum value of each sample and take the average of them. We then have an estimated value for the probability that the correct answer is chosen. The score is then binomially distributed with this estimated probability and number of questions as parameters,

$$\mathrm{Bin}(n, P(V_1 \cap R_1 \cup V_2 \cap R_2 \cup \ldots \cup V_m \cap R_m)).$$

Let $\hat{p} = P(V_1 \cap R_1 \cup V_2 \cap R_2 \cup \ldots \cup V_m \cap R_m)$, then

$$E(p_i | R_i) = n\hat{p}$$

is the expected score and,

$$Var(p_i | R_i) = n\hat{p}(1 - \hat{p})$$

is the variance of the score for the traditional MCT.

B.1 Quadratic score function

Let the quadratic score function be defined as $Q(\boldsymbol{r}, \boldsymbol{d}) = 1 - \sum_{i=1}^{m} (r_i - d_i)^2$, where r_i are the probabilities given by the participant. Assuming the participant's behaviour is rational, the conditional expected score is

$$E(Q(\boldsymbol{p}, \boldsymbol{d}) | \boldsymbol{p}) = \sum_{i=1}^{m} p_i^2$$

To find the subjective expected score we need to consider the distribution of $\boldsymbol{p}$, which is assumed to be Dirichlet distributed. By using the expressions for moments

and mean of the Dirichlet distribution from Appendix A.1 we find that the expected value of the quadratic score function is

$$\begin{aligned} E_p(Q(\boldsymbol{p},\boldsymbol{d})) &= E_p(E(Q(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}) \\ &= E_p\Big(\sum_{i=1}^{m} p_i^2\Big) \\ &= \sum_{i=1}^{m} E_p\big(p_i^2\big) \\ &= \sum_{i=1}^{m} \frac{(\alpha+1)}{m(m\alpha+1)} \\ &= \frac{(\alpha+1)}{(m\alpha+1)}. \end{aligned}$$

The conditional variance of the subjective expected score given the probability distribution $\boldsymbol{p}$ is

$$\mathrm{Var}(Q(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}) = E(Q(\boldsymbol{p},\boldsymbol{d})^2|\boldsymbol{p}) - E(Q(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p})^2,$$

where

$$\begin{aligned} E(Q(\boldsymbol{p},\boldsymbol{d})^2|\boldsymbol{p}) &= E\left(\left(1-\sum_{i=1}^{m}(p_i-d_i)^2\right)^2\middle|\boldsymbol{p}\right) \\ &= E\left(1-2\sum_{i=1}^{m}(p_i-d_i)^2+\left(\sum_{i=1}^{m}(p_i-d_i)^2\right)^2\middle|\boldsymbol{p}\right) \\ &= 1-2\sum_{i=1}^{m}E(p_i^2-2p_id_i+d_i^2|\boldsymbol{p})+\sum_{i=1}^{m}E\big((p_i-d_i)^4|\boldsymbol{p}\big) \\ &\quad +2\sum_{i<j}\sum E((p_i-d_i)^2(p_j-d_j)^2|\boldsymbol{p}) \\ &= 1-2\sum_{i=1}^{m}p_i^2+4\sum_{i=1}^{m}p_i^2-2\sum_{i=1}^{m}p_i+\sum_{i=1}^{m}p_i^4-4\sum_{i=1}^{m}p_i^4+6\sum_{i=1}^{m}p_i^3 \\ &\quad -4\sum_{i=1}^{m}p_i^2+1+2\sum_{i<j}\sum E\big((p_i^2-2p_id_i+d_i^2)(p_j^2-2p_jd_j+d_j^2)|\boldsymbol{p}\big) \\ &\quad \vdots \\ &= -2\sum_{i=1}^{m}p_i^2-3\sum_{i=1}^{m}p_i^4+6\sum_{i=1}^{m}p_i^3+2\sum_{i<j}\sum(-3p_i^2p_j^2+p_i^2p_j+p_ip_j^2). \end{aligned} \tag{B.4}$$

Hence, using what we found in Equation (2.3) and Equation (B.4), the conditional variance of the score is

$$\begin{aligned}
\mathrm{Var}(Q(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}) &= E(Q(\boldsymbol{p},\boldsymbol{d})^2|\boldsymbol{p}) - E(Q(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p})^2 \\
&= -2\sum_{i=1}^{m} p_i^2 - 3\sum_{i=1}^{m} p_i^4 + 6\sum_{i=1}^{m} p_i^3 + 2\sum_{i<j}\sum(-3p_i^2p_j^2 + p_i^2p_j + p_ip_j^2) - \left(\sum_{i=1}^{m} p_i^2\right)^2 \\
&= -2\sum_{i=1}^{m} p_i^2 - 3\sum_{i=1}^{m} p_i^4 + 6\sum_{i=1}^{m} p_i^3 + 2\sum_{i<j}\sum(-3p_i^2p_j^2 + p_i^2p_j + p_ip_j^2) \\
&\quad - \sum_{i=1}^{m} p_i^4 - 2\sum_{i<j}\sum p_i^2p_j^2.
\end{aligned}$$

From a general formula for expectations of moments of the Dirichlet distribution (A.1), we find

$$\begin{aligned}
E(p_i) &= \frac{1}{m} \\
E(p_i^2) &= \frac{(\alpha+1)}{m(m\alpha+1)} \\
E(p_i^3) &= \frac{(\alpha+1)(\alpha+2)}{m(m\alpha+1)(m\alpha+2)} \\
E(p_i^4) &= \frac{(\alpha+1)(\alpha+2)(\alpha+3)}{m(m\alpha+1)(m\alpha+2)(m\alpha+3)} \\
E(p_ip_j) &= \frac{\alpha}{m(m\alpha+1)} \\
E(p_ip_j^2) &= E(p_i^2p_j) = \frac{\alpha(\alpha+1)}{m(m\alpha+1)(m\alpha+2)} \\
E(p_i^2p_j^2) &= \frac{\alpha(\alpha+1)^2}{m(m\alpha+1)(m\alpha+2)(m\alpha+3)} \\
\mathrm{Var}(p_i) &= \frac{(m-1)}{m^2(m\alpha+1)}
\end{aligned} \tag{B.5}$$

The subjective variance of the score can thus be found by using the equations in (B.5),

$$
\begin{aligned}
\operatorname{Var}(Q(\boldsymbol{p},\boldsymbol{d})) &= E_p \operatorname{Var}(Q(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}) + \operatorname{Var}_p E(Q(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}) \\
&= E_p\Bigg(-2\sum_{i=1}^{m} p_i^2 - 3\sum_{i=1}^{m} p_i^4 + 6\sum_{i=1}^{m} p_i^3 + 2\sum_{i<j}\sum (-3p_i^2p_j^2 + p_i^2p_j + p_ip_j^2) \\
&\qquad - \sum_{i=1}^{m} p_i^4 - 2\sum_{i<j}\sum p_i^2p_j^2\Bigg) + \operatorname{Var}_p\Big(\sum_{i=1}^{m} p_i^2\Big) \\
&= E_p\Bigg(-2\sum_{i=1}^{m} p_i^2 - 3\sum_{i=1}^{m} p_i^4 + 6\sum_{i=1}^{m} p_i^3 + 2\sum_{i<j}\sum (-3p_i^2p_j^2 + p_i^2p_j + p_ip_j^2) \\
&\qquad - \sum_{i=1}^{m} p_i^4 - 2\sum_{i<j}\sum p_i^2p_j^2\Bigg) + E_p\Bigg(\Big(\sum_{i=1}^{m} p_i^2\Big)^2\Bigg) - E_p\Big(\sum_{i=1}^{m} p_i^2\Big)^2 \\
&= -2\frac{(\alpha+1)}{(m\alpha+1)} - 3\frac{(\alpha+1)(\alpha+2)(\alpha+3)}{(m\alpha+1)(m\alpha+2)(m\alpha+3)} + 6\frac{(\alpha+1)(\alpha+2)}{(m\alpha+1)(m\alpha+2)} \\
&\qquad + (m-1)\Bigg(2\frac{\alpha(\alpha+1)}{(m\alpha+1)(m\alpha+2)} - 3\frac{\alpha(\alpha+1)^2}{(m\alpha+1)(m\alpha+2)(m\alpha+3)}\Bigg) \\
&\qquad - \frac{(\alpha+1)(\alpha+2)(\alpha+3)}{(m\alpha+1)(m\alpha+2)(m\alpha+3)} - (m-1)\frac{\alpha(\alpha+1)^2}{(m\alpha+1)(m\alpha+2)(m\alpha+3)} \\
&\qquad + \frac{(\alpha+1)(\alpha+2)(\alpha+3)}{(m\alpha+1)(m\alpha+2)(m\alpha+3)} + \frac{(m-1)\alpha(\alpha+1)^2}{(m\alpha+1)(m\alpha+2)(m\alpha+3)} \\
&\qquad - \frac{(\alpha+1)^2}{(m\alpha+1)^2} \\
&= -\frac{2(\alpha+1)}{(m\alpha+1)} - \frac{3(\alpha+1)(\alpha+2)(\alpha+3) + 3(m-1)\alpha(\alpha+1)^2}{(m\alpha+1)(m\alpha+2)(m\alpha+3)} \\
&\qquad + \frac{6(\alpha+1)(\alpha+2) + 2(m-1)\alpha(\alpha+1)}{(m\alpha+1)(m\alpha+2)} - \frac{(\alpha+1)^2}{(m\alpha+1)^2} \\
&= -\frac{(\alpha+1)\big(2m\alpha+2+\alpha+1\big)}{(m\alpha+1)^2} - \frac{3(\alpha+1)\big((\alpha+2)(\alpha+3) + (m-1)\alpha(\alpha+1)\big)}{(m\alpha+1)(m\alpha+2)(m\alpha+3)} \\
&\qquad + \frac{2(\alpha+1)\big((3\alpha+6) + (m-1)\alpha}{(m\alpha+1)(m\alpha+2)} \\
&= (\alpha+1)\Bigg(-\frac{(2m+1)\alpha+3}{(m\alpha+1)^2} - \frac{3(m\alpha^2+(m+4)\alpha+6)}{(m\alpha+1)(m\alpha+2)(m\alpha+3)} + \frac{2((m+2)\alpha+6)}{(m\alpha+1)(m\alpha+2)}\Bigg)
\end{aligned}
$$

B.2 Logarithmic score function

Let the logarithmic score function be defined as

$$L(\boldsymbol{r},\boldsymbol{d}) = \sum_{i=1}^{m} \big(d_i \ln(r_i)\big).$$

Assuming the participant's behaviour is rational, the conditional expected score is

$$E(L(\boldsymbol{r},\boldsymbol{d})|\boldsymbol{p}) = \sum_{i=1}^{m} E(d_i \ln(r_i)) = \sum_{i=1}^{m} p_i \ln(r_i).$$

The optimal decision is to set $\boldsymbol{r} = \boldsymbol{p}$, thus

$$E(L(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}) = \sum_{i=1}^{m} p_i \ln(p_i).$$

We assume that $\alpha_1 = \ldots = \alpha_m = \alpha$, thus we can evaluate the expectation of one p_i and take the summation. We choose to find the expectation of p_1, then

$$\begin{aligned}
E_p(L(\boldsymbol{p},\boldsymbol{d})) &= E_p(E(L(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}) \\
&= E_p\left(\sum_{i=1}^{m} p_i \ln(p_i)\right) \\
&= \sum_{i=1}^{m} \int \ldots \int \frac{\Gamma(m\alpha)}{\Gamma(\alpha)^m} p_1 \ln(p_1) p_1^{\alpha-1} \ldots p_m^{\alpha-1} dp_1 \ldots dp_{m-1} \\
&= \sum_{i=1}^{m} \int \ldots \int \frac{\Gamma(m\alpha)}{\Gamma(\alpha)^m} p_1 \ln(p_1) p_1^{\alpha-1} \ldots p_m^{\alpha-1} dp_1 \ldots dp_{m-1},
\end{aligned}$$

where we can use that

$$\frac{\Gamma(m\alpha)}{\Gamma(\alpha)^m} = \frac{\Gamma(m\alpha)}{\Gamma(\alpha)^m} \frac{\Gamma(m\alpha+1)}{\Gamma(m\alpha+1} \frac{\Gamma(\alpha)^{m-1}\Gamma(\alpha+1)}{\Gamma(\alpha)^{m-1}\Gamma(\alpha+1)} = \frac{(m\alpha-1)!(\alpha)!}{(\alpha-1)!(m\alpha)!} \frac{\Gamma(m\alpha+1)}{\Gamma(\alpha)^{m-1}\Gamma(\alpha+1)}$$

thus,

$$\begin{aligned}
E_p(L(\boldsymbol{p},\boldsymbol{d})) &= E_p(E(L(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}) \\
&= m \int \ldots \int \frac{(m\alpha-1)!(\alpha)!}{(\alpha-1)!(m\alpha)!} \frac{\Gamma(m\alpha+1)}{\Gamma(\alpha)^{m-1}\Gamma(\alpha+1)} p_1 \ln(p_1) p_1^{\alpha-1} \ldots p_m^{\alpha-1} dp_1 \ldots dp_{m-1} \\
&= m \int \ldots \int \frac{\alpha}{(m\alpha)} \frac{\Gamma(m\alpha+1)}{\Gamma(\alpha)^{m-1}\Gamma(\alpha+1)} p_1 \ln(p_1) p_1^{\alpha-1} \ldots p_m^{\alpha-1} dp_1 \ldots dp_{m-1} \\
&= \int \ldots \int \frac{\Gamma(m\alpha+1)}{\Gamma(\alpha)^{m-1}\Gamma(\alpha+1)} \ln(p_1) p_1^{(\alpha+1)-1} p_2^{\alpha-1} \ldots p_m^{\alpha-1} dp_1 \ldots dp_{m-1}
\end{aligned}$$

From the expected value formula for $\ln p_i$ in Section A.1,

$$E_p(L(\boldsymbol{p},\boldsymbol{d})) = \psi(\alpha+1) - \psi(m\alpha+1).$$

The subjective variance of the score for the logarithmic score function is

$$\mathrm{Var}_p(L(\boldsymbol{p},\boldsymbol{d})) = E_p \mathrm{Var}(L(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}) + \mathrm{Var}\, E(L(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}).$$

Let

$$\begin{aligned}\operatorname{Var}(L(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}) &= \operatorname{Var}\left(\sum_{i=1}^{m} d_i \ln p_i\right)\\ &= E\left(\left(\sum_{i=1}^{m} d_i \ln p_i\right)^2\right) - E\left(\sum_{i=1}^{m} d_i \ln p_i\right)^2\\ &\vdots\\ &= \sum_{i=1}^{m} p_i(\ln p_i)^2 - \sum_{i=1}^{m} p_i^2(\ln p_i)^2 - 2\sum_{i<j}\sum p_i p_j \ln p_i \ln p_j,\end{aligned}$$

and let

$$\begin{aligned}\operatorname{Var}_p E(L(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}) &= \operatorname{Var}_p\left(\sum_{i=1}^{m} p_i \ln p_i\right)\\ &= E_p\left(\left(\sum_{i=1}^{m} p_i \ln p_i\right)^2\right) - E_p\left(\sum_{i=1}^{m} p_i \ln p_i\right)^2\\ &= E_p\left(\sum_{i=1}^{m} p_i^2(\ln p_i)^2 + 2\sum_{i<j}\sum p_i p_j \ln p_i \ln p_j\right) - \left(\sum_{i=1}^{m} E_p(p_i \ln p_i)\right)^2\end{aligned}$$

then,

$$\begin{aligned}\operatorname{Var}_p(L(\boldsymbol{p},\boldsymbol{d})) &= E_p \operatorname{Var}(L(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p}) + \operatorname{Var} E(L(\boldsymbol{p},\boldsymbol{d})|\boldsymbol{p})\\ &= E_p\left(\sum_{i=1}^{m} p_i^2(\ln p_i)^2 + 2\sum_{i<j}\sum p_i p_j \ln p_i \ln p_j\right) - \left(E_p \sum_{i=1}^{m}(p_i \ln p_i)\right)^2\\ &+ E_p\left(\sum_{i=1}^{m} p_i(\ln p_i)^2 - \sum_{i=1}^{m} p_i^2(\ln p_i)^2 - 2\sum_{i<j}\sum p_i p_j \ln p_i \ln p_j\right)\\ &= \operatorname{Var}_p(\ln p_i).\end{aligned}$$

From the covariance formula for $\ln p_i$ in Appendix A.1,

$$\operatorname{Var}_p(L(\boldsymbol{p},\boldsymbol{d})) = \psi_1(\alpha+1) - \psi_1(m\alpha+1).$$

The expressions for the expected score and the variance of the score for the logarithmic score function can also be found by an alternative way as shown in Appendix C, which display the same results as shown here.

Appendix C

Alternative method for the expected score and variance of the logarithmic score function

The probability of an alternative being correct, from now referred to as the event R_i, is $P(R_i) = \frac{1}{m}$, where m is the number of alternatives. Intuitively, from the participant's point of view, the probability that alternative i is correct is equal to the probability he/she assigns to alternative i is $P(R_i|p_i) = p_i$. Using this, and Equation (B.1) we find that

$$\begin{aligned} f_{p_i|R_i}(\boldsymbol{p}) &= \frac{P(R_i|p_i)f_{p_i}(\boldsymbol{p})}{P(R_i)} \\ &= m\frac{\Gamma(m\alpha)}{\Gamma(\alpha)\Gamma((m-1)\alpha)}p_i^{(\alpha+1)-1}(1-p_i)^{(m-1)\alpha-1} \end{aligned}$$

The distribution of $\ln p_i|R_i$ is the marginal Dirichlet distribution, which we see from Equation (A.2) is just the beta distribution, where $\ln p_i|R_i \sim \text{Beta}(\alpha+1, (m-1)\alpha)$. Thus, the moment generating function of $\ln p_i|R_i$ is

$$\begin{aligned} M_{lnp_i|R_i}(t) &= E(e^{tlnp_i}|R_i) \\ &= \int \frac{m\Gamma(m\alpha)}{\Gamma(\alpha)\Gamma\big((m-1)\alpha\big)}p_i^{(\alpha+t+1)-1}(1-p_i)^{(m-1)\alpha-1}dp_i \\ &= \frac{m\Gamma(m\alpha)}{\Gamma(\alpha)\Gamma\big((m-1)\alpha\big)}\frac{\Gamma(\alpha+t+1)\Gamma\big((m-1)\alpha\big)}{\Gamma(m\alpha+t+1)} \\ &= \frac{m\Gamma(m\alpha)}{\Gamma(\alpha)}\frac{\Gamma(\alpha+t+1)}{\Gamma(m\alpha+t+1)} \end{aligned} \tag{C.1}$$

Note also that the t'th conditional moment of p_i, $E(p_i^t|R_i) = M_{lnp_i|R_i}(t)$. We use the cumulant generating function to find the logarithm of the conditional moment generating function (Kendall et al., 1991),

$$\begin{aligned} K_{\ln p_i|R_i}(t) &= \ln M_{\ln p_i|R_i}(t) \\ &= \ln\Gamma(m\alpha) + \ln(m) - \ln\Gamma(\alpha) - \ln\Gamma\big((m-1)\alpha\big) \\ &\quad + \ln\Gamma(\alpha+t+1) + \ln\Gamma\big((m-1)\alpha\big) - \ln\Gamma(m\alpha+t+1) \\ &= \ln\Gamma(m\alpha) + \ln(m) - \ln\Gamma(\alpha) + \ln\Gamma(\alpha+t+1) - \ln\Gamma(m\alpha+t+1) \end{aligned} \tag{C.2}$$

To find the expected score of the logarithmic score function, we need Equation (C.1) and use the cumulant generating function in Equation (C.2). Let $\psi(t)$ denote the digamma function, where $\frac{d}{dt}\ln\Gamma(t) = \psi(t)$, then we get the following expression for the expected value

$$\begin{aligned} E(\ln p_i|R_i) = K'(0) &= \frac{d}{dt}\ln\Gamma(\alpha+t+1) - \frac{d}{dt}\ln\Gamma(m\alpha+t+1)|_{t=0} \\ &= \psi(\alpha+1) - \psi(m\alpha+1). \end{aligned}$$

Let $\psi_1(t)$ denote the trigamma function, then $\frac{d^2}{dt^2}\ln\Gamma(t) = \psi_1(t)$, then we get the following expression for the variance

$$\begin{aligned} \mathrm{Var}(\ln p_i|R_i) = K''(0) &= \frac{d^2}{dt^2}\ln\Gamma(\alpha+t+1) - \frac{d^2}{dt^2}\ln\Gamma(m\alpha+t+1)|_{t=0} \\ &= \psi_1(\alpha+1) - \psi_1(m\alpha+1). \end{aligned}$$

Appendix D

R functions

D.1 Reparameterisation of responses to original and creating a dataframe

```
# post process incoming responses
quiz <- "torunn-quiz/"

allresponses <- dir(quiz, pattern="response-*")
allresponses<-allresponses[-c
    (3,48,59,60,68,69,8,16,22,38,44,54,73,82,89)] #Every participant
    with sex=Ikke valgt is dropped(8,16,22,38,44,54,60,73,82,89) time
    spent on test is low (3,48,59,60,68,69)
m <- length(allresponses) # number of responses
n <- 27 # number of questions
allprobs <- array(NA,dim=c(m,n,4)) # store all reported probabilities
    in a m by n by 4 array
questionposition <- array(NA,dim=c(m,n))
data <- data.frame() # other variables
for (j in 1:m) {
  load(paste(quiz, allresponses[j], sep=""))
  if (nrow(response$probs) != n)
    error("Incorrect number of questions in response")
  probs <- matrix(NA,n,4)
  for (i in 1:n) { # permute back to original non-random ordering
    probs[response$order[i],response$answerorder[i,]] <- response$
    probs[i,]
  }
  allprobs[j,,] <- probs
  questionposition[j,response$order] <- 1:n # also store the random
    position of each question
  for (varname in names(response)[-(1:3)])
    data[j,varname] <- response[[varname]]
}
# Change all character variables in data to factors
for (varname in names(data))
  if (class(data[,varname])=="character")
    data[,varname] <- factor(data[,varname])
```

```
# score function rescaled to 0 to n interval where 0 corresponds to
    uniform probabilities and n perfect knowledge
data$logarithmic <- n + apply(log(allprobs[,,1]), 1, sum)/log(4) #
    compute logarithmic scores
data$quadratic <- n - apply((allprobs - rep(c(1,0,0,0),each=m*n))^2, 1
    , sum)*4/3 # compute quadratic scores

# subjective expectations
data$Equadratic <-  n - (n - apply(allprobs^2,1,sum))*4/3 #Subjective
    expected quadratic score

Elogarithmic<-c()
for ( i in 1:m){
  tmp<-0
  for( j in 1:n){
    tmp<-tmp+sum(allprobs[i,j,allprobs[i,j,]!=0]*log(allprobs[i,j,
    allprobs[i,j,]!=0])/log(4)) #Expected score=0 when p_i=0
    }
  Elogarithmic[i]<-n+tmp
  }
data$Elogarithmic<-Elogarithmic #Subjective expected logarithmic score
```

D.2 Dirichlet sampling

```
#Analysis of Expected value and variance using the dirichlet
    distribution

library("MCMCpack", lib.loc="/Library/Frameworks/R.framework/Versions/
    3.2/Resources/library")
m<-4
n<-27
a<-1.5^((-13):15)#Different values of alpha
k<-length(a)
dirp<-vector()
for(j in 1:k){
dirp[j]<-mean(apply(rdirichlet(n=1e+5,alpha=rep(a[j],m)),1,max))
}
#Expected score for the traditional MC-test
dirE<-n*(m*dirp-1)/(m-1)
#Variance of the score for the traditional MC-test
dirV<-m^2*n*dirp*(1-dirp)/((m-1)^2)


#Logarithmic theoretical expectation and variance

dirEL<-n+n*(digamma(a+1)-digamma(m*a+1))/(log(m))
dirVL<-n*(trigamma(a+1)-trigamma(m*a+1))/((log(m))^2)

#Quadratic theoretical expectation and variance
dirEQ<-n/3+n*4*(a+1)/(3*(m*a+1))
dirVQ<-n*16*(-2*(a+1)/(m*a+1)-3*(a+1)*(a+2)*(a+3)/((m*a+1)*(m*a+2)*(m*
    a+3))+6*(a+1)*(a+2)/((m*a+1)*(m*a+2))-3*(m-1)*a*(a+1)^2/((m*a+1)*(
    m*a+2)*(m*a+3))+2*(m-1)*a*(a+1)/((m*a+1)*(m*a+2))-(a+1)^2/((m*a+1)
    ^2))/9
```

```
#Mean squared error
MSEoftradandlog<-dirV+(dirE-dirEL)^2
MSEofquadandlog<-dirVQ+(dirEQ-dirEL)^2
```

D.3 Maximum likelihood estimation

```
# Reorganize the data such that questionposition can be used as a
    covariate
# reorganize the 74x27x4 array allprobs into the 1998x4 matrix p
p <- aperm(allprobs,3:1)
dim(p) <- c(4,m*n)
dim(p)
p <- t(p)
# repeat each row of data n times to form a 1998x11 data.frame and add
     a column containing questionspositions
bigdata <- data[rep(1:m,each=n),]
bigdata$questionposition <- as.vector(t(questionposition))
bigdata$p <- p
dim(bigdata)
names(bigdata)

fitmodel <- function(aformula, bformula, data, method="BFGS") {
  lnL <- function(par) {
    eta.a <- as.vector(Xa %*% par[1:na]) # "linear predictor a"
    eta.b <- as.vector(Xb %*% par[na + 1:nb]) # "linear predictor b"
    p <- data$p^exp(eta.a) + exp(eta.b) # note that eta.a and eta.b
    are recycled columnwise
    p <- p/rowSums(p) # also the rowwise sums are recycled columnwise
    sum(log(p[,1])) # first alternative correct is always the "
    observed outcome"
  }
  Xa <- model.matrix(aformula, data)
  na <- ncol(Xa)
  Xb <- model.matrix(bformula, data)
  nb <- ncol(Xb)
  start <-rep(-0.5, na + nb)
  names(start) <- c(paste("a",colnames(Xa),sep="."), paste("b",
    colnames(Xb),sep="."))
  fit <- optim(start, lnL, control = list(fnscale = -1), hessian =
    TRUE, method = method)
  fit
}
fit_0<-fitmodel(~ 1, ~ 1,bigdata)
fit_1<-fitmodel(~minimumscore+sex+feedback+sex:feedback, ~
    scorefunction+feedback+ scorefunction:feedback,bigdata)
st.error_0<-sqrt(-diag(solve(fit_0$hessian)))
st.error_1<-sqrt(-diag(solve(fit_1$hessian)))

#calculating p_prime of the suggested model and data
Xa <- model.matrix(~sex+minimumscore+feedback+sex:feedback, bigdata)
na <- ncol(Xa)
Xb <- model.matrix(~ scorefunction+feedback+feedback:scorefunction,
    bigdata)
nb <- ncol(Xb)
eta.a <- as.vector(Xa %*% fit_1$par[1:na]) # "linear predictor a"
eta.b <- as.vector(Xb %*% fit_1$par[na + 1:nb]) # "linear predictor b"
```

```
p_prime <- bigdata$p^exp(eta.a) + exp(eta.b)
bigdata$p_prime <- p_prime/rowSums(p_prime)
```

D.4 Residuals

```
# Score functions (takes a matrix p and a vector of indices of correct
    alternatives as input)
logarithmic <- function(p,correct) {
  subset <- cbind(1:nrow(p),correct)
  x<-p[subset]
  nrow(p)+sum(log(x))/log(4)
}
quadratic <- function(p,correct) {
  subset <- cbind(1:nrow(p),correct)
  d <- matrix(0,nrow(p),ncol(p))
  d[subset] <- 1
  nrow(p) - 4*sum((p-d)^2)/3
}

# given a matrix of reported probabilities p, a matrix of estimated
    probabilities
# and a score function fn, compute (using simulation)
# the probability that the score takes a value smaller than s (the
    observed score)
uresid <- function(fn, p, p.est=p, s=0, nsim=1e+4) {
  n <- nrow(p)
  m <- ncol(p)
  s.sim <- numeric(nsim)
  for (i in 1:nsim) {
    correct <- apply(p.est, 1, function(x) sample(1:4,size=1,prob=x))
    # simulate the outcome given p.est
    s.sim[i] <- fn(p,correct) # compute the corresponding score based
    on the reported probability
  }
  list(u=mean(s.sim<s), # estimated probability that the score is
    smaller than s
       inf.prob=mean(is.infinite(s.sim)), # estimated probability that
     the score takes a value of -Inf
       s=s.sim) # estimated (samples from) distribution of the score
}
U<-rep(0,m)#residuals of the estimated probability distribution based
    on the model
U_orig<-rep(0,m)#residuals of the original probability distribution
Est.log.Score<-rep(0,m)#The estimated score
Est.quad.Score<-rep(0,m)
Log_prime<-rep(0,m)
Quad_prime<-rep(0,m)
for(j in 1:m){
  J<-j*27
  p<-bigdata$p[(J-26):J,]#empirical data of the probability
    distributions
  p.est<-bigdata$p_prime[(J-26):J,]#estimated new probability
    distributions based on the model
  Result_log<-uresid(logarithmic, p, p.est, s=data$logarithmic[j])
  Result_quad<-uresid(quadratic, p, p.est, s=data$quadratic[j])
  Log_prime[j]<-n + sum(log(p.est[,1]))/log(4) # compute logarithmic
```

```
    scores based on p_prime
  Quad_prime[j]<- n - sum((p.est - rep(c(1,0,0,0),each=n))^2)*4/3 #
    compute quadratic scores based on p_prime
  Est.log.Score[j]<-mean(Result_log$s) #mean value of the estimated
    logarithmic scores
  Est.quad.Score[j]<-mean(Result_quad$s) #mean value of the estimated
    quadratic scores
  if(data$scorefunction[j]=="logarithmic"){
    U[j]<-Result_log$u
    U_orig[j]<-uresid(logarithmic, p, s=data$logarithmic[j])$u
  }
  else{
    U[j]<-Result_quad$u
    U_orig[j]<-uresid(quadratic, p, s=data$quadratic[j])$u
  }
}
#Estimated expected score
data$Est.log.Score<-Est.log.Score
data$Est.quad.Score<-Est.quad.Score
data$Log_prime<-Log_prime
data$Quad_prime<-Quad_prime
```

Appendix E

Data

Table E.1: Mean taken over all questions, alternatives and participants

Question	a	b	c	d
1	0.851	0.061	0.037	0.051
2	0.533	0.184	0.171	0.112
3	0.747	0.121	0.055	0.077
4	0.432	0.134	0.313	0.122
5	0.447	0.252	0.162	0.140
6	0.655	0.073	0.206	0.066
7	0.149	0.301	0.454	0.096
8	0.968	0.011	0.004	0.017
9	0.767	0.056	0.131	0.046
10	0.532	0.135	0.159	0.174
11	0.807	0.082	0.049	0.062
12	0.879	0.033	0.053	0.035
13	0.595	0.102	0.219	0.085
14	0.523	0.183	0.122	0.173
15	0.590	0.166	0.052	0.192
16	0.690	0.031	0.040	0.239
17	0.666	0.244	0.040	0.050
18	0.446	0.139	0.137	0.278
19	0.942	0.013	0.011	0.034
20	0.444	0.354	0.155	0.046
21	0.540	0.214	0.095	0.151
22	0.320	0.238	0.258	0.183
23	0.799	0.066	0.097	0.039
24	0.649	0.108	0.118	0.125
25	0.791	0.078	0.084	0.047
26	0.296	0.262	0.206	0.235
27	0.296	0.370	0.160	0.174

Table E.2: Mean taken over all questions, alternatives and participants for the traditional MCT

Question	a	b	c	d
1	0.905	0.068	0	0.027
2	0.608	0.189	0.122	0.081
3	0.838	0.081	0.041	0.041
4	0.514	0.108	0.297	0.081
5	0.514	0.257	0.108	0.122
6	0.784	0.027	0.176	0.014
7	0.122	0.324	0.527	0.027
8	0.973	0.014	0	0.014
9	0.838	0.027	0.108	0.027
10	0.541	0.149	0.135	0.176
11	0.892	0.054	0.014	0.041
12	0.932	0.014	0.027	0.027
13	0.662	0.041	0.270	0.027
14	0.568	0.149	0.108	0.176
15	0.568	0.135	0.054	0.243
16	0.770	0.014	0	0.216
17	0.662	0.270	0.027	0.041
18	0.446	0.108	0.149	0.297
19	1	0	0	0
20	0.459	0.459	0.054	0.027
21	0.676	0.189	0.014	0.122
22	0.378	0.284	0.243	0.095
23	0.865	0.041	0.095	0
24	0.797	0.054	0.068	0.081
25	0.865	0.041	0.081	0.014
26	0.297	0.257	0.149	0.297
27	0.311	0.459	0.135	0.095

Table E.3: Participants with minimum score and feedback

	Feedback TRUE	Feedback FALSE
MinimumscoreTRUE	15	15
Minimumscore FALSE	40	19

Table E.4: Participant's sex and score function

	Logarithmic	Quadratic
Female	12	17
Male	19	31
None	3	7

Table E.5: Participant's sex and minimum score

	Minimumscore TRUE	Minimumscore FALSE
Female	9	20
Male	20	30
None	1	9

Table E.6: Participant's sex and feedback

	Feedback TRUE	Feedback FALSE
Female	16	13
Male	33	17
None	6	4

Table E.7: Score function and feedback

	Feedback TRUE	Feedback FALSE
Logarithmic	16	18
Quadratic	39	16

Table E.8: Score function and minimum score

	Minimumscore TRUE	Minimumscore FALSE
Logarithmic	14	20
Quadratic	16	39

Appendix F

Descriptive statistics

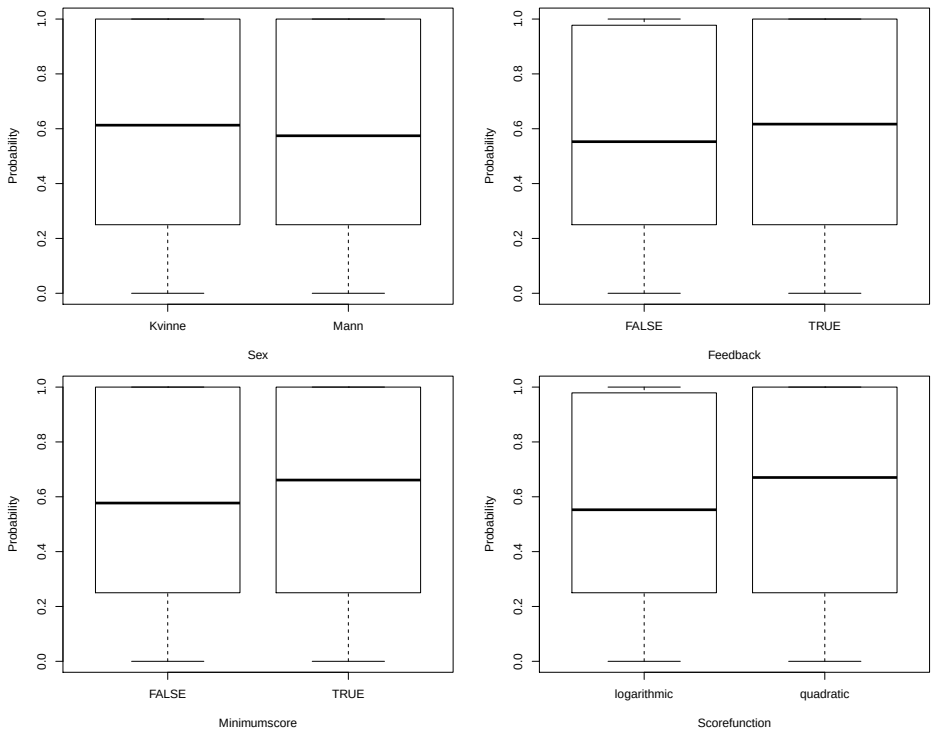

Figure F.1: Box plot of covariates Sex, Feedback, Minimum score and Score function.

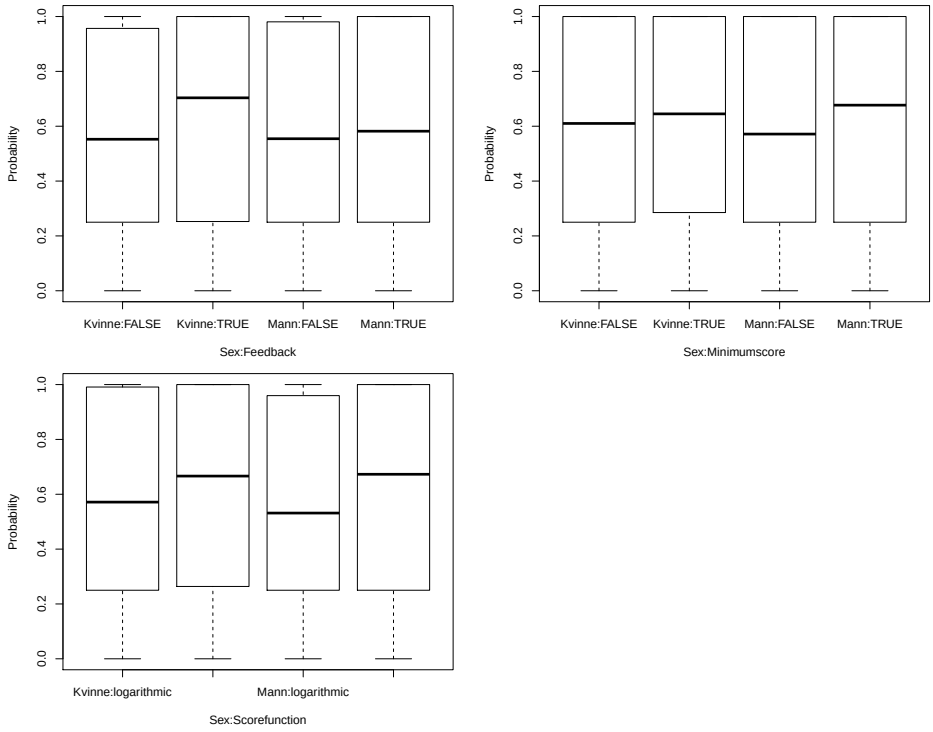

Figure F.2: Box plot of covariates Sex: Feedback, Sex:Minimum score, Sex: Score function and Minimum score: Feedback. In the upper left corner we have four boxes for the interaction covariate Sex: Feedback. The boxes have different levels of the interaction covariate, namely False: False, False: True, True:False and last True: True. We have the same order of the boxes for the boxplots in the upper right corner and lower left corner.

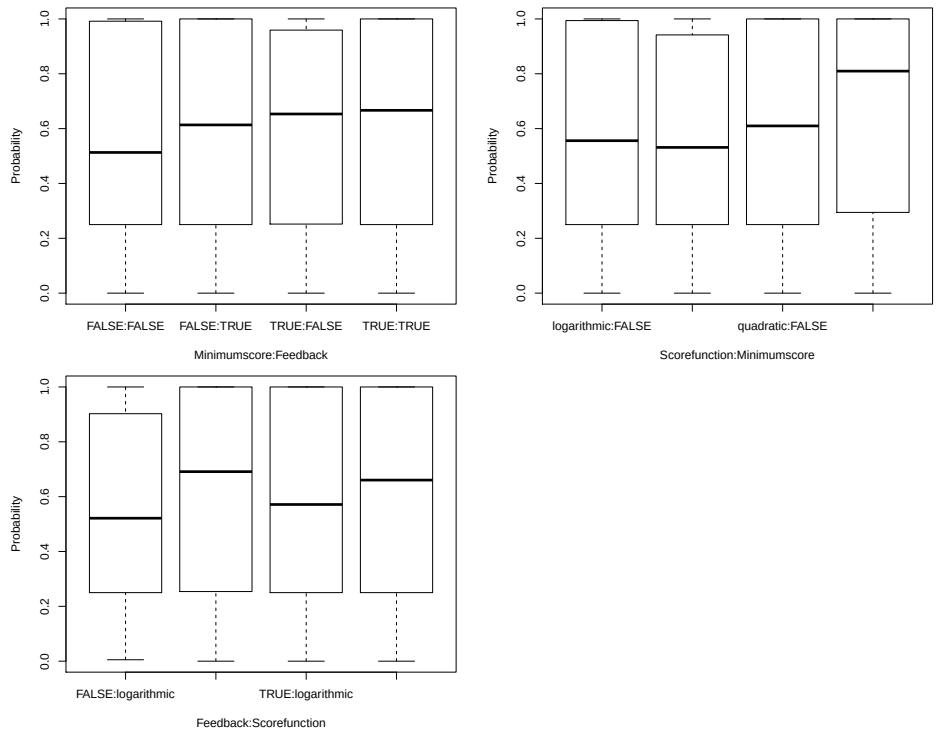

Figure F.3: Box plot of covariates Minimum score: Feedback, Minimum score: Score function and Feedback: Score function. In the upper left corner we have four boxes for the interaction covariate Minimum score: Feedback. The boxes have different levels of the interaction covariate, namely False: False, False: True, True:False and last True: True. We have the same order of the boxes for the boxplots in the upper right corner and lower left corner.

Appendix G

Quiz questions

1. Dersom

$$A = \{x : 0 \leq x \leq 4\},$$
$$B = \{x : 2 \leq x \leq 6\},$$
$$C = \{x : x = 0, 1, 2, ...\}$$

   hva er

$$A \cap B \cap C?$$

   (a) $A \cap B \cap C = \{2, 3, 4\}$,
   (b) $A \cap B \cap C = \{0, 1, 2, 3, 4, 5, 6\}$
   (c) $A \cap B \cap C = \{7, 8, 9, ...\}$
   (d) $A \cap B \cap C = 2, 3, 4, 5$

2. I en uniform diskret sannsynlighetsmodell har vi for enhver hendelse $A \subseteq S$ at

   (a) $P(A) =$(antall gunstige utfall)/(antall mulige utfall)
   (b) $P(A) = 1/(b - a)$
   (c) $P(A) = 1/k$
   (d) $P(A) = 0$

3. Hvilket år er neste skuddår?

   (a) 2020
   (b) 2018
   (c) 2017
   (d) 2019

4. Hvilken formel er korrekt for antall ordnede utvalg uten tilbakelegging av størrelse r fra n objekter?

(a)
$$\frac{n!}{(n-r)!}$$

(b)
$$n^r$$

(c)
$$\binom{n}{r}$$

(d)
$$\binom{n+r-1}{r}$$

5. På en prøve er det 6 spørsmål, hvert spørsmål har fire svaralternativer. Hvor mange permutasjoner av svar finnes det?

(a)
$$4^6$$

(b)
$$6^4$$

(c)
$$\binom{6}{4}$$

(d)
$$\frac{6!}{(6-4)!}$$

6. Blant de 15 elevene i klasse C skal det velges to tillitsvalgte. Hvor mange mulige kombinasjoner av tillitsvalgte finnes?

(a)
$$\binom{15}{2}$$

(b)
$$15^2$$

(c)
$$\frac{15!}{(15-2)!}$$

(d)
$$\binom{15+2-1}{2}$$

7. I butikken er det jordbær, pærer og klementiner. Vi skal kjøpe 4 frukter. På hvor mange måter kan dette gjøres?

(a)
$$\binom{3+4-1}{4}$$

(b)
$$\binom{4}{3}$$

(c)
$$3^4$$

(d)
$$\binom{4+3-1}{3}$$

8. Hvem er statsminister i Norge?

   (a) Erna Solberg
   (b) Jens Stoltenberg
   (c) Kjell Magne Bondevik
   (d) Siv Jensen

9. Den betingede sannsynligheten for B gitt A er

   (a) $P(B|A) = P(A \cap B)/P(A)$
   (b) $P(B|A) = P(A \cup B)/P(A)$ v$P(B|A) = P(A \cap B)/P(B)$
   (c) $P(A|B) = P(A \cap B)/P(A)$

10. Hva heter Donald Trumps gründer-datter?

    (a) Ivanka Trump
    (b) Melania Trump
    (c) Ivana Trumpt
    (d) Tiffany Trump

11. Hva heter hovedrollen i høstens sesong av skam?

    (a) Isak
    (b) Noora
    (c) Eva
    (d) William

12. Hvem er rektor på NTNU?

    (a) Gunnar Bovim
    (b) Sem Sæland

(c) Thorbjørn Digernes

(d) Eivind Hiis Hauge

13. Hvilket år ble NTNU dannet?

(a) 1996

(b) 1968

(c) 1910

(d) 1955

14. Hvem ledet den 100 timer lange p3-aksjonen på Trondheim Torg?

(a) Tuva Fellmann, Ronny Brede Aase, Niklas Baarli og Silje Nordnes

(b) Tuva Fellmann, Ronny Brede Aase, Markus Neby og Silje Nordnes

(c) Tuva Fellmann, Ronny Brede Aase, Markus Neby og Chirag Patel

(d) Tuva Fellmann, Ronny Brede Aase, Chirag Patel og Silje Nordnes

15. Hvem vant Tour de France 2016?

(a) Chris Froome

(b) Adam Yates

(c) Thor Hushovd

(d) Mark Cavendish

16. Hva er det som angis i enheten radian?

(a) Størrelsen til en vinkel

(b) Solens intensitet

(c) Diameter

(d) Vinkelbuen

17. Hvem ble tildelt Nobel fredspris i år?

(a) Juan Manuel Santos

(b) Malala Yousafzai

(c) Barack Obama

(d) May-Britt og Edvard Moser

18. Hvilket år hadde serien South Park TV-premiere?

(a) 1997

(b) 2001

(c) 1989

(d) 1996

19. Hva er det som angis med måleenheten newton?

    (a) Kraft
    (b) Akselerasjon
    (c) Varme
    (d) Friksjon

20. I hvilke byer foregår handlingen i den siste Sex og Singelliv-filmen?

    (a) New York og Abu Dhabi
    (b) New York og Dubai
    (c) New York og Sharjah
    (d) New York og Trondheim

21. Hva het kongssønnen som birkebeinerne Torstein Skevla og Skjervald Skrukka reddet fra baglerne i 1206?

    (a) Håkon Håkonsson
    (b) Harald Haraldsson
    (c) Christian Christiansson
    (d) Olav Olavsson

22. Hvilke tre land har vært hardest rammet av Ebola?

    (a) Sierra Leone, Liberia og Guinea
    (b) Sierra Leone, Nigeria og Guinea
    (c) Sierra Leone, Mali og Liberia
    (d) Liberia, Mali og Guinea

23. Hva heter hovedstaden i Australia?

    (a) Canberra
    (b) Sydney
    (c) Melbourne
    (d) Perth

24. Hvem eier brusmerket Solo?

    (a) Ringnes
    (b) Coca Cola
    (c) Hansa
    (d) E.C. Dahls

25. Hvilken by kom The Beatles fra?

(a) Liverpool
(b) Manchester
(c) London
(d) Birmingham

26. Hvilket år ble internett tilgjengelig for vanlige folk?

(a) 1995
(b) 1992
(c) 1989
(d) 1997

27. Hvem er den personen som har blitt googlet mest i Norge i 2015?

(a) Caroline Berg Eriksen
(b) Martin Ødegaard
(c) Pablo Escobar
(d) Jens Stoltenberg