Identifying variables impacting external validity of conjoint analysis

a meta-analysis in FMCG branch

MSc. Thesis

Economics and Business -
Specialization: Data Science and Marketing Analytics

Erasmus University Rotterdam

Erasmus School of Economics

Student: Loïs Schipper

Student number: 443675

Date: 14/07/2021

Supervisor: Prof. Dr. D. Fok

Second assessor: Dr. R. Karpienko

External supervisor: R. Don

Abstract

Conjoint analysis often does not work as well as theory suggests. A lack of external validity means that there is a difference between preference shares and actual market shares. Improving external validity could help marketers in making more accurate predictions and consequently help to make better decisions. Throughout this study, the main objective is to find variables with a significant influence on external invalidity and find similarities between product categories. Difference is addressed as the magnitude of absolute difference, which holds the total absolute error of a study. It is also addressed as the difference on a product level, which is the difference of preference minus market share for a product. Using regression techniques on both differences, this study identifies the following significant variables: price, the number of respondents, and a high purchase frequency significantly decrease difference; whereas distribution and the number of products in the market increase difference. Seven product categories, such as cigarettes, batteries, and dairy products, are used as dummy variables and interaction effects with price, distribution, and volume to find similarities between studies. However, there are not enough significant results stating that product categories significantly impact price, distribution, and volume, indicating these variables are not context-dependent. Lastly, black-box models predict actual market shares of new data, using the conjoint prediction and all previously mentioned variables as input. The Random Forest predicts the data very well, but only slightly better compared to the baseline. It shows that there is still room for improvement in the estimation of market shares.

Keywords: conjoint analysis, external validity, stated preference, meta-analysis, FMCG

Acknowledgement

First, I would like to show my gratitude to Prof. Dr. Dennis Fok for his guidance through the process and his academic insights.

Secondly, I would like to thank Remco Don for the practical insights and the weekly meetings. Without either of these, this thesis would not have been the same.

Also, I want to thank the colleagues at SKIM for all help, (coffee) breaks, and conversations, which made writing this thesis so much better.

Finally, a massive thanks to my family for always being there to hear me out and putting up with me through the period of writing this thesis. In particular, I would like to take the chance to thank my parents for encouraging me to pursue my Masters and a final big thanks to my dad, Peter Schipper, for always being ready to act as a sparring partner when I needed one.

Contents

List of Figures

List of Tables

1 Introduction

In recent years, conjoint analysis has become one of the most-used methods for researching the market and the customers' preferences in that market. As the method makes it easy to show products and trade-offs of (important) attributes and their different levels to the respondent, customers get to make a more real-life decision while filling out a survey than other forms of surveys. The method is more closely related to real-life decision-making compared to surveys that do not show trade-offs, as this is a common feature in product selection in a real-life decision.

In a survey, respondents manually select the products in a simulated environment. They explicitly show which products, attributes, and levels have the highest preference by selecting these most often. The resulting preferences are called *stated preferences*. On the other hand, there are *revealed preferences* where customers have not specifically indicated what product they prefer; their (purchase) behaviour and real market data gives an idea of preference. A product that is sold very often is believed to have a higher level of preference compared to a similar alternative that is rarely bought (Fifer et al., 2014). Surveys, and the resulting stated preferences, are needed as revealed preferences might be subject to correlated variables. They cannot give insights in situations with slightly different conditions, and it is sometimes difficult to examine all variables when there is not enough variation in revealed preferences (Kroes and Sheldon, 1988).

Even though conjoint analysis is supposed to work relatively well according to the theory, it happens more often than not that the outcome of the conjoint differs from the actual market shows due to impacting variables (Feit et al., 2010).

An aspect that possibly interferes with the accuracy of the conjoint analysis is the time spent making the decision. For example, customers might take days or weeks to decide which photo camera they would prefer while making multiple such decisions in a matter of minutes when taking a survey. An increase in task size and complexity may also compromise consumers' judgment; they get tired or bored and might rush through the questions. Moreover, increasing the difficulty of a task might make for a worse interpretation of all of the products due to a limitation in cognitive processing capabilities, which causes an information overload (Louviere and Timmermans, 1992). On the other hand, choices might be simplified in a survey, and consumers can get influenced by other factors in real life, either marketing or non-marketing related, that are difficult to capture

in a study (Allenby et al., 2005).

These aspects impact such that the resulting model has a lower validity than desired. Different measures for validity are used in expressing the performance of studies: internal validity, which shows how well the observed results represent the actual results in the sample; face validity; predictive validity; and external validity. The last shows whether or not the model can be generalized to data outside of the sample (Fitzner, 2007) and is most interesting when looking into the performance of conjoint analysis for the outside world.

