Karlsruhe Institute of Technology

Experimental Analysis of Smart Meter Information

Bachelor Thesis
of

Christoph Großbaier

At the Department of Economics and Business Engineering
Institute of Information Systems and Management (IISM)
Information & Market Engineering

Reviewer: Prof. Dr. rer. pol. Christof Weinhardt
Advisor: M.Sc. Anders Dalén

XX. Monat 20XX

Contents

List of Tables

List of Figures

List of Acronyms

HIT Human Intelligence Task

AMT Amazon Mechanical Turk

DPA Dynamic Programming Algorithm

GA Greedy Algorithm

JSON JavaScript Object Notation

CSV Comma-Separated Values

GUI Graphical User Interface

NP Non-Parametric Statistical Tests

LMM Linear Mixed Model

1. Introduction

> ”Most domestic energy use, most of the time, is invisible to the user”
>
> Darby (2006)

When it comes to energy consumption in domestic households, there is a limited understanding about how much energy is used for different purposes - we only have a vague idea about how much energy we consume during our daily routines. Just compare the energy consumption with the weekly trip to the grocery store. In a normal grocery store, every item has its price tag, offering full transparency about what we pay to use it. Now imagine a grocery store, where you can buy the energy usage of the fridge, the room lightings, etc. None of these offered items would indicate an individual price marking. The bill is not presented directly, but would come on a monthly or even annually basis and would show the accumulated price of all items bought. You would not have any idea how or when you bought these items or how much they were individually. Furthermore, you would not have any idea if your consumption was low or high compared to your peers, or how your spendings developed over time (Kempton and Layne, 1994). As a result, we only get a limited understanding about what impact a change of daily behaviour has on our energy bill (Darby, 2006). Furthermore a potential benefit of replacing a non-energy efficient household device with a new energy efficient appliance is unknown.

This lack of timely information prevents consumers from using energy more intelligently and efficiently (Darby, 2000). Studies have shown that introducing metered energy consumption in domestic households, and providing regular feedback and suggestions, have measurable effects on total energy use and is worth pursuing. Darby analysed the results of 38 feedback studies and came to the conclusion, that ”direct feedback, almost or in combination with other factors, is the most promising single type”. Direct feedback is available on demand and users can access information about their energy usage via direct displays, smart meters and interactive feedback. The learning approach is in a ”looking or paying” sense, since direct feedback requires an active attitude from the customer. Some of the surveyed studies showed that direct feedback in combination with some form of advice or information enabled a cost savings of up to 10%.

Smart meters are the most prominent example of providing direct feedback to households. The value of the smart meters' home installation is that each unit can tell households, on demand, how high the energy consumption is at the moment. Moreover, it can add additional information like the consumption of individual appliances and the efficiency of installed devices compared to a potential new one. The household's benefit from the information provided by smart meters, however, depends on whether the information is tangible to the user. Therefore, we define the research question for this thesis:

> What information do consumers need to know about the objects they own or manage to increase their energy efficiency?

We try to define a level of information provided by smart meters that is sufficient for users to help them reduce their energy consumption, but does not overburden them with too much information, resulting in the danger of an information overload.
In order to find the optimal level of information load in a smart meter environment, we design an behavioural experiment that describes a situation in which a user of a smart meters faces the following task: the user must choose out of several options in order to reduce his or her energy consumption. Thereby, the optimal choice is influenced by two driving factors, the personal benefit and the cost of the option (e.g. kWh).
In our experiment, we transfer this choice environment into a knapsack-problem. In this type of optimization problem, one must choose between several items that show two characteristics: a benefit and a cost. The goal is to maximize the cumulated benefit of all chosen items without extending a given cost restraint.
This thesis is a first step to design this experiment. We abstract from the energy context and introduces a plain version of the experiment. So we do not refer to energy cost and appliance's benefit when referring to cost and benefit, but try to identify the driving forces behind the experiment without the affiliation with the energy context. A second step will then use the results from this experiment and transfer the experiment to an energy context.
By finding the optimal level of information detail for smart meters, we can contribute to develop smart meters that will help consumers to reduce their energy consumption in the future.
The remainder of this thesis is organized as follows. Section 2 gives a brief overview of the underlying literature, describing the current research about smart meters and provides scientific background about the phenomenon "information overload". Afterwards, we define our main hypotheses. Section 3 introduces the used experiment and gives details about the descriptives of the gained data. Section 4 evaluates the outcomes of the experiment and tests the hypotheses. Finally, Section 5 concludes the findings of the experiment and gives an outlook about potential fields of future research.

2. Literature Review

This section aims to provide an overview over the current research on information feedback on energy consumption. In particular, we will concentrate on direct feedback provided by smart meters. In addition, the phenomenon of "information overload" is introduced and an information overload in a smart meter environment is described. A last step will then define the hypotheses that are derived from the literature review and applied on our experiment.

2.1. Smart Meter

According to Darby (2008), a smart meter is defined as

> *a meter that stores information and gives accurate consumption data at specified intervals to suppliers and consumers.*

The full potential of smart meters enables benefits for consumers, suppliers and regulators by defining a new level of communication between these parties. This can lead to a positive impact on overall consumption, on load management and on consumer retention.

Benefits for consumers include access to on-demand information and, in combination with user-friendly feedback, the opportunity to reduce the energy bill. Suppliers can benefit from better demand-management by providing incentives for customers to reduce peak-time consumption or to save the expenses of manual meter reading. Smart meters also open the market to a "smart home" environment, which includes a remote energy management system for individuals to control aspects such as lights and heating. This new relationship between suppliers and consumers is not only beneficial to consumers and suppliers, but can also support regulators in their pursuit of reducing carbon emissions (Darby, 2008).

Potential Information provided to consumers

The data provided by the smart meter can cover a wide range of different aspects[1] of household energy consumption such as:

[1] Refer to (Anderson and White, 2009).

- Power consumptions of individual devices
- Energy costs of individual devices
- Household baseload consumption
- Range of household's energy consumption levels
- Individual definition of a high level of consumption
- Typical daily consumption or spending
- Patterns of energy consumption over the week
- Link between individual and collective energy consumption

This wide variety of information delivers a detailed insight into the energy consumption of an household and might help consumers to reduce and optimize their energy consumption. Providing more and more information to an household, however, might run into the risk that the provided data is not intelligible to users any more. According to Anderson and White (2009), an ideal smart meter does not only have to make energy usage visible, but should also capture the user's interest by providing useful information while minimizing data that might be deemed too detailed or too complex.
Previous smart meter experiments support Anderson's hypothesis. They came to the conclusion that the potential for receiving energy feedback depends on the capability of the user to understand and process the given information. Feedback that contains too much information can overwhelm the user and reduce his or her capability to process the information (Henryson et al., 2000). Consequently, adding information or tools may complicate a decision rather than making it easier (Darby, 2006) and users might face an *information overload* (Fischer, 2008). Therefore, Anderson and White (2009) underline the necessity to further research the question,

> How can an information overload be avoided in a smart meter environment?

2.2. Information Overload

Research on information overload has a long tradition in marketing science. The basic idea behind information overload is that people tend to make poorer and less effective decisions when being presented with too much information at any given time (Streufert and Driver, 1965). This results in the question, how much information is too much for individuals?
But before describing different levels of information load, one must first define how to find an accurate measure of information load, e.g. to make the outcome of different experiments comparable.

2.2.1. Measure of information load

The traditional approach measures the information load by counting the number of alternatives and attributes presented to the consumer (Chen et al., 2009). This two-dimensional approach is however criticized in more current literature and new measuring techniques of information have evolved. Information load was now not only considered as a product of alternatives and attributes, but as being influenced by many factors, including the information structure (Lurie, 2004), information quality (Keller and Staelin, 1987), time pressure, the diversity of information dimension (Payne, 1982) and information repetitiveness (Hwang and Lin, 1999). More recently, the distribution of attributes and alternatives has been added into the pool of factors which might influence the information load (Lurie, 2004). All in all, the issue of how to measure information load is a highly debated topic in the literature. And even though the magnitude of the influence from both attributes and alternatives is still a focus of current research, these factors are still commonly used as a base for information load experiments.

2.2.2. Impact of information overload

Since there is no common academic ground about how to define information load, the research on the impact of an information overload returns mixed and inconsistent results. One of the first main results from studies conducted by the traditional approach is the inverted U-shaped relationship between the amount of information and the choice accuracy (Jacoby et al., 1974). Jacoby concluded that there is an optimal information load for consumers to make the best and most effective decision and consumers *can* be overloaded with information. Nevertheless, individuals *will not* be overloaded since they "are highly selective in how much and just which information they access". Thus, facing an information overload leads individuals to focus selectively on the information they feel is important. Later studies, however, have questioned Jacoby's thesis, in particular the inverted U-shape relation [2], since Jacoby's results could not be re-produced. Nevertheless, an impact of information load on choice accuracy was still detected by the majority of experiments.

2.3. Information Overload in a knapsack problem

In this paper, we compare the situation where a consumer needs to make an efficient decision on which of the appliances belongs to an energy-efficient portfolio with the knapsack problem. The knapsack problem describes an optimization problem in which a person chooses out of a set of items with individual weights and benefits. The aim of the optimization is to maximize the cumulated benefit of all the chosen items while not exceeding a given weight restriction.

By transferring the abstract level of the knapsack problem to an energy efficiency context, one can imagine that household appliances have a benefit and a weight for a consumer. The benefit of the appliance is how good the appliance meets the requirements of its owner, and the weight could be its energy consumption. The optimization problem in this context is to pick the best selection of appliances that both meet the personal requirements and do not exceed an individual budget. This can include replacing or eliminating an existing appliance. The information overload in a knapsack problem is produced by a high number of items to choose from, ergo a high number of alternatives, and a constant number

[2] Refer to Malhotra et al. (1982)

of attributes, weight and benefit. The parameter that is manipulated is the number of attribute levels, the so-called information granularity.

2.4. Hypotheses

The experiment evaluates the choice accuracy of the participants, the time it takes them to make a decision, the time it takes an individual to complete one round of the experiment, and the number of decisions per round. Choice accuracy is measured as how close the cumulated benefit of the knapsack is to the optimal solution of the optimization. The closer the total benefit of the knapsack gets to the optimal solution, the better the choice accuracy. In order to analyse the choice accuracy in a knapsack setting, we consider three different points in time. First, the total benefit of the first solution (i.e. the first time the knapsack is filled and no further items can be added). Second, the total benefit of the best solution (i.e. the set with the highest cumulated value). Third, the total benefit of the final solution (i.e. the last set represented in the knapsack). Furthermore, we evaluate the number of decisions made in one round and the average time it took a participant to make a decision.

Information granularity

The goal of this thesis is to answer how participants respond to different levels of information granularity. Information granularity in the experiment is designed as the number of colours representing the benefit. Two or three different colours are considered a low level of information granularity, seven colours are a mediate level and eleven and fifteen colours are defined as a high level of information granularity. For the experiment we follow the hypothesis of Jacoby et al. (1974) which forms an inverted U-shape relation between information load and choice accuracy. Consequently, we argue that a mediate level of information load will have the best choice accuracy.
Hypothesis 1 (H1a-H1c): The total benefit of the first(a), best(b), final(c) solution is best on a mediate level of information detail.

Jacoby et al. (1974) indicated that the relationship of the information granularity[3] is curvilinearly correlated to the time it takes a person to make a decision. From a low level to a mediate level of information granularity, the time spent on one decision increases with the information load. After a specific level of information granularity, a decision becomes more and more complex (Hendrick et al., 1968). Therefore, individuals tend to give up trying to compare alternatives, and instead make their choices impulsively. In doing this, individuals simplify their information processing by ignoring some portion of the provided information (Malhotra et al., 1982). As a result, the time it takes to make an individual decision is high on a mediate level of information detail. However, the number of decisions is the lowest for the medium level since the choice is more accurate. Therefore an individual concentrates on the quality of the choice rather than by the quantity.

[3] Jacoby et al. refer to Hendrick et al. (1968) who defined complexity by the number of different dimensions of an attribute. In our case, this corresponds to the number of possible colours per item, our definition of information granularity.

Hypothesis 2 (H2): The average amount of time per decision is the highest on a mediate level of information detail.
Hypothesis 3 (H3): The number of decisions is the lowest on a mediate level of information detail.

Learning effect

The second effect to evaluate is whether participants learn from playing the game and can adapt to an information overload. Jacoby et al. (1974) emphasized that individuals develop the ability to accommodate large amounts of information. Thus, we expect participants to learn to cope with similar information loads the more they experience it. In the experiment, participant play three rounds of the knapsack problem with a constant information load. We therefore argue that participants will improve their choice accuracy with each round.
Hypothesis 4 (H4a-H4c): The first(a), best(b), final(c) solution increases with the number of repetitions of the task.
In addition, we expect participants to make quicker decisions with a growing experience, and a decreasing amount of decisions since the quality of the decisions increase and therefore fewer decisions are necessary.
Hypothesis 5 (H5): The average amount of time per decision decreases with the number of repetitions of the task.
Hypothesis 6 (H6): The number of decisions decreases with the number of repetitions of the task.
Since the total time is the product of the average decision time and the number of decisions, we argue that the total time it takes participants to reach a solution decreases with the amount of rounds played.
Hypothesis 7 (H7a-H7c): The first(a), best(b), final(c) solution is reached faster with the number of repetitions of the task.

Table 2.1.: Treatments and Hypotheses

Treatment	**# Colours**	**Hypotheses**	
1	2	H1a-c:	Treatment 1+2 $<$ 3 $>$ 4+5
2	3	H2:	Treatment 1+2 $<$ 3 $>$ 4+5
3	7	H3:	Treatment 1+2 $>$ 3 $<$ 4+5
4	11	H4a-c:	Round 1 $<$ 2 $<$ 3
5	15	H5:	Round 1 $>$ 2 $>$ 3
		H6:	Round 1 $>$ 2 $>$ 3
		H7a-c:	Round 1 $>$ 2 $>$ 3

3. Experiment

The knapsack problem introduced in Section 2.3 is used to test the described hypotheses in our experiment. A Graphical User Interface (GUI) is designed to provide an intuitive approach to participants. The experiment is conducted using the Amazon Mechanical Turk (AMT) website and it reached 400 participants.

3.1. Knapsack Problem

The knapsack problem used in this thesis is the 0-1-knapsack problem. This problem maximizes the cumulated value of items under a given weight restriction.

$$\begin{gathered}\max_{x_j} \sum_{j=1}^{n} v_j x_j \quad s.t. \sum_{j=1}^{n} w_j x_j \leq W \\ x_j \in \{0,1\} \quad for \quad j = 1, \ldots, n \\ v_j := \text{value of item j} \quad w_j := \text{weight of item j}\end{gathered} \tag{3.1}$$

Table 3.1 shows the given parameters for the considered 0-1 knapsack problem. The width of the knapsack, or the pixel width, represents the given weight restriction for the problem. There are 100 items to choose from and the weight of each item ranges from 20 to 80 units, and the benefit is defined as a number between 0 and 80.

Since the knapsack problem is an NP-hard problem, individuals are faced with a high rate of complexity and are further challenged with finding a solution in reasonable time.

Parameter	Fixed value
W	694
n	100
w_j **range**	[20,80]
v_j **range**	[0,80]

Table 3.1.: Fixed parameters for the knapsack problem

A "greedy approach" is used to model a potential game strategy used by participants to tackle the optimization. The optimal solution is calculated by a pseudo-polynomial time algorithm using dynamic programming.

Greedy Algorithm

The greedy approach can be split into two steps. First, the items are sorted according to the ratio of their benefit and weight in descending order. Second, the items are added to the knapsack as long as it is not full.

Algorithm 1 Greedy Algorithm

```
/* items:=all available items, W:= knapsack limit  */
Input: W, items
Result: Greedy Algorithm (GA) solution to knapsack problem
Sort items according to the ratio in descending order
knapsackSize=0, knapsackValue=0
for item in items do
    if knapsackSize+item.weight ≤ W then
        knapsackValue+= item.benefit
        knapsackSize += item.weight
    end
end
return knapsackValue
```

Dynamic Programming Algorithm

Since the greedy approach does not guarantee to find the most efficient, optimal solution for the knapsack problem, a different approach is used to calculate the optimal solution. The Dynamic Programming Algorithm (DPA) introduced by Kleinberg and Tardos (2005) divides the problem into sub problems and can be solved in O(nW) time with W being the maximum weight of the knapsack.

Algorithm 2 Dynamic Programming Algorithm

```
/* items:=all available items, W:= knapsack limit  */
Input: W, items
Result: DPA solution to knapsack problem
i = 100
R = Matrix (100 x 694), each element with value 0
for item in items do
    for j=0 to limit do
        if item.weight ≤ j then
            R[i][j] = max(item.benefit+R[i+1][j-item.weight], R[i+1][j])
        end
    end
    i - -
end
return R[i][limit];
```

Comparability of the problems' difficulty

To ensure the value of the statistical analysis across treatments and rounds, the underlying knapsack problems are designed to be sufficiently similar. For this purpose, a benchmark is introduced to compare the difficulty of knapsack problems and to design equally difficult problems.
The *benchmark* is computed as the relative difference between the computed solution by the DPA and the GA. The greater the benchmark, the harder the problem. The setup algorithm runs 100 times to ensure a sufficiently high benchmark.

$$benchmark = \Big(\frac{\text{DPA solution}}{\text{GA solution}} - 1\Big) * 100\% \tag{3.2}$$

Comparability across treatments and rounds

Four different knapsack problems are created in our experiment, one for each round plus the Trial round. To ensure the **comparability across rounds**, the benchmark is used to identify hard problems. For each round, the benchmark is calculated and the values are checked to be sufficiently similar.
In order to compare the performance of participants in **different treatments**, the underlying knapsack problems are the same for each individual round. Consequently, the

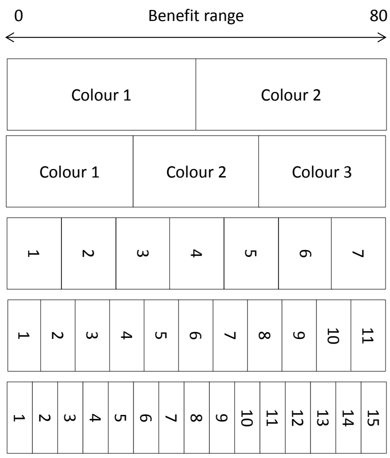


Figure 3.1.: Colours assigned to benefit range

distribution of the benefits and weights of each box is the same, the only difference is the assignation of colours. In the 2-colour treatment, one colour covers a wide range of benefit values, whereas in the 15-colour setup, one colour represents a smaller range of colours (refer to Figure 3.1). Consequently, there are two opposing trends when increasing the number of colours. The more colours there are, the more alternatives there are to choose from, so the complexity of the task might increase. In contrast to that, the range of the benefit within one colour gets smaller, so participants can better distinguish between different benefit values.

Setup

The setup of the game[1] is prepared before uploading the game. Each treatment is assigned 4 rounds with 100 boxes per round. The first round is the Trial round[2] while the other 3 rounds are shown during the actual game. Each box is randomly[3] assigned a weight in the range of 20 to 80 units. This value represents the width of the boxes in pixel in the game.

3.2. Interface

The items the participant can choose from are modelled as boxes. The width of each box represents its weight; every box has the same height. Thus, the wider the box, the greater the weight. The colour of the box represents the benefit.

Colour scale

The colour scale used for each treatment aims to provide an intuitive range of colours that are also easily distinguishable. A traffic light colour scale was first favoured since it is a well-known and intuitive colour pattern. Trial runs, however, showed that participants had problems distinguishing the colours when they were dealing with 11 or 15 colours. Green boxes were especially hard to differentiate for test participants. Furthermore, a colour-blindness test would have been necessary to identify colour-blind users. As a result, we use a grey scale as shown in Figure 3.2. This colour scale is both intuitive and accessible to colour-blind users and the grey levels are easy to distinguish.

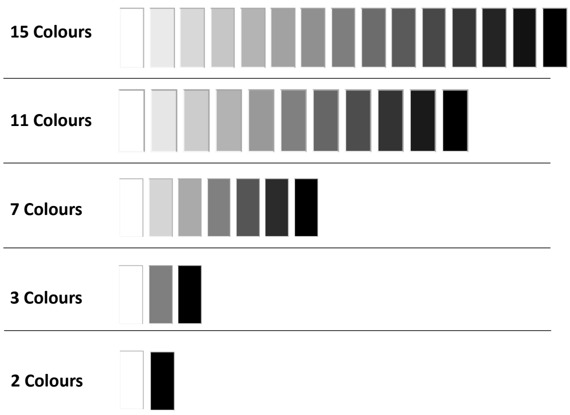


Figure 3.2.: Grey scale per number of colours

The colour scale is presented individually per treatment to the participant on the left of the interface (Figure 3.2). Participants can choose from 100 boxes for each round they play. The large number of boxes ensures an information overload, but still enables participants to achieve a satisfying result. The knapsack is illustrated as a wide rectangle and is limited by its width. Boxes can be added to and removed from the knapsack by clicking

[1] For the algorithm, refer to Appendix A.
[2] No data was conducted from the Trial round.
[3] The used random function returns a randomly determined value in a given range. The return values are uniformly distributed (Source code).

on them. In the cases that a selected box doesn't fit into the knapsack because it exceeds its capacity, the knapsack will shake and appear in red, not allowing the selected box to enter the knapsack.

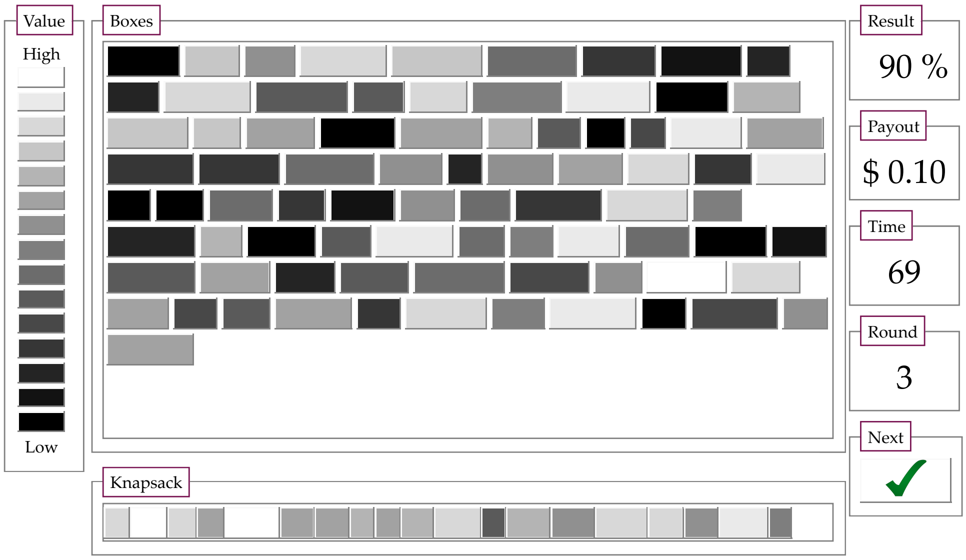


Figure 3.3.: Exemplary interface (Treatment 5, Trial 1)

3.2.1. Description of the toolboxes

Result

The performance of the participants is measured by comparing the current value of the knapsack to the optimal solution calculated by the DPA introduced in Section 3.1.

$$x = \frac{\text{Current value of all boxes in knapsack}}{\text{DPA solution}} \tag{3.3}$$

Payout

In correspondence to the result, participants receive monetary compensation for playing the current round. Every time the bonus bar[4] is reached, the toolbox turns green to get the participant's attention to further work on improving the result.

Time

The time restriction is set to 100 seconds per round. This rather long duration is chosen because participants should not experience a significant time pressure. The main goal is to isolate the colour effect; a potential time pressure effect is not the focus of the research since time has an impact on the decision quality (Hahn et al., 2006). A time limitation, however, must be implemented to ensure that it is necessary to play the round in one sitting (refer to subsection 3.3.1). The *Next* button is available to give participants the opportunity to skip the current round when they have already reached a personally satisfying result and do not want to wait for the round to be finished.

[4]Refer to Subsection 3.2.2.

Round

Participants must complete three rounds. Multiple rounds are chosen to identify a potential learning effect; the limitation to three rounds keeps the total time for the experiment to a level that is attractive for an AMT experiment.

3.2.2. Payout scheme

The implemented payout scheme has two components – a guaranteed payout for each completed round and a payout dependent on the participant's performance. A guaranteed payout is necessary to advertise the experiment on AMT and an attractive offer draws a larger participant pool. The bonus leverage is increased in Trial 2 so the differences in the payout schemes must be taken into account in the analysis.
To ensure that participants would not only hit the *Next*-button in each round and still get the minimum payout, a restriction is designed to only pay out participants who clicked at least one box per round.

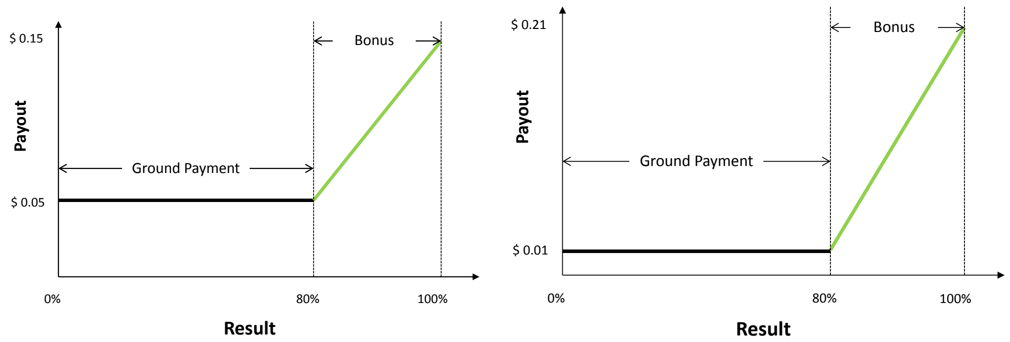


Figure 3.4.: Payout scheme for Trial 1 and 2

The reason for choosing a bonus system is to get participants to compare boxes of different weights and benefits and to make them try different combinations for the knapsack. In order to incentivize such behaviour, a bonus system can help. Trial rounds show that getting up to 80% is possible when participants make an effort and try out different solutions. Hence, the performance incentive starts at 80% and offers the participants in the first Trial an additional bonus of up to $0.10 when reaching 100%, resulting in a potential total payout of $0.15 per round. Nevertheless, the experience of Trial 1 shows a large number of participants who did not reach the bonus bar. Therefore, we increased the leverage of the bonus system in Trial 2, from a ground payment of $0.01 to a possible total payout of $0.21 when reaching 100%.

3.3. Online Implementation

For the setup of an online environment, the Python web framework Django 1.4.3 is used to implement the communication between GUI and the server. The experiment is advertised on Amazon Mechanical Turk (AMT) as a Human Intelligence Task (HIT).

3.3.1. Amazon Mechanical Turk

A previous research study (Schmidt, 2012) using a similar interface and setup reached 28 participants. While the results helped to prove that more colours to choose from does not increase the average payout, the overall experiment lacked a wider statistical foundation. Consequently, we try to increase the foundation by using the online labour market Amazon Mechanical Turk (AMT).

The AMT platform is the most active online labour market for conducting behavioural experiments. It enables researchers to conduct experiments very quickly, cheaply (Rand, 2012), and in good quality (Gardner et al., 2012). Moreover, the data gained is at least as reliable as the data obtained via traditional methods (Buhrmester et al., 2011). AMT connects employers with potential workers who are paid according to the satisfactory completion of the assigned task. Furthermore, workers can be motivated to perform well by a bonus structure. Since individuals can participate entirely over the computer, the experience is quite similar to participating in computer simulations.

Researchers can act as employers on the AMT website by offering experiments as tasks otherwise known as a Human Intelligence Task (HIT). Completing an HIT usually results in a payment of less than $1.00 for less than 5 minutes of work. Users are from around the world, but mainly concentrated in the United States and India (Rand, 2012). These findings are supported by our experiences.

Limitations of Amazon Mechanical Turk

Although the AMT offers an enormous potential for the purposes of our study, there are several limitations to keep in mind.

First, in contrast to experiments conducted in a lab environment, AMT experiments have limited control over what kind of individuals take part in the experiment. In particular, the lack of control over *non-random attrition* (Rand, 2012) defines a limitation for our experiment in terms of the participant structure. Individuals who are overwhelmed with the complexity of the task might drop out early, not finish the HIT and may not be included in the statistical analysis. This can have an impact on the pool of participants among different treatments and can therefore limit our ability to compare results. As a result, we have tried to make the description of the game and the game itself as simple and intuitive as possible in order to reduce the risk of early drop-outs. Trial runs previous to offering the HIT revealed that using the Trial round helped individuals to get a better understanding of the game and the task.

Second, researchers cannot be sure what participants are actually doing during the experiment on the AMT platform. This fact especially reduces the benefit to cognitive load experiments similar to our experiment since our experiment necessitates the full attention of each individual. We try to tackle this issue by setting a time restriction per round so individuals have an incentive to stay on task. Moreover, we log the time that individuals spend on different parts of the experiment and evaluate the participant's actions by analysing the recorded data. We are therefore able to better understand each individual's actions. Last, we take this limitation into consideration in the choice of a statistical model.

Third, the statistical analysis is based on the assumptions that every observation is independent. In the setup of our experiment, individuals are able to conduct the experiment over and over by re-directing to the start page after finishing a session. Even though repeat responses appear to be a minor concern according to Berinsky et al. (2012), it can still potentially dilute the observation independence and the statistical value of the learning effect. We therefore record the IP address of each user when they are directed to the welcome page in order to exclude those participants who play several games successively. However, this method does not exclude those individuals who change their IP in-between experiments. Moreover, the AMT platform has other disadvantages against a lab experiment in terms of user support during the experiment, and the lack of control over the English language competencies of individuals.

After summarizing the benefits and limitations of the AMT platform, we conclude that AMT is feasible for our experiment when the design and evaluation of the experiment takes the limitations into consideration.

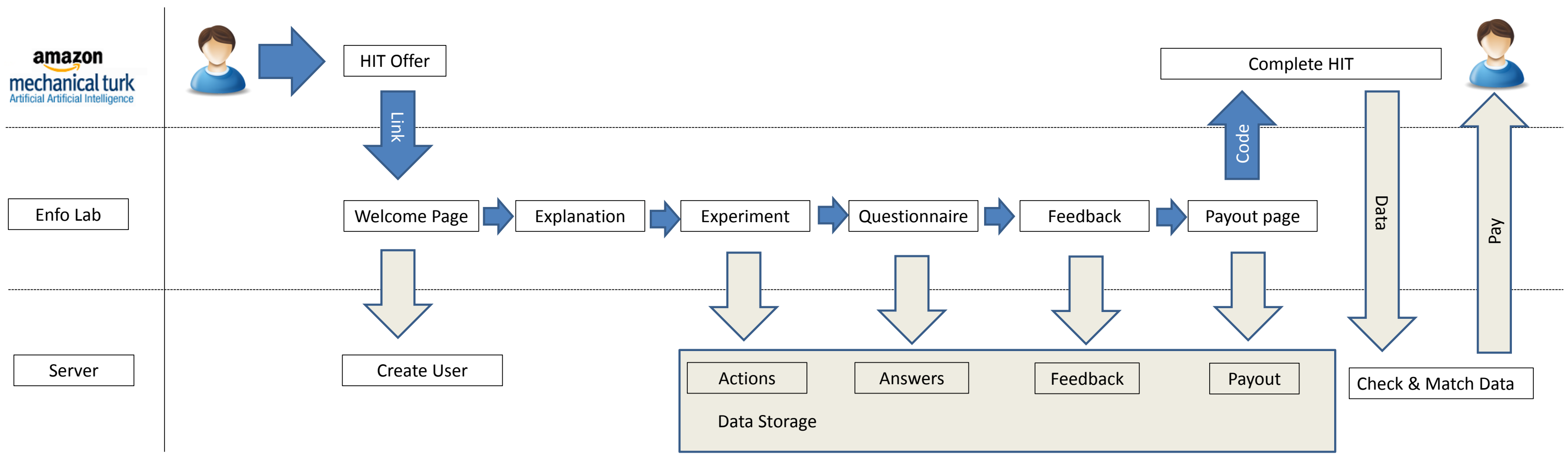
3.3.2. System design

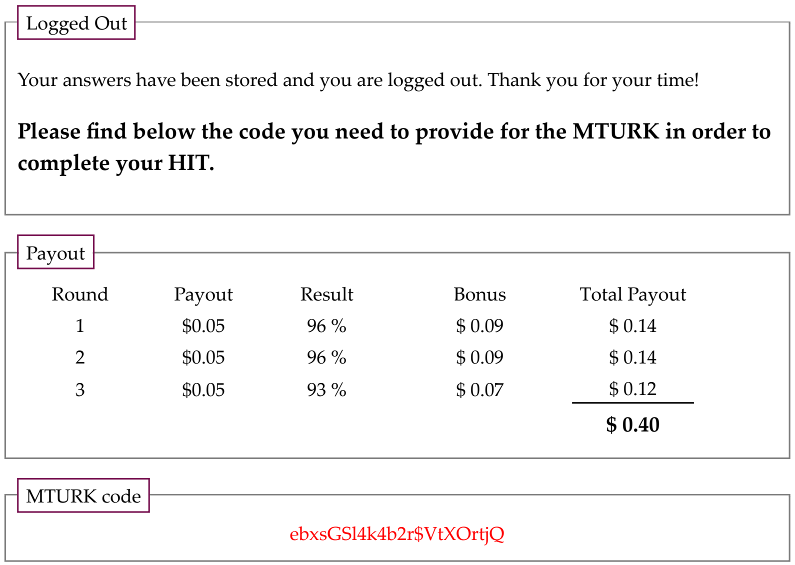
Rand (2012) indicates that in order to conduct an experiment using a game environment on AMT, researchers must provide participants with instructions, making sure they understand the rules of the game. Next, researchers must process the data and pay the participants according to their earnings. We follow this setup for our experiment (refer to Figure 3.5).

Once the AMT user has decided to complete the HIT, he or she is redirected to an external web page on which the experiment is run. When arriving at the page, the program creates an individual user who is randomly assigned to one treatment and an individual code is generated. Additionally, a time stamp is saved in order to track the time a user will spend on different stages throughout the experiment.

The participants first go through the explanation of the game where they are introduced to the concept of the knapsack problem and the implemented graphical interface. Before the participants begin playing their three rounds, they can get used to the interface by playing a Trial round which has no time restriction and offers information boxes to explain the different parts of the interface.
The participants then start the game by playing the three rounds of the experiment. Every action that includes adding or removing a box to or from the knapsack is saved on the server, including time, box ID, box benefit, box weight, current payout, the type of action (remove or add) and whether or not the knapsack is full. The JavaScript Object Notation (JSON) is used to transfer the determined data to the server. The collected data can be extracted from the server via a Comma-Separated Values (CSV) file.
After completing the game, participants are presented with a questionnaire where they must answer six questions about their experience playing the game and the strategy they

Figure 3.5.: Design of the online environment





Logged Out

Your answers have been stored and you are logged out. Thank you for your time!

Please find below the code you need to provide for the MTURK in order to complete your HIT.

Payout

Round	Payout	Result	Bonus	Total Payout
1	$0.05	96 %	$ 0.09	$ 0.14
2	$0.05	96 %	$ 0.09	$ 0.14
3	$0.05	93 %	$ 0.07	$ 0.12
				$ 0.40

MTURK code

ebxsGSl4k4b2r$VtXOrtjQ

Figure 3.6.: Final page with payout

16

used. To avoid participants rushing through the questionnaire, a minimum time of 15 seconds is introduced. If participants spend less time on the questionnaire, they are redirected to the beginning of the questionnaire. In addition, we include an additional question in Trial 2 that is directed to evaluate the participant's attention, by asking them to give a specific answer to the question.
Subsequent to the questionnaire, participants have the opportunity to give feedback by filling in a text box. This feedback is used to have an additional feedback channel to identify problems in any aspect of the interface.
The final payout for the experiment (Figure 3.6) is shown to the participants after they have completed their three rounds and the questionnaire.
When the minimum requirement of clicking at least one box per round is fulfilled, participants are presented the payout for each round and the total payout. Furthermore, the AMT code is provided. The user then goes back onto the AMT system, enters the AMT code and completes his or her HIT.
At the end, the data on the server is compared to the HITs completed, and the payout is made via the AMT system.

3.4. Data acquisition

The 1st Trial was offered on AMT between the 20th and the 21st of February 2013 and reached 200 participants, the 2nd Trial with the same amount of participants ran between the 5th and the 8th of March 2013. A total of 31.291 actions were recorded on the server. Before the data is statistically analysed using IBM's **SPSS** and **R**, the recorded data is cumulated on a per round basis for the purposes of this study. For each round per user, a data point is created that includes information on the user, treatment and corresponding number of colours, round, number of decisions, first result, best result, final result and finally the time for each result. Out of a potential 1.200 rounds per 400 participants,

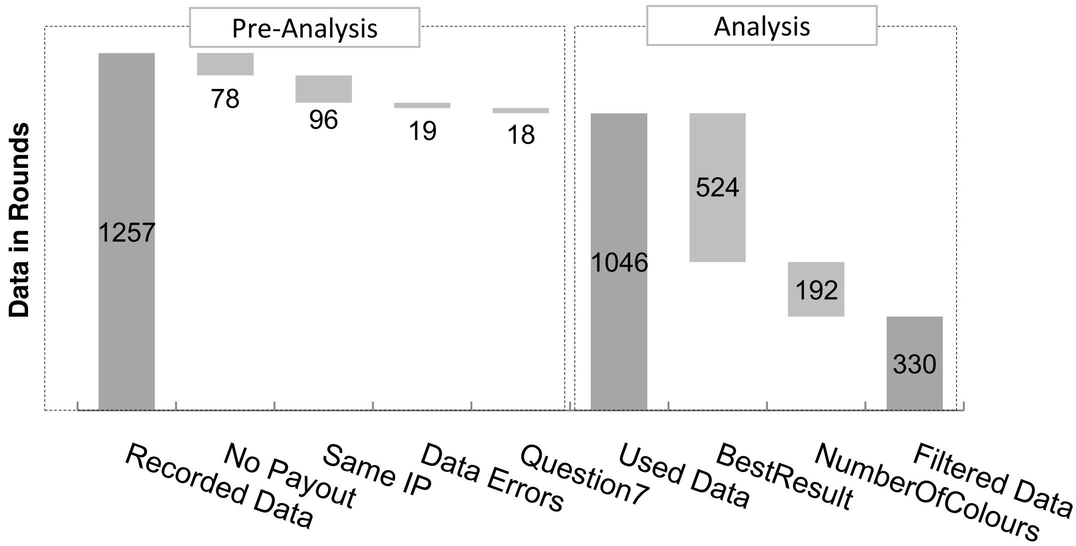


Figure 3.7.: Generated Data in Rounds

there are 1257 played rounds recorded (refer to Figure 3.7). The fact that the recorded

number of rounds outnumbers the potential 1.200 rounds, can be explained using two cases: first by the functionality test rounds conducted by the research team while the game was running and second by participants playing one or two rounds, then dropping out of the experiment. These cases can be identified by looking at the payout of the user, since there was no payout made to users who did not complete all rounds. Consequently, we will exclude the rounds of those users who did not achieve a payout greater than 0.
The next step is to identify the users who played the game again and again, and therefore have skewed the observation independence as described in Section 3.3.1. 32 users are found who share both their IP with at least one other user and who achieved a payout greater than 0. We do not exclude those users who committed only once to the experiment and therefore only reached a payout greater than 0 once.
The data of 6 participants showed errors, e.g. a total game time per round greater than 100s or a result greater than 100%. Since we were not able to trace the origin of these errors, we will exclude these users from the data set. As a result, 349 users with a total of 1.046 rounds build the data for the statistical analysis.
Within the data used for the analysis, there are two further filters applied. These filters are used to focus on the participants who made an effort to succeed in the game. As stated in Section 3.5, a performance lower than 80% per round can be explained by a lack of effort or understanding on the participant's side. By including the whole range of performances in the parameter estimation, the statistical noise might dilute the inductive value of the study. Consequently, we only concentrate on participants who were able to achieve a result greater than or equal to 80% for each round. This filter leaves a data set consisting of 174 users with a total of 522 rounds recorded. We will refer to this filter as **Filter 1**.
In a second step, users who are assigned to a treatment group with less than 7 colours are excluded. This is due to statistical implications of the data since significant results can be seen for the three remaining treatments. After applying both filters, the data set is made up of 110 users with 330 rounds recorded. We will refer to this data set as **Filter 2**.

Experiment stages

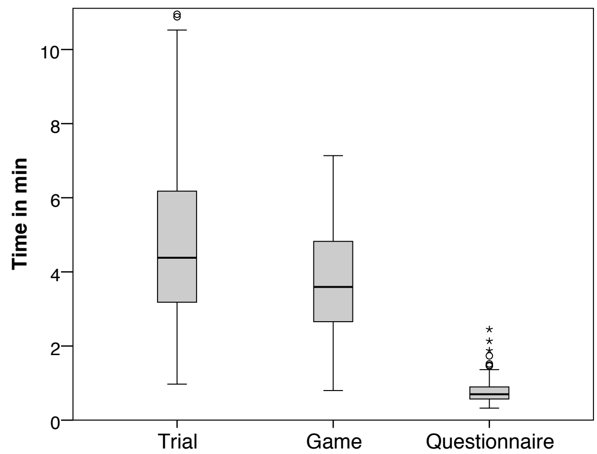


Figure 3.8.: Time spent on different stages of experiment[5]

The tracked time logs for the participants (Figure 3.8) show that individuals spent a considerable amount of time on the explanation and the Trial round. The total game length is defined by 5 minutes, and the majority of participants complete the game in less than the maximum time. A value greater than 5 minutes can be explained by the information box that pops up before the beginning of the game. The countdown for the current round only starts after participants click on the information box to start the round, resulting in a total game duration longer than 5 minutes when participants do not click instantly on the info box. The questionnaire is the shortest of the stages. We set a minimum time of 15 seconds, corresponding to 0.25 minutes, to fill out the questionnaire - the majority of individuals spent between 30 and 60 seconds on the questionnaire section.

Feedback from participants

53% of all users used the opportunity to give feedback via text. The majority gave positive feedback on the technical functionality of the game. These results might, however, be skewed since the feedback page is only reached by participants who did not have any major technical problems while going through the experiment.
Only a small number of participants gave feedback on areas of improvement, including the suggestion to offer a better and earlier visual notification when the time is running out. One participant suggested to introduce a more colourful colour scale or background music.

3.5. Data Descriptives

This section aims to provide a descriptive overview of the unfiltered data. First, the treatment statistics are introduced, second the descriptive results from the game are analysed and third the answers from the participants are examined.

Treatments and payout

The participants were randomly and uniformly assigned to one treatment once they directed to the welcome page. The low share of treatment 1 is related to the fact that the treatment was added in Trial 2 (Table 3.2).

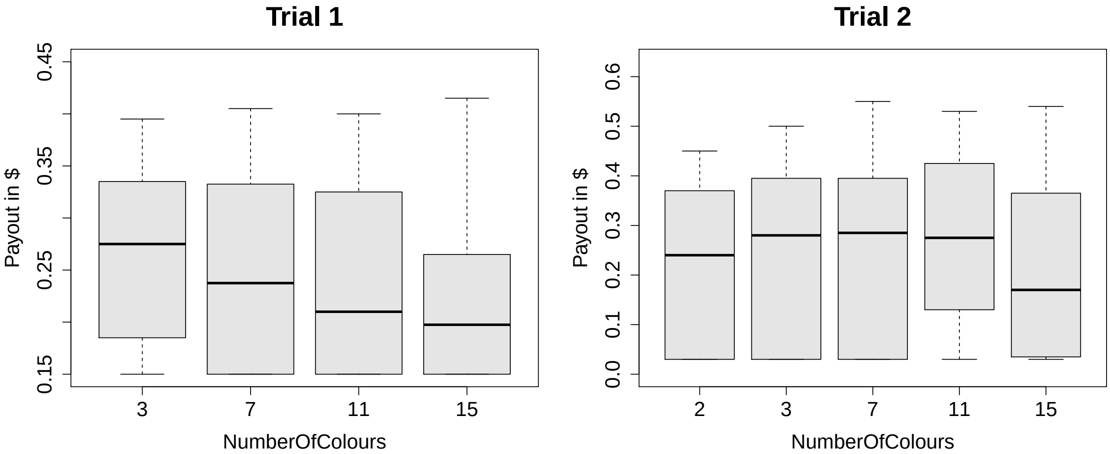
Table 3.2.: Distribution of Treatments

Treatment	Users	Share	Share of Dropouts[6]	Share Minimum Payout
1	33	10%	37%	31%
2	88	25%	40%	25%
3	78	22%	40%	31%
4	72	21%	39%	28%
5	77	22%	46%	27%
Total	348	100%	40%	28%

[5] Outliers with a time greater than 11 minutes on each stage are excluded for the purposes of this graph due to better readability.
[6] See Appendix B for the formulas.

The distribution of the payout presented in Figure 3.9 shows that the medians of the payout per treatment are decreasing with the *NumberOfColours* in Trial 1, whereas an inverted U-Shape can be detected for the medians in Trial 2, with a peak at the medium 7-colour level.
28% of all users achieved the minimum payout for each Trial, and 72% reached the bonus bar. The highest payout is $0.44 accomplished in Trial 1, and $0.55 in Trial 2.

Figure 3.9.: Distribution of payout among treatments for each Trial



Performance

The characteristics of the histograms for all result types show a similar shape[7]. The average result lies between 75% for the *FirstResult* and 79% for the *BestResult*. The relative standard deviation is high with values greater than 23%, and the standard deviation is just below 20. The population for all result types are negatively skewed with a value between –1.29 and –1.58, and the kurtosis is positive in the range of 1.4 to 2.4. The modal group can be found between 85% and 95%.
As indicated by the histograms, there is a lot of noise in the data, since values below 80% are likely to be related to a minimum effort. Therefore, the statistical potential is emphasized by excluding the low performers. The distribution of the different result types among different treatments is exemplary shown in Figure 3.10 (right) for the *FinalResult* and indicates a decreasing interquartile range with an increasing number of colours. The medians for all treatments are above the bonus bar and show a tendency to form a U-Shape with a peak at the treatment with three colours.

Significant outliers can be found in all treatments except for the treatment with two colours. Outliers with a *FinalResult* lower than 20% can be explained by a lack of effort and understanding on behalf of the participant, as opposed to an experience of information overload. A result greater than this value can namely be achieved by simply clicking on random boxes to fill up the knapsack.

[7]Figure 3.10 (left) shows the histogram for the *FinalResult*, see Appendix C for Histograms and Boxplots for *FirstResult* and *BestResult*.

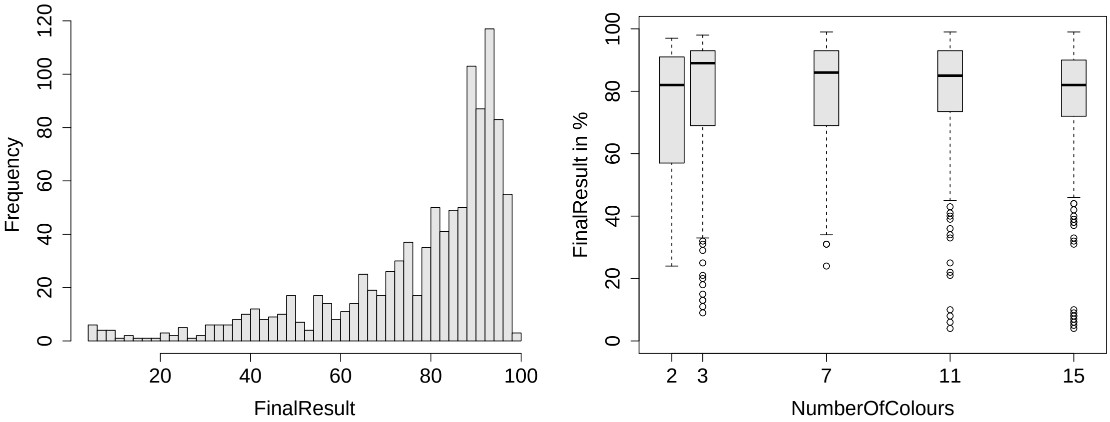


Figure 3.10.: FinalResult - Histogram and Box plot

Time

The histograms of the different time types differ. Whereas *FirstTime* shows a positively skewed distribution, *BestTime* seems to be closer to a uniform distribution. *FinalTime* has a negatively skewed population and *DecisionTime* looks like the results are normally distributed. The majority of participants reached the *FirstResult* in under 40 seconds, and finished the round mostly in over 80 seconds. Since only a minority played the full time, the use of the "Next"-button seemed to be popular. No specific time can be identified when participants were most likely to reach their *BestResult*.

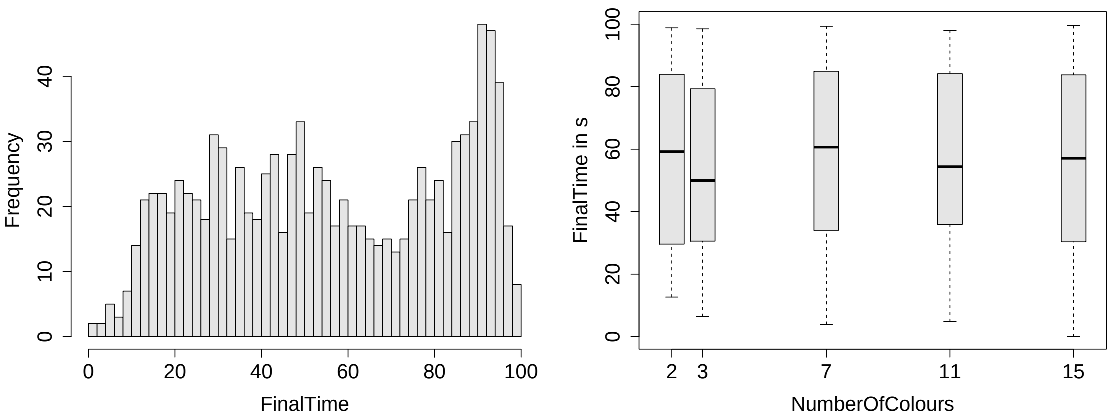


Figure 3.11.: FinalTime - Histogram and Box plot

The box plot of *FirstTime* shows similar medians across treatments. The interquartile range is smaller for treatments 2 and 4, and bigger for treatment 3. The standard deviation is the lowest for all time types.
The interquartile range is similar among all treatments for the *BestTime*. The values for the medians define a U-shape, with treatment 2 as the apex. The medians for *FinalTime* vary across treatments, yet the interquartile range does not seem to be dependent on the

number of colours.

Decision time and number

The box plots for *DecisionTime* show a wide range, from below 0 seconds to up to 6 seconds for all treatments. The medians are similar and treatment 3 has a wider interquartile range. The majority of participants clicked between 5 and 40 times on boxes per round. The box plot for *DecisionNumber* shows outliers for every treatment, resulting in a range of less than 10 clicks up until 130 clicks per round.

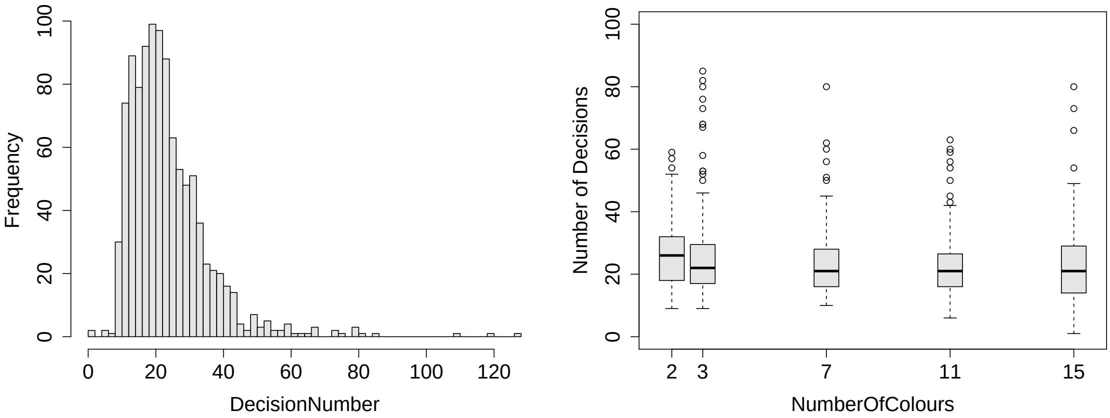


Figure 3.12.: DecisionNumber - Histogram and Box plot

Questionnaire

Figure 3.13 summarizes each treatment's medians for all the answers. In general, the medians for all treatments show similar values, so the different treatments do not seem to have an influence on the answers.

The mental demand for the task is described as high(5) for all *NumberOfColours* and the pace of the task is defined as medium to rather high. Most of the participants rate their performance as rather successful and define their effort for accomplishing their level of performance as high.

There are no significant signs for negative emotions like insecurity or irritation triggered by the experiment, since all treatments indicate a low level of negative emotions. Most individuals define "colour" (2) as the main box attribute they are looking at to reach their end result. A high number of individuals take into consideration a combination of the box size and box colour(3) to reach their end result. The box size (1) is favoured by only a small minority of participants, and other strategies(4) or no strategy(5) have only a minor influence on individuals.

Correlation of variables

Figure 3.14 illustrates the correlation between all the variables. The bigger the size of the circles, the greater the magnitude of the correlation. The green colour represents a positive

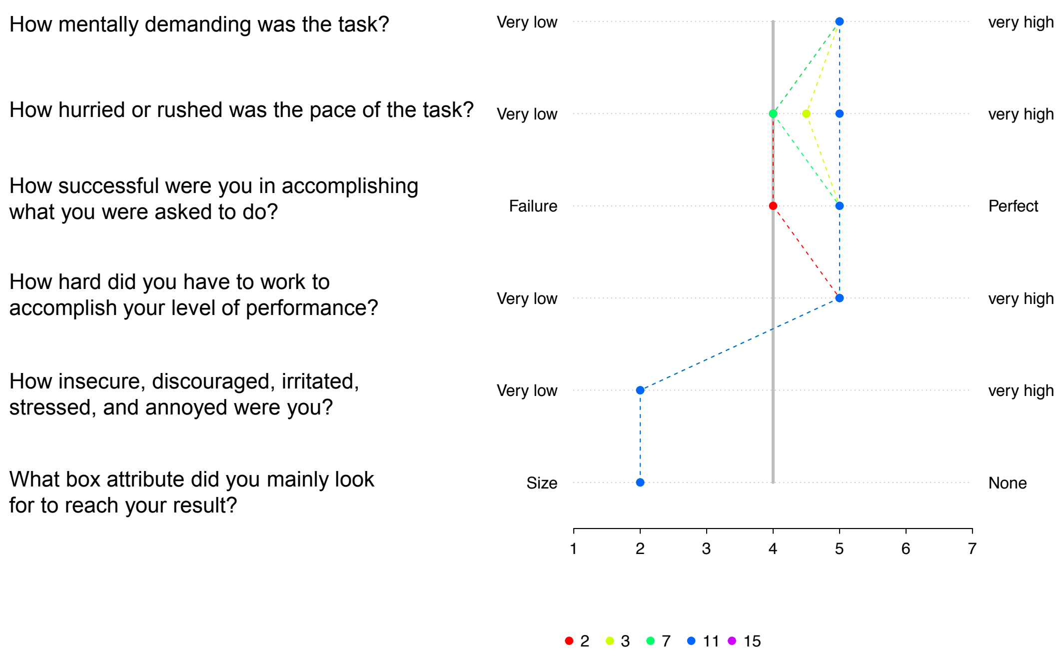


Figure 3.13.: Profile Plot

correlation and the red colour represents a negative linear correlation, so big green circles refer to a high positive linear correlation.
The two main correlation groups can be identified as the different result and time types since they both show a high correlation. The correlation between these two groups is smaller than within the groups, but is still positive. This result indicates that the more time a participant takes for completing the round, the more successful he or she will be, and vice versa.
The linear correlation between the time and result variables, and the independent variables *Round* and *NumberOfColours*, is only small, yet the *DecisionTime* shows a higher positive correlation for the *NumberOfColours* and a more negative correlation for *Round*. In other words, the more colours that are added to the game, the more time it takes the participant, on average, to make a decision, and the decision time decreases over the rounds played.
The *DecisionNumber* is positively correlated to the result and time variables, showing the highest correlation value for *FinalTime* and *BestTime*.

The correlation between the mental demand of the task (Question 1) and the level of how hard participants work to accomplish the task (Question 4) is the highest for all answers' correlations, even though it indicates only a medium correlation. Smaller but still positive correlations with the mental demand, can be identified for the pace of the task (Question 2) and the level of negative emotions (Question 5).
A small negative correlation is detectable for Question 1, Question 4 and Question 5, as well as the performance variables. Therefore, participants with a higher result indicate a lower mental demand, a lower level of effort and a lower level of negative emotions. Interestingly, the correlation matrix does not show a high correlation between how successful the participants are and how successful they feel (Question 2).

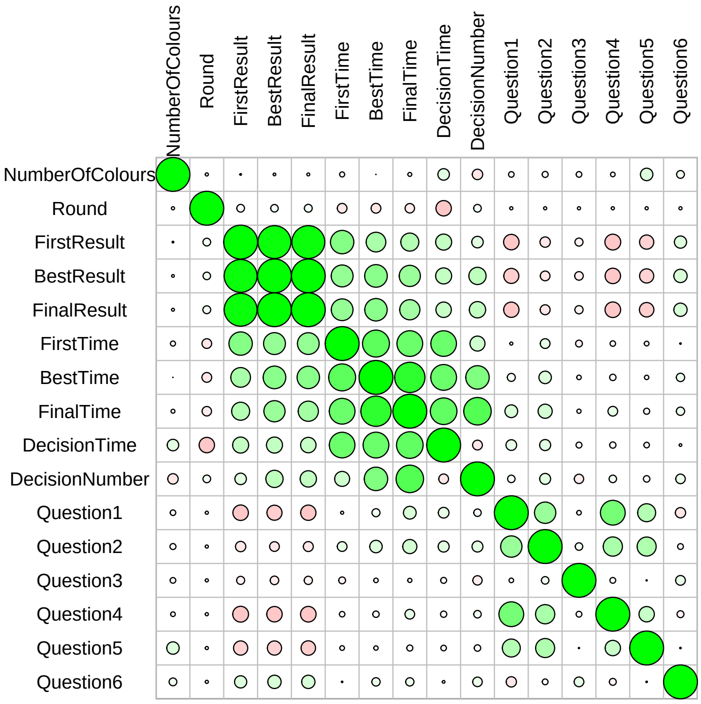


Figure 3.14.: Correlation of variables

4. Evaluation

The results for the knapsack experiment are carefully evaluated in this section. In the first step we discuss the choice of statistical models, second we introduce Non-Parametric Statistical Tests (NP) and the Linear Mixed Model (LMM) that are used to examine the recorded data, and in the last step we present the results for the parameter estimations.

4.1. Choice of statistical model

In order to evaluate the data and extract statistical stress-able results, the appropriate choice of a statistical model is crucial. Siegel (1957) defined three main criteria to identify the most suitable statistical model:

1. The statistical model of the test should fit the conditions of the research.
2. The measurement requirement of the test should be met by the measures used in the research.
3. From among those tests with appropriate statistical models and appropriate measurement requirements, that test should be chosen which has greatest power-efficiency[1].

The conditions of the research are defined by a combination of two between-subjects and one within-subjects factors. *Usergroup* and *Trial* separate the treatments and the within-subjects factor *Round* defines the number of observations recorded for each subject. Table 4.1 gives an overview of the examined variables. All variables fulfil the measurement requirements since all dependent variables are numerical and on a scale, and all independent variables are numerical.

Two statistical models are chosen to examine the data.

The application of Non-Parametric Statistical Tests (NP) is aimed at answering whether

[1]The power of a test is being defined as the probability that the test will reject the null hypothesis when in fact it is false and should be rejected (Siegel, 1957). Thus, a statistical test is considered to be good if it has small probability of rejecting H_0 when H_0 is true, but a large probability of rejecting H_0 when it is false.

Independent	Dependent		
NumberOfColours	*FirstResult*	*FirstTime*	*DecisionTime*
Round	*BestResult*	*BestTime*	*DecisionNumber*
Trial	*FinalResult*	*FinalTime*	

Table 4.1.: Overview - Independent and dependent variables

or not there is an influence from the independent variables *NumberOfColours*, *Round* and *Trial* on the dependent variables. In particular, the different filters are tested to identify the data sets that show an influence on *NumberOfColours*, *Round* and *Trial*.
Subsequently, a Linear Mixed Model (LMM) is used to give parameter estimations to the combination of those dependent and independent variables that show an influence in the NPs.

4.2. Non-Parametric Statistical Tests (NP)

Non-Parametric Statistical Testss are statistical models that do not specify restrictive conditions (Siegel, 1957). As a result, NP only make minimal assumptions regarding the underlying distribution of the data, so they do not require the data to be normally distributed. Furthermore, the tests rank the values instead of looking at the values themselves which makes them robust against outliers.
Two different categories of tests are used; an independent-samples test, analysing the dependent variables that are grouped by *NumberOfColours* and *Trial*; and a test for related samples, comparing *Rounds* for the same set of users.

4.2.1. Tests of Independent samples: *NumberOfColours* & *Trial*

Two tests are conducted for the independent samples. The Kruskal-Wallis Test of Independent Samples[2], comparing the distributions per *NumberOfColours*, as well as the Mann-Whitney Test, which compares the two *Trials* against each other.
Null and alternative hypotheses are defined in the following way:

Kruskal-Wallis- & Mann-Whitney-Test
H_0: *The distribution of the dependent variables are the same across different usergroups.*
H_A: *At least one distribution is different.*

Trial

The Mann-Whitney-Test retains the null hypothesis for all dependent variables for the influence of *Trial* (Table 4.2), so the different bonus systems do not seem to have an influence on the data.

NumberOfColours

The results for the Kruskal-Wallis-Test are dependent on the applied filters (Table 4.2). Except of *DecisionTime*, there is no influence of the *NumberOfColours* detectable for

[2]For the SPSS source code of the two tests, refer to this Link.

the unfiltered data, yet there is an undefined, but detectable influence for Filter 1. The Kruskal-Wallis-Test for Filter 2 shows an influence for the result variables. *FirstTime* does not seem to be affected by the *NumberOfColours*.

Table 4.2.: Results for the NP Independent & Dependent Tests

Reject hypotheses	*Trial* Any Filter	*NumberOfColours* Unfiltered	Filter 1	Filter 2	*Round* Any Filter
FirstResult			$\surd$	$\surd$	$\surd$
BestResult			$\surd$	$\surd$	$\surd$
FinalResult			$\surd$	$\surd$	$\surd$
FirstTime					$\surd$
BestTime			$\surd$		$\surd$
FinalTime			$\surd$		$\surd$
DecisionTime		$\surd$	$\surd$		$\surd$
DecisionNumber			$\surd$		$\surd$

4.2.2. Tests of dependent samples: ***Round***

A related-Samples Friedman's Two-Way Analysis of Variance by Ranks test (Friedman, 1937) is applied to compare the values in-between *Rounds*. The null and alternative hypotheses are defined in the following way:

Friedman's Test
$\mathbf{H_0}$: *The distributions of the dependent variables are the same across rounds.*
$\mathbf{H_A}$: *At least one distribution is different.*

Friedman's Test results are shown in Table 4.2 and reject the null hypothesis for all dependent variables. The applied filter does not have an influence on the findings.

4.2.3. Implications for parameter estimations

The findings for the NPs underline the magnitude of the statistical noise in the unfiltered data. According to the Kruskal-Wallis-Test, no influence is detectable by *NumberOfColours* when no filter is used. As a result, we only use Filter 1 and Filter 2 for the parameter estimation.
Furthermore, not every dependent variable seems to be influenced by the independent variables for the filtered data, so the parameter estimations for those variables might be non-significant.

4.3. Linear Mixed Model (LMM)

For the purposes of testing the proposed polynomial relationship between the dependent variables and the *NumberOfColours* as well as the logarithmic influence of *Round*, we examine the data using a Linear Mixed Model. This model is a powerful modelling tool that allows the analysis of complex datasets with hierarchical structures (Galecki, 2013). As a result, the Linear Mixed Model is an appropriate choice with a high power-efficiency for the conditions of our research.

The term *Mixed Model* refers to the use of both fixed and random effects in the same analysis (Seltman, 2012). Whereas fixed effects are essential to evaluate the potential polynomial relationship between the independent and dependent variables, random effects aggregate the influences which are not relevant for the purposes of this study.
Fixed effects are usually related to treatments (*NumberOfColours* and *Trial*), whereas subject effects are defined as random effects (*User*).
Subject effects include the individual characteristics of each participant. These effects (e.g. experience with similar tasks or the general ability to perform well in the given task), have an influence on the result and time variables, but are not the focus of our research. As described in Section 3.3, one implication for cognitive load experiments on Amazon Mechanical Turk (AMT) is the lack of control over what participants are doing during the experiment. By taking into account these individual influences, and separating them from the focus of the research, the noise of the data can be reduced and can return more significant results.

Normality Assumption

LMM imply the assumption that the data is normally distributed. Yet, the normality assumption is not met by the original data, so a data transformation is necessary.
We use the Box-Cox transformation[3] (Sakia, 1992) that applies a shifted power transformation to adjust the standard deviation and the mean to the requested values for the normal distribution. Since the method is using a range of power transformations, the efficiency of normalizing and variance equalizing for both positively- and negatively-skewed variables can be improved (Osborne, 2010). In order to find the optimal input parameter λ for the transformation, we implement Osborne's SPSS algorithm.
We find suitable λ-values for the variables *FirstTime* and *DecisionTime*. Nevertheless, no adequate values can be found for all result types, namely *BestTime*, *FinalTime* and *DecisionNumber*. Figure 4.1 provides an explanation for the lack of appropriate λ-values in the *FirstResult*. No λ-values can be found that reduce the skewness and kurtosis of the transformed data to a limit at which one can assume a normal distribution. Similar results are returned for *BestResult*, *FinalResult*, *BestTime* and *FinalTime*. In addition, a test of Normality proves our previous findings that a Box-Cox-transformation is only suitable for *FirstTime* and *DecisionTime*. Since we have a data set smaller than 2000 elements, the Shapiro-Wilk test (Shapiro and Wilk, 1965) is used to test the null hypothesis which states that the observed population does not come from a normal distribution. The

[3]Refer to Appendix B for the formula.

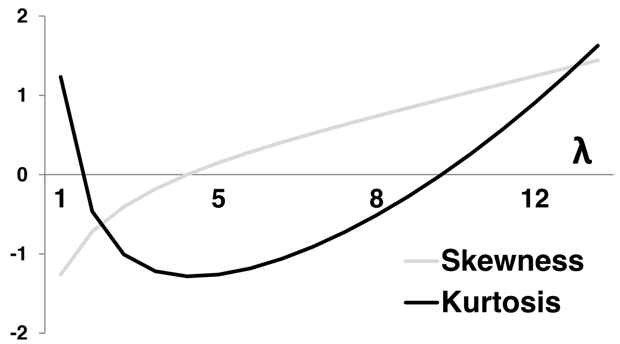


Figure 4.1.: Transformation *FirstResult* - Skewness and Kurtosis for different λ values

p-value is greater than 0.05 only for *FirstTime* and *DecisionTime*, so we must retain the null hypothesis for all other variables and conclude that only the data transformation of *FirstTime* and *DecisionTime* comes from a normal distribution.

Conclusion

The validity of the results based on the LMM depends on whether or not its conditions are met (Siegel, 1957). Since the normality assumption is not fulfilled for the majority of the variables, the addressed power of the LMM might be diluted.
According to Graybill (1976), we can either ignore the violation of the assumptions and proceed with the analysis as if all assumptions were satisfied, or we can use a distribution-free procedure.
A distribution-free procedure in the form of non-parametric tests is used in Section 4.2. Yet, these tests do not give information about potential parameter estimations for an implemented model. Therefore, we continue analysing the data using the LMM.

4.3.1. Choice of a LMM model

In order to specify a mixed model, several steps, each of which requires an informed decision (Seltman, 2012), are necessary.
First, the identification of fixed effects requires the specification of which influences affect the average performances for all individuals. We assume that the level of *NumberOfColours* and the *Round* number affects each participant. The Non-Parametric Statistical Tests in Section 4.2 do not show any significant influence on the population mean from the *Trial*, so we do not take this variable into account as a fixed effect.
Second, we must determine whether the fixed effects are sufficient without a corresponding random effect. We assume that the performance is dependent on the individual characteristics of the participant. In other words, participants have a relatively equal sensitivity to *NumberOfColours* and *Round*, but perform on different levels due to the individual characteristics of each participant. Consequently, every participant has his or her own regression line, with a personalized intercept and an equal slope (Figure 4.2). Due to the lack of influence by the *Trial*, a nested classification of the random intercept is not necessary that would require a combination of *Trial* and *User* for the random intercept.
Third, correlations among repeated measurements must be taken into account. We assume

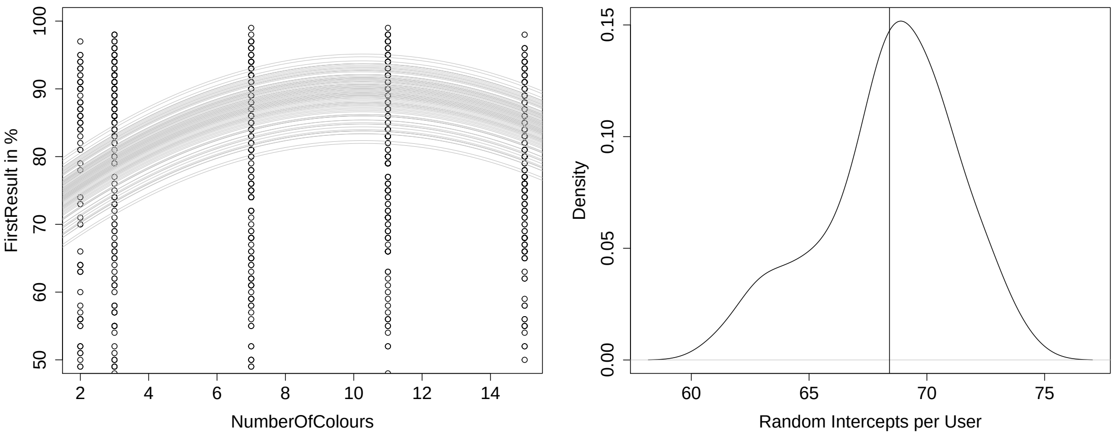


Figure 4.2.: User-individual Intercepts for the *FirstResult*-parameters

that all rounds a user plays are equally correlated with each other and the total variation per round, $\sigma^2 = \sigma_\gamma^2 + \sigma_\varepsilon^2$, can be partitioned into the "shared" (user) component, σ_γ^2 and the "unshared" (round) component, σ_ε^2.
The *Compound symmetry covariance matrix* is defined as:

$$\begin{aligned} \mathrm{cov}(Y_{ij}, Y_{ik}) &= \sigma_\gamma^2 + \sigma_\varepsilon^2 \cdot \mathcal{I}(k = j) \\ Y_{ij} &:= Total\ Variance\ of\ User\ i\ in\ Round\ j \end{aligned} \tag{4.1}$$

Equation 4.2 shows the used single-level LMM with random intercepts.

$$\begin{aligned} DependentVariable_{i,j} = & \\ & \boldsymbol{const} \\ & + \boldsymbol{\beta_1} * NumberOfColours + \boldsymbol{\beta_2} * (NumberOfColours)^2 \\ & \quad + \boldsymbol{\beta_3} * log_{10}(Round) + \boldsymbol{r_i} + \epsilon_{i,j} \end{aligned} \tag{4.2}$$

Indexes	$i := Participant,\ j := Round$
Fixed effect 1	$\boldsymbol{\beta_1}$ * NumberOfColours + $\boldsymbol{\beta_2}$ * $(NumberOfColours)^2$
Fixed effect 2	$\boldsymbol{\beta_3} * log_{10}(Round)$
Random effect	$\boldsymbol{r_i}$:= Random intercept per User i
Error	$\epsilon_{i,j}$

Implementation in R

For the implementation of the LMMs we use the **nlme**-package of the software programming language **R**[4] . The package is especially designed for LMMs (Pinheiro et al., 2013) and fits a linear mixed-effects model in the formulation described by Laird and Ware (1982).
The model described in 4.2 is implemented using the following code:

[4]Refer to R Core Team (2012).

```
> LMM <- lme(DV ~ 1 + NumberOfColours + NumberOfColours2 + RoundL,
+               random=~1 | User,
+               correlation=corCompSymm(form = ~ RoundL| User))
```

4.3.2. Results for the LMM

According to the Non-Parametric Statistical Tests findings, only the filtered data is considered for the parameter estimations.
The results for the fixed effects of the LMM can be interpreted in the same way as an ANOVA regression. Nevertheless, we must take into account that the intercept represents the mean over all subjects and each individual subject has its own individual intercept (Seltman, 2012). The population made up by all individual intercepts approximately follows a normal distribution, with the computed overall estimate as the mean and the variance of the random intercepts (Figure 4.2).

Filter 1: Users with *BestResult* $\geq$ 80% for each round

NumberOfColours
The LMM returns significant results for the *FirstResult*, the *DecisionTime* and the *DecisionNumber*. Moreover, they all have significant parameter estimations for all parameters. So the hypothesis that there is an ∩-relation between the performance of a participant and the *NumberOfColours* is only supported for the *FirstResult*. In addition, the results for *DecisionTime* and *DecisionNumber* show a polynomial relation with *NumberOfColours*, as expected by the hypothesis. The *DecisionTime* reaches its maximum and the *DecisionNumber* its minimum for a medium level of information granularity. No significant results are signalized for the *First-*, *Best-*, and *FinalTime*.

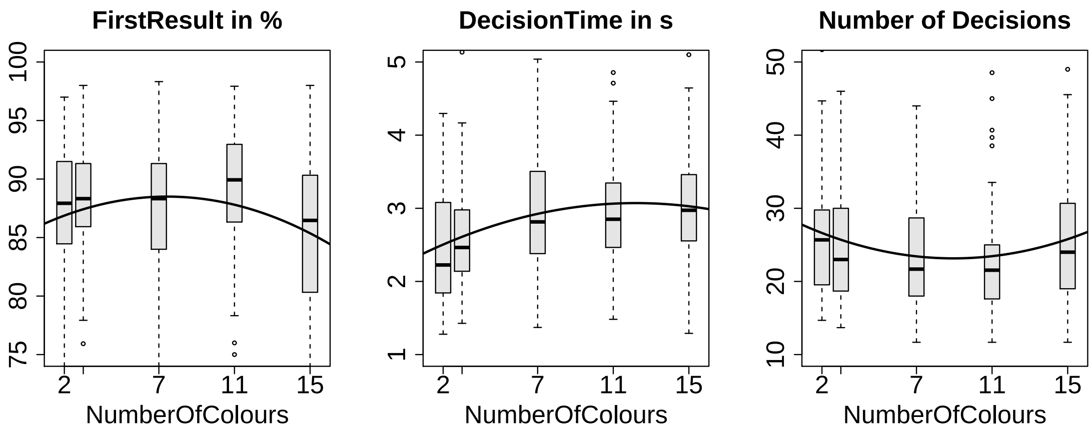


Figure 4.3.: Results for Filter 1: Information Granularity[5]

Round
The learning effect, hence the influence of the *Round*, is significant for all variables. So we can conclude that there is a learning effect when playing several rounds in the experiment, and the slope of this learning effect diminishes with the number of rounds played, as

[5]The values for the dependent variables are adjusted so they exclude the predicted effect by *Round*.

	First Result	Best Result	Final Result	First Time	Best Time	Final Time	Decision Number	Decision Time
(Intercept)	85.46^{***}	89.82^{***}	89.40^{***}	44.09^{***}	53.50^{***}	63.38^{***}	2.26^{***}	28.91^{***}
	(1.29)	(0.92)	(0.97)	(4.30)	(4.44)	(5.23)	(0.18)	(2.14)
NumberOfColours	0.82^{*}	0.41	0.43	1.83	0.84	0.31	0.13^{**}	-1.29^{*}
	(0.36)	(0.26)	(0.27)	(1.22)	(1.23)	(1.49)	(0.05)	(0.61)
NumberOfColours2	-0.05^{**}	$-0.03^{\cdot}$	$-0.03^{\cdot}$	-0.10	-0.02	0.03	$-0.01^{\cdot}$	0.07^{*}
	(0.02)	(0.01)	(0.02)	(0.07)	(0.07)	(0.09)	(0.00)	(0.03)
RoundL	$2.24^{\cdot}$	2.84^{***}	2.99^{***}	-17.39^{***}	-16.16^{***}	-14.54^{***}	-1.21^{***}	4.86^{***}
	(1.29)	(0.76)	(0.81)	(3.37)	(3.27)	(3.25)	(0.11)	(1.40)
σ: Intercept User in Trial	3.52	2.84	2.99	13.62	12.81	17.58	0.59	7.11
σ: Residual	5.50	3.19	3.40	14.09	14.72	13.45	0.47	5.82

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.3.: LMM-Results for Filter 1

32

indicated by the logarithmic relation. Surprisingly, the number of decisions increases with the number of rounds, as supposed to the expected decline.
Since the time it takes participants to reach the *FinalResult* is represented by the product of *DecisionNumber* and *DecisionTime*, two opposing trends can be identified when describing the round effect on the *FinalTime*. Whereas the number of decisions increases, the time to take a decision decreases. The LMM returns a negative parameter for *FinalTime*, indicating that the decrease in the decision times outbalances the increase in the number of decisions.

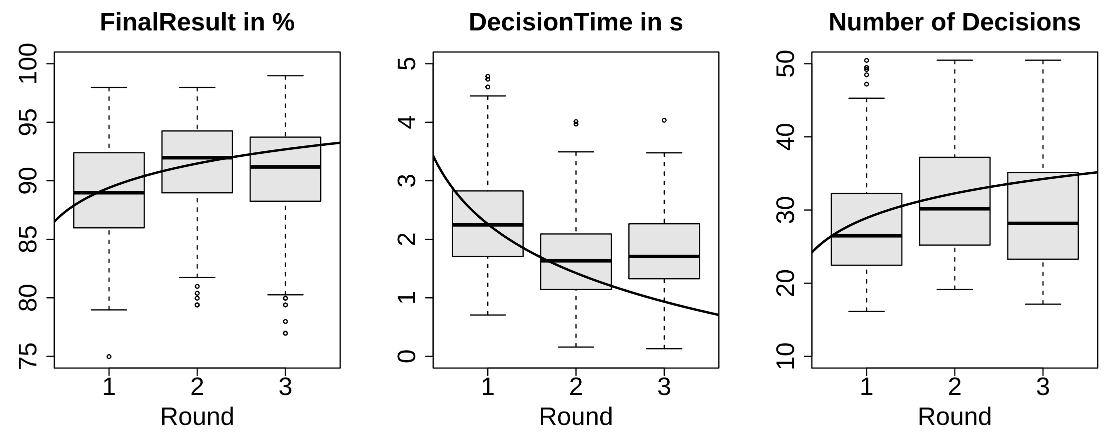


Figure 4.4.: Results for Filter 1: Learning effect[6]

User

The standard deviations of the *User* random intercepts are high, and underline the influential magnitude of the individual performance characteristics of the participants. In particular, the *DecisionNumber* and *DecisionTime* seem to be most affected by the individual, since the relative standard deviation is the highest for all variables.

Filter 2: Filter 1 + *NumberOfColours* $\geq$ 7

NumberOfColours

When only including the treatment groups that were assigned to 7, 11 and 15 colours, the parameter estimations become significant for *FirstResult*, *BestResult* and *FinalResult*. The $\cap$-relation with *NumberOfColours* is consequently confirmed when looking at this range of *NumberOfColours*. These findings, however, do not support our hypothesis that the performance will be best on a medium information granularity, since a peak can be found for the 11-colours treatment. In addition, none of the previous findings for *DecisionNumber* and *DecisionTime* are confirmed in Filter 2.

Round

A learning effect can also be detected for this data set, with an exception of *FirstResult* - the LMM does not return significant results for its parameters.

[6]The values for the dependent variables are adjusted so they exclude the predicted effect by *NumberOfColours*.

	First Result	Best Result	Final Result	Decision Time	Decision Number
(Intercept)	68.41***	78.95***	76.55***	3.34***	36.34***
	(6.87)	(5.13)	(5.34)	(0.94)	(10.85)
NumberOfColours	4.08**	2.47*	2.88**	−0.07	−2.78
	(1.36)	(1.01)	(1.05)	(0.19)	(2.14)
NumberOfColours2	−0.20**	−0.12*	−0.14**	0.00	0.14
	(0.06)	(0.05)	(0.05)	(0.01)	(0.10)
RoundL	2.26	3.73***	3.64***	−1.31***	6.26***
	(1.43)	(0.96)	(1.03)	(0.15)	(1.63)
AIC	2134.39	1897.16	1936.59	719.54	2288.81
BIC	2164.69	1927.46	1966.89	749.84	2319.11
Log Likelihood	−1059.20	−940.58	−960.30	−351.77	−1136.41
σ: Intercept User	4.16	3.22	3.31	0.61	7.10
σ: Residual	4.77	3.17	3.43	0.49	5.33

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.4.: LMM-Results for Filter 2

User

The standard deviation for the *User* random intercepts again show high values yet again, so the impact of excluding the variance caused by individual characteristics is underlined.

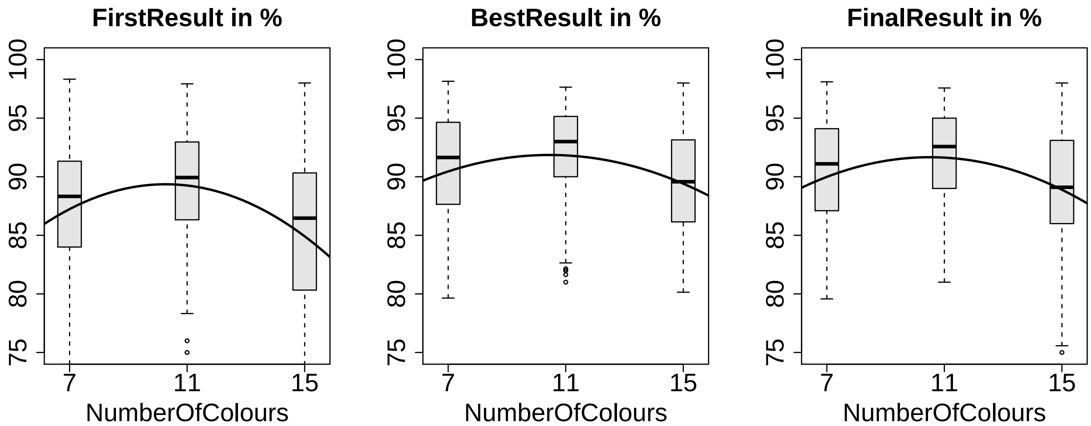


Figure 4.5.: Results for Filter 2: Information Granularity[7]

[7]The values for the dependent variables are adjusted so they exclude the predicted effect by *Round*.

5. Conclusion

In this section, the conclusions from the evaluation section are drawn and potential limitations of the result are discussed. Furthermore, lessons learned from the experiment are described and potential fields of further research are examined.

Amazon Mechanical Turk (AMT) and the quality of data

As indicated by the findings from the Non-Parametric Statistical Tests, the data generated on AMT includes a lot of noise. Many participants failed to get a result higher than 80%, a level which can be considered fairly easy to achieve. Even increasing the leverage of the bonus and the total possible payout did not show to have an influence on the participants. So we must conclude that offering the experiment as an HIT on AMT will return a lot of data that may not be fully usable for a statistical analysis.

The *FirstResult* is the best on a medium level of information granularity, *DecisionTime* shows a maximum and *DecisionNumber* a minimum on the medium level

When focusing the statistical analysis on the participants who made an effort to succeed in the game, the results indicate that an information overload has materialized. In fact, the highest performance is achieved on a medium level of information granularity for the first result that the participants offer. These findings are particularly interesting because we can argue that users of smart meters have an interest in achieving a satisfying result in a minimum amount of time, since they divide their time between many things.
The findings for the *DecisionTime* indicate that users facing a medium level of information granularity take the most time to make a decision. These results show a potential information overload and can be explained by Jacoby's findings.
Individuals facing an information overload might not directly experience an information overload while making their choice. This could be because they only concentrate on one part of the given information prior to their decision. In other words, individuals confronted with an information overload are selective in their information selection and stop *"far short*

of overloading themselves" (Jacoby, 1984).
A participant on a medium level of information granularity is just at the border of an information overload - the individual can still cope with the amount of information. Since the given information provides more detail than a lower level of information granularity, we can define three consequences:

1. The choice is supported by a better understanding of the underlying parameter (Benefit) and therefore the choice accuracy increases ⇒ *FirstResult increases.*
2. Due to the amount of information to process, it takes the individual more time to make the decision ⇒ *DecisionTime increases.*
3. Since the choice accuracy increases, the individual makes fewer but better decisions ⇒ *DecisionNumber decreases.*

When an individual is provided with more information than the optimal level, the information processing gets selective, so the actual amount of information that is considered for the choice might decrease. Therefore, both the choice accuracy and the decision time decrease, and the individuals might not base their decisions on a profound assessment of the options, but rather play around with different choices. Consequently, the number of decisions increases.
In addition, when excluding the lower levels of information granularity, the results of the statistical model indicate that the best performance can be achieved on a medium-to-high level of information granularity. These findings are not described in our original hypothesis, but have implications for the further development of the experiment.
A future experiment might include setups with more colours, in order to examine if the best performances further decrease with an increasing number of colours.

Learning Effect has an influence on both the performance and the game time

Participants improve their performance with a growing number of played rounds, yet the magnitude of the learning effect diminishes over the number of rounds. So for the identified information overload in reaching the *FirstResult*, we can conclude that individuals find strategies to cope with information overload when they get more experienced in this situation. A potential question to ask in future research is how these findings are reflected in a setup that includes more rounds.

Individual characteristics affect the performance

The individual characteristics of the participants have proven to influence their performances. Even though participants might show a similar sensitivity to the level of information granularity, their performances are affected by individual factors. Taking into account these effects when setting up a statistical model helps us to exclude influences such as limited attention while playing the game, or an individual talent for succeeding in these types of experiments. As a result, limitations of the Amazon Mechanical Turk and their negative consequences on statistical results can be tackled by the design of the statistical model.

A higher leverage on the bonus does not improve performances.

The adjusted bonus system in Trial 2 has not proven to have an influence on the performances of the participants. Reasons might be related to an unclear explanation so individuals were not aware of the monetary incentives when reaching the bonus bar of 80%.

Limitations of the experiment

In order to make the experiment as intuitive and straight-forward as possible, a considerable level of abstraction is necessary. This helps participants to better understand the given task, but also simplifies the information environment that the real-life smart-meter user is usually confronted with[1]. The experiment has proven that an information overload exists in the simplified game environment. So we can argue that these results lead to the conclusion that an information overload would occur under a more complex experiment setup, e.g. analysing the data provided by a smart-meter. In contrast to that, the undetectable information overload in the simplified experiment for the best and final performances of a participants does not point to the conclusion that there will not be an information overload in a more complex experiment environment.

Research Question!!!!

[1]Refer to Jacoby (1984)

6. Declaration

Ich versichere hiermit wahrheitsgemäß, die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die Satzung des Karlsruher Instituts für Technologie (KIT) zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, den DD. MM. 20XX

Christoph Großbaier

Appendix

A. Algorithm

Algorithm 3 Setup of rounds

```
/* rounds:= 4 rounds; 4 usergroups */
Input: rounds, usergroups
Result: 4 Rounds per each usergroup
benchmarkCurrent = 0
forall the rounds do
    /* 100 iterations */
    for i=0 to 100 do
        /* 100 Boxes */
        for j=0 to 100 do
            width = uniform random in range of 20 and 80
            benefit = uniform random in range of 0 and 80
            create Box with width and benefit
        end
        benchmark = (DPA Solution / GA Solution - 1) x 100%;
        if benchmarkCurrent ≤ benchmark then
            benchmarkCurrent = benchmark
            /* itemsHardProblem:= save current boxes  */
            boxesHardProblem = boxes
        end
    end
    forall the  boxes in boxesHardProblem do
        forall the usergroups do
            new Box (box.benefit, box.weight, current round)
            add colour according to benefit and number of colours
        end
    end
end
```

B. Formulas

$$Share\ of\ dropout = \frac{p_i(user|dropout)}{p_i(user)}, \quad i = 1, ..5,\ user := number\ of\ users\ logged\ in \tag{6.1}$$

$$Share\ of\ minimum\ payout = \frac{p_i(user|payout = 0)}{p_i(user)}, \quad i = 1, ..5,\ user := number\ of\ users \tag{6.2}$$

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda}, & if \lambda \neq 0, \\ log(y_i), & if \lambda = 0 \end{cases}, \quad y > 0 \tag{6.3}$$

C. Descriptive statistics

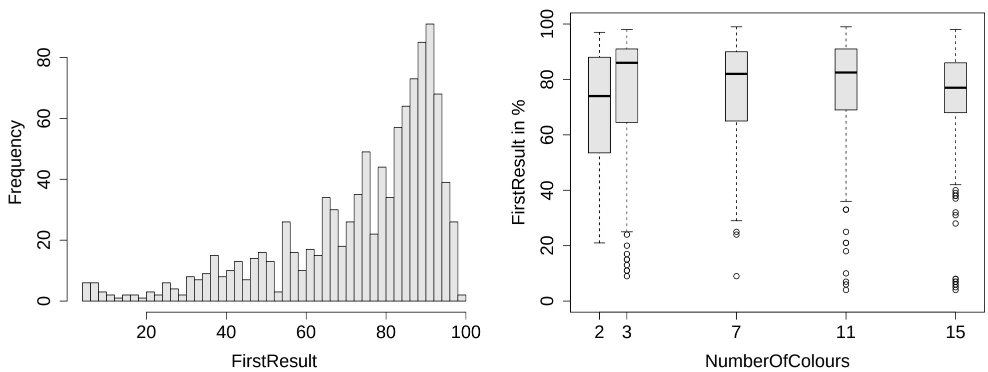


Figure C.1.: FirstResult - Histogram and Box plot

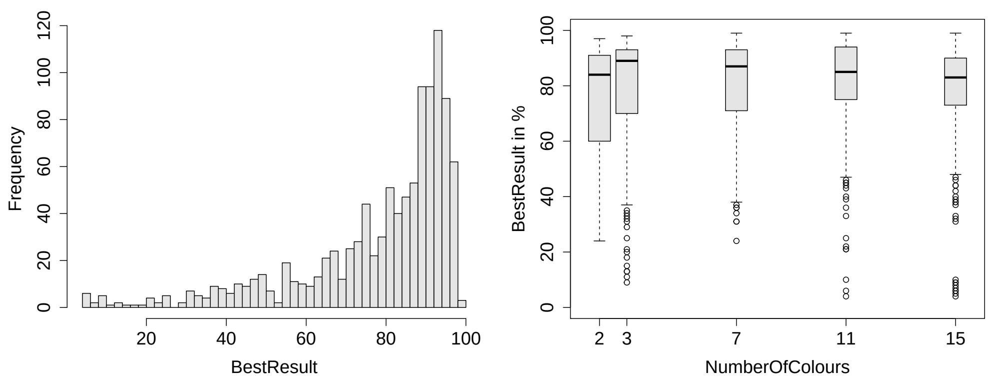


Figure C.2.: BestResult - Histogram and Box plot

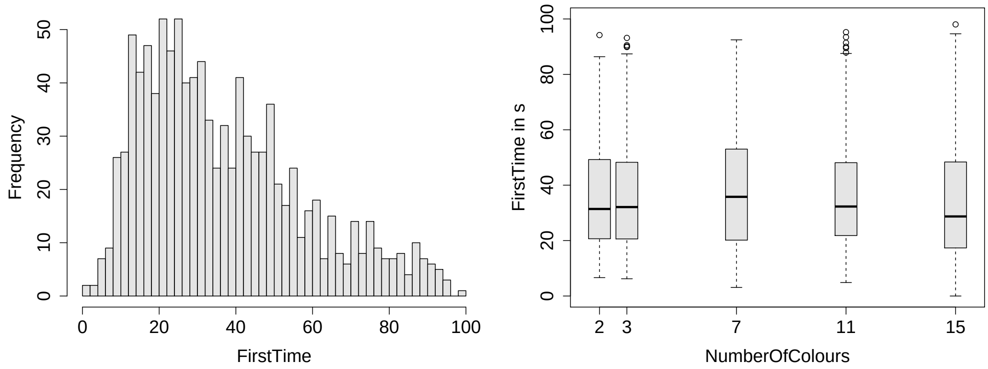


Figure C.3.: FirstTime - Histogram and Box plot

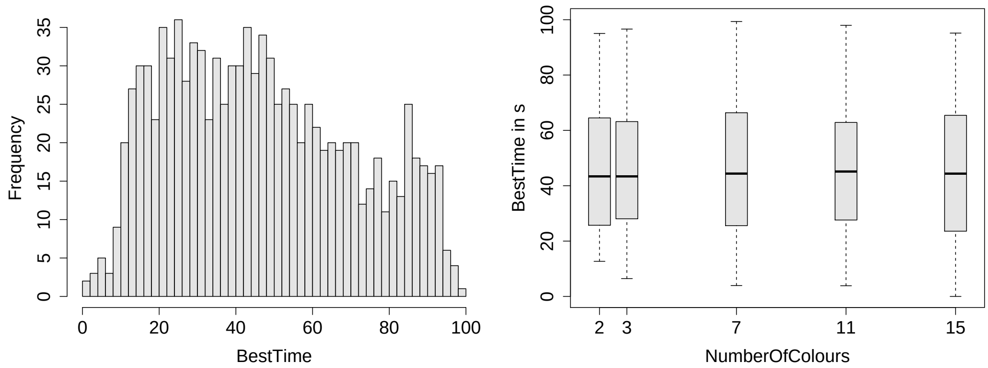


Figure C.4.: BestTime - Histogram and Box plot

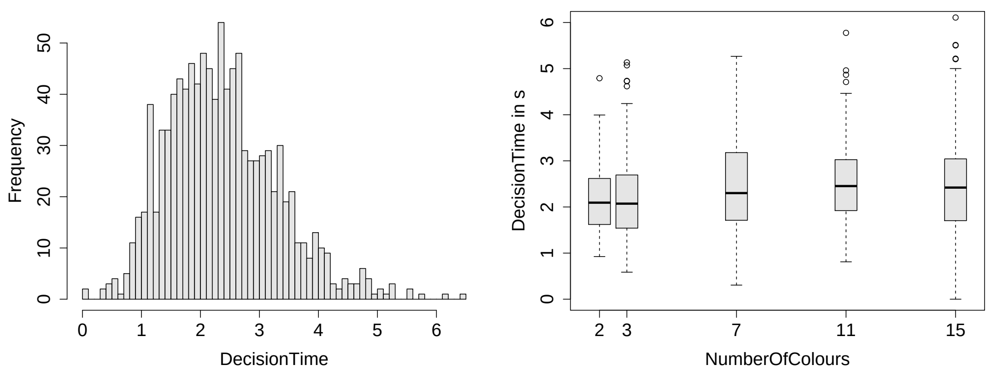
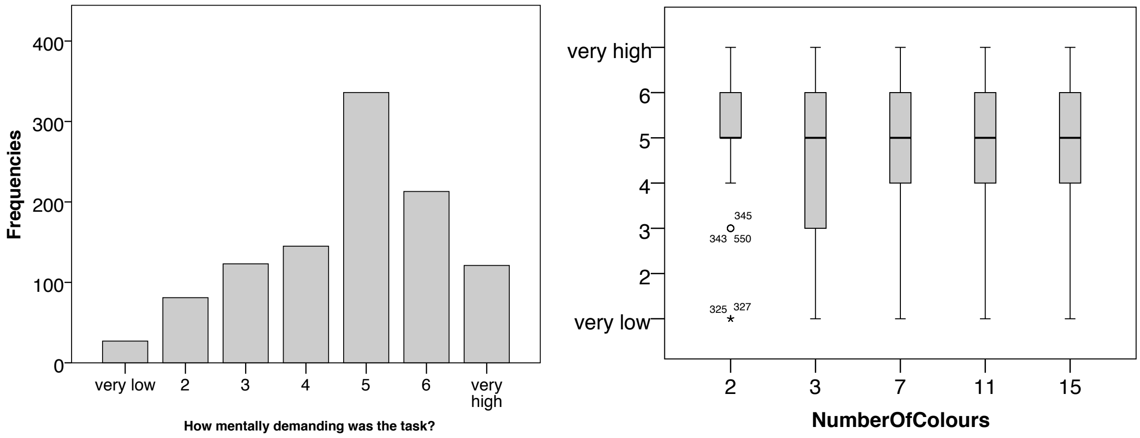


Figure C.5.: DecisionTime - Histogram and Box plot

Figure C.6.: Question 1 - very low (1) to very high (7)



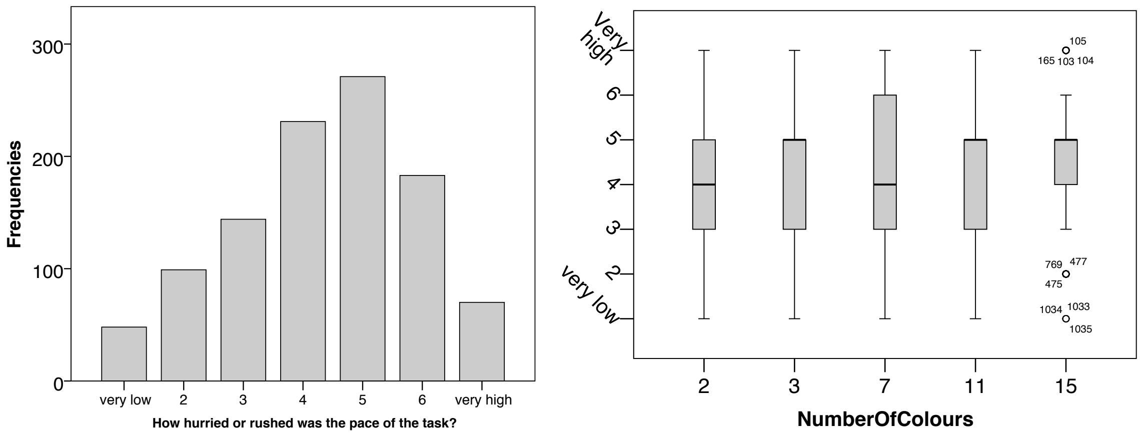


Figure C.7.: Question 2 - very low (1) to very high (7)

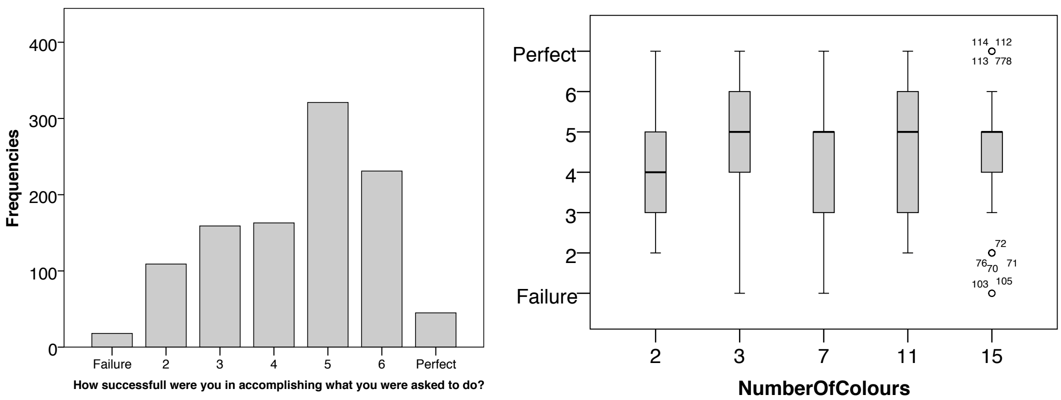


Figure C.8.: Question 3 - Failure (1) to Perfect (7)

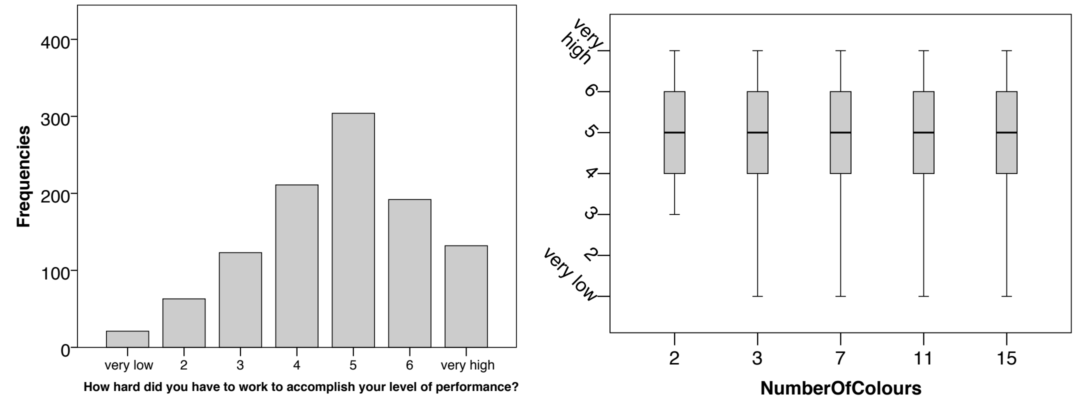


Figure C.9.: Question 4 - very low (1) to very high (7

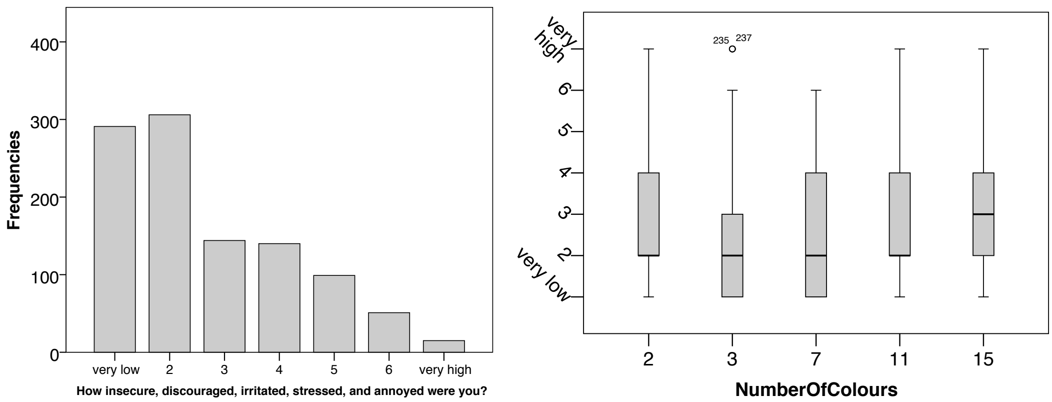


Figure C.10.: Question 5 - very low (1) to very high (7)

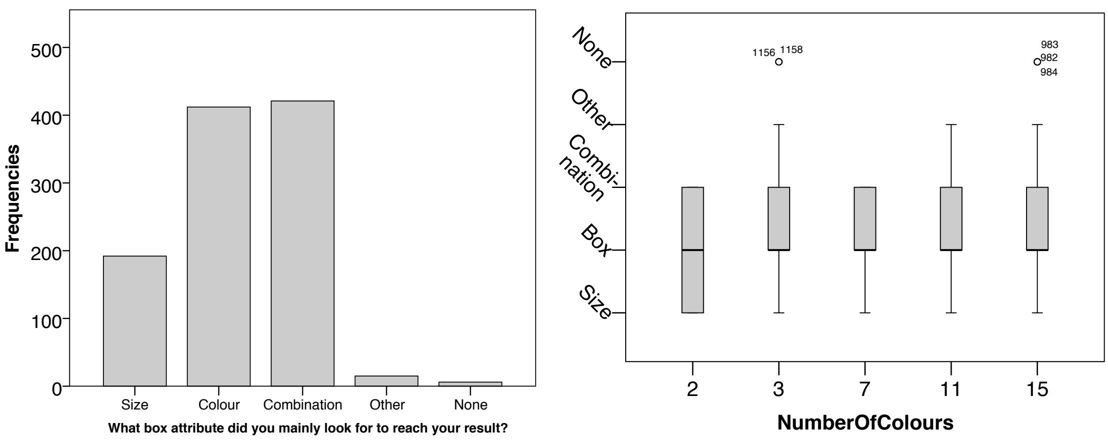


Figure C.11.: Question 6 - Size (1), Colour (2), Combination (3), Other (4), None (5)

Bibliography

Anderson, W. and V. White (2009). Exploring consumer references for home energy display functionality. *Centre for Sustainable Energy, Report to the Energy Saving Trust*.

Berinsky, A. J., G. A. Huber, and G. S. Lenz. (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*.

Buhrmester, M. D., T. Kwang, and S. D. Gosling (2011). Amazon's mechanical turk - a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science 6*, 3–5.

Chen, Y.-C., R.-A. Shang, and C.-Y. Kao (2009, January). The effects of information overload on consumers' subjective state towards buying decision in the internet shopping environment. *Electron. Commer. Rec. Appl. 8*(1), 48–58.

Darby, S. (2000). Making it obvious: designing feedback into energy consumption. *Proceedings of the 2nd International Conference on Energy Efficiency in Household Appliances and Lighting. Italian Association of Energy Economists / EC-SAVE programme.*.

Darby, S. (2006). The effectiveness of feedback on energy consumption. *A review for DEFRA of the literature on metering, billing and direct displays 1*.

Darby, S. (2008). Why, what, when, how, where and who? Developing UK policy on metering, billing and energy display devices. *Buildings*, 70–81.

Fischer, C. (2008). Feedback on household electricity consumption: a tool for saving energy? *Energy Efficiency 1*, 79–104.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association 32*(200), pp. 675–701.

Galecki, A. (2013). Linear mixed-effects models using r : A step-by-step approach.

Gardner, R. M., D. L. Brown, and R. Boice (2012). Using amazon's mechanical turk website to measure accuracy of body size estimation and body dissatisfaction. *Body Image 9 (4)*, 532–534.

Graybill, F. (1976). The theroy and applications of the linear model. *Review of Economics and Statistics 62*, 147–148.

Hahn, M., R. Lawson, and Y. G. Lee (2006). The effects of time pressure and information load on decision quality. *Psychology and Marketing 9*(5), 365–378.

Hendrick, C., J. Mills, and C. A. Kiesler (1968). Decision time as a function of the number and complexity of equally attractive alternatives. *Journal of Personality and Social Psychology 8(3, Pt.1)*, 313–318.

Henryson, J., T. Hakansson, and J. Pyrko (2000). Energy efficiency in buildings through information - swedish perspective. *Energy Policy 28*(3), 169 – 180.

Hwang, M. I. and J. W. Lin (1999). Information dimension, information overload and decision quality. *Journal of Information Science 25*(3), 213–218.

Jacoby, J. (1984). Perspectives on information overload. *Journal of Consumer Research 10*(4), pp. 432–435.

Jacoby, J., D. E. Speller, and C. A. Kohn (1974). Brand choice behavior as a function of information load. *Advances in Consumer Research Volume 01*, 381–383.

Keller, K. L. and R. Staelin (1987, September). Effects of quality and quantity of information on decision effectiveness. *Journal of Consumer Research 14*(2), 200–213.

Kempton, W. and L. L. Layne (1994). The consumer's energy analysis environment. *Energy Policy 22*(10), 857 – 866.

Kleinberg, J. and E. Tardos (2005). *Algorithm Design*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Laird, N. M. and J. H. Ware (1982). Random effects models for longitudinal data. *Biometrics 38*, 963–974.

Lurie, N. (2004). Decision making in information-rich environments: the role of information structure. *Advances in Consumer Research 29*, 91–92.

Malhotra, N. K., A. K. Jain, and S. W. Lagakos (1982). The information overload controversy: An alternative viewpoint. *Journal of Marketing 46*(2), pp. 27–37.

Osborne, J. W. (2010). Improving your data transformations: Applying the box-cox transformation. *Practical Assessment, Research & Evaluation 15 (12)*, 1–6.

Payne, J. W. (1982). Contingent decision behavior. *Psychological Bulletin 92.2*, 382–402.

Pinheiro, J., D. Bates, and S. DebRoy (2013, 01). Package nlme: Linear and nonlinear mixed effects models. Online.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology 299*, 172–179.

Sakia, R. M. (1992). The box-cox transformation technique: A review. *Journal of the Royal Statistical Society. Series D (The Statistician) 41*(2), pp. 169–178.

Schmidt, C. (2012). Making smart decisions - experimental analysis of smart meter information. *Department of Economics and Business Engineering - Institute of Information Systems and Management (IISM) - Information & Market Engineering*.

Seltman, H. J. (2012). *Experimental Design and Analysis.* College of Humanities and Social Sciences at Carnegie Mellon University.

Shapiro, S. S. and M. B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika 52*(3/4), pp. 591–611.

Siegel, S. (1957). Nonparametric statistics. *The American Statistician 11*(3), 13–19.

Streufert, S. and M. J. Driver (1965). Conceptual structure, information load and perceptual complexity. *Psychonomic Science 3(6)*, 249–250.