While there is much research on the internal validity of conjoint analysis, there is less known about the external validity of the conjoint analysis. Many papers (e.g., Hainmueller et al., 2014; Laurent, 2000; Rogers and Soopramanien, 2009) state that the causes of external invalidity should be investigated. However, new research, as well as most literature, are focused more on the calibration of conjoint analysis, creating newer types of conjoint analysis (such as Filter choice-based conjoint and Build-your-own options), or looking at the usage of different modelling methods (e.g., Liu and Tang, 2015). These methods and improvements help minimise the difference in stated and revealed preferences and, consequently, in the estimation errors of predictions, eventually leading to lower external invalidity. However, only a small number of papers address the factors directly influencing the external validity.

Improving the external validity by looking at the influence of different factors, such as purchase frequency of the product, the number of competitors, or the distribution of the product, has not been discussed that much, while this could be the key to creating more accurate models with the results of conjoint analysis. Therefore, this thesis focuses on answering the following question:

> *Which variables have a significant influence on external validity, and are these variables similar within product categories?*

Finding how and which variables affect consumer behaviour, and, more importantly, how these variables affect consumers such that choices made in a conjoint study and actual purchase behaviour diverge, is of high relevance for both the scientific and the societal field. As there are not many papers that lay the groundwork for this topic, researching this would make way for more economists to look into this subject and expand current research. On the other hand, this research is relevant for marketers in decision-making.

Improving the external validity of conjoint analysis results in more accurate predictions, which will increase the accuracy of a company's scenario analysis. Marketing managers can use the insights from this research to support their decisions better. A meta-analysis is proposed to identify key differences between the results of conjoint analysis and the actual market to answer the research question, considering different segments within the Fast-moving consumer good industry (FMCG).

This proposal is built up as follows: in the following section, the theory behind this problem is discussed. Section 3 holds the data description and research methodology. This methodology is then applied to the data, and the results are given in Section 4. Concluding this thesis, the implications of the results and the connection to the literature, together with limitations and recommendations for feature research, are given in Section 5.

2 Theoretical framework

2.1 Conjoint analysis

Since the introduction of conjoint analysis in 1964 by Luce and Tukey, the method has gained popularity among marketers. The method deduces attribute importance, preferences of attribute levels, and the trade-offs between them. Researchers use surveys to let consumers choose between multiple options of a product. The products that respondents choose from are build up out of relevant attributes, decided on by the researcher. These attributes are what consumers would call 'product features' or 'characteristics'. Those aspects are what consumers base their purchase decision on (Green et al., 2001). Attributes, such as the number of megapixels in a photo camera or the number of calories in a bag of chips, can have multiple levels.

There are multiple ways for collecting data that gives insights into preference. An often-used method is choice-based conjoint analysis. Participants choose the most preferred option from different stimuli: products or descriptions of products shown to a respondent, thus a specific combination of attributes and levels. A respondent chooses one of the shown options, which reveals that he or she prefers that specific combination of the attribute levels over the other combinations shown.

Rating-based conjoint analysis let customers rate stimuli on a pre-determined scale, where higher is a more preferred combination of attribute levels. After rating multiple

stimuli, the stimuli are sorted in a ranking list, where higher means better. The most notable difference with the choice-based conjoint analysis is that respondents do not choose between options and only get to see one stimulus at a time (Orme, 2004).

Besides rating- and choice-based conjoint analysis, there are a few other designs that can be used for conjoint analysis: MaxDiff is a design where respondents select the most important and the least important attribute out of a list; Filter Choice-based conjoint is similar to the traditional choice-based design, but respondents now have the option to filter out irrelevant attribute levels; Build-your-own designs ask respondents to choose desired levels for attributes where each upgrade holds a monetary value, resulting in an estimation of willingness to pay. Build-your-own questions are often used in combination with Adaptive choice-based conjoint, where stimuli are shown based on previous choices. (Cunningham et al., 2010; Orme, 2009).

Performing multiple tasks in a survey where multiple stimuli are shown, part-worth utilities can be derived (Louviere, 1988). Part-worth utilities can be described as the value that a respondent gives to a certain attribute level. The summation of the part-worths results in the preference of a stimulus (Green and Srinivasan, 1978; Luce and Tukey, 1964). Guadagni and Little (2008) and Moore (1980) denote this as

$$\begin{aligned} u_i &= v_i + \epsilon_i, \\ u_i &= \sum_{l=1}^{N} \hat{\beta}_l x_{il} + \epsilon_i, \end{aligned} \tag{1}$$

where u_i describes the preference or utility for stimulus, or product, i. Product i is taken out of a set of alternatives, S. As part-worths can be seen as coefficients in the estimation of total preference u_i, they are denoted by $\hat{\beta}$, where $\hat{\beta}_l$ thus represents the utility estimate of the l^{th} attribute level. Variables x_{il} through x_{iN} hold the description of product i. These variables hold the coding of the attributes of the product. This coding is done with continuous variables, such as price, and dummy variables for non-continuous attributes. The dummy variables represent attribute levels, where a variable is coded 1 when a attribute levels is present in the product, while coded 0 for all levels that are not present. The error-term is represented by ϵ and possibly varies for each choice task. This error is a random component and is likely the result of unobserved variance. It is assumed that this variable is independently distributed. Lastly, N is the number of all possible features of a product by summing all K levels for all J attributes, resulting in

$N = \sum_{j=1}^{J} K_j$ (Guadagni and Little, 2008; Louviere, 1988; Moore, 1980).

The preference calculation as stated in Equation 1 thus gives the preference of a consumer for a product i. When consumers behave rationally, we can assume that they will always select the product that gives them the highest utility. Preference can be summarized using a utility function, therefore it is safe to assume that rational consumers select products with the highest level of preference (Mas-Colell et al., 1995). The probability of choosing alternative i, p_i, is expressed by selecting the product with the highest preference, resulting in

$$p_i = P(u_i \geq u_i', \ \ \forall\, u_i' \in S, i' \neq i). \tag{2}$$

Using Equations 1 and 2, choice probability, p_i, can be rewritten as a multinomial logit (Guadagni and Little, 2008; Theil, 1969):

$$p_i = e^{v_i} \Big/ \sum_{i' \in S} e^{v_i'} \ . \tag{3}$$

2.2 Validity and conjoint analysis

Validity is the umbrella term that holds different measures of the reliability of a model or method. The most interesting measures to look at when assessing the performance of conjoint analysis are face, predictive, internal, and external validity (Fitzner, 2007).

Face validity measures if the survey appears to be measuring what it claims to be measuring. A survey where the respondents know very well what is being tested has high face validity, while a study that has a vague goal has low face validity (Nevo, 1985). Another measure for face validity is *validity by hypothesis*, which looks more to the method rather than the survey. A high validity results from the knowledge that using the method in previous studies, or similar situations, resulted in a highly valid or effective method. Based on the previous outcomes, it is hypothesized that the method will be valid for this objective. In the case of a conjoint study, the method is expected to be highly face valid when previous conjoint analysis showed promising results for a similar study, product and market (Mosier, 1947).

Predictive validity shows how well the model can predict the outcome. A general definition of this validity measure is 'the correlation between a prediction based on a test and some outside variable of interest' (Rogers and Soopramanien, 2009).

Internal validity shows to what extent the observed results, caused by the interested

variables, explain (a change in) the actual results, however only within the data used. Internal validity in the conjoint analysis is based on the utilities of respondents. The results from the conjoint analysis give stated preferences. When a conjoint analysis is highly internally valid, a person's stated preferences for characteristics are causing an effect in their purchase probability, expressed as their individual preference shares, for all respondents (Darmon and Rouziès, 1999).

While internal validity is tested and deduced from the sample data set, external validity looks beyond the sample set. It does not look at an individual's preferences but looks at the total preference shares resulting from the conjoint analysis compared to the actual market shares. It reflects the dynamics and consumer behaviour of the real world, opposed to simulated answers in a lab experiment or survey. Formally, external validity is about generalizing the results of a study. The question, however, is: "generalizing to what?". External validity refers to two fields, both generalizing *to* a particular person, setting or moment in time, as well as generalizing *across* different people, settings or moments (Lucas, 2003).

Having a model with high external validity shows that the critical causal links in the real system, thus the real world, are well represented in the model (Laurent, 2000). Having a high external validity is of high importance, as this can lead to improved marketing decision-making (Rogers and Soopramanien, 2009). It is also crucial as having low external validity can be responsible for the gap between the self-stated decision heuristics and an actual purchase decision (Bremer et al., 2017).

Theoretically, conjoint analysis is supposed to be extremely good in presenting the real world, with results that can be generalized easily. The generalization of the conjoint analysis results would be to say that the stated preferences, expressed in preference shares for all products, should be seen as market shares. If this is not the case, by systematically not being the same as the actual market shares, external invalidity is at play. If the difference is not systematic or significant, this could result from variance rather than bias. In that case, there is not necessarily invalidity at play (Feit et al., 2010).

2.3 External validity and conjoint studies

Even though conjoint studies are preferred for mapping out the market as they simulate real-life decisions, low external validity might still be the biggest downside of the

conjoint analysis. It results as the analysis is only an estimation of the real world (Yang et al., 2018). Low external validity also occurs as conjoint analysis shows information about *choice behaviour*, not necessarily about *purchase behaviour* (Louviere, 1988). While the model, in theory, shows how consumers select and deal with trade-offs, the preferences and actual purchases can be inconsistent (Feit et al., 2010). Effectively, the stated preference and the revealed preference do not align. Also, the simulation of products is not always a perfect representation of the real world (Yang et al., 2018).

Theory suggests multiple ways to fill the gap between the preference shares resulting from the conjoint analysis and the actual market shares. Theory, calibration of conjoint analysis, new types of conjoint analysis, and different modelling methods all try to decrease the gap between the preference shares and market shares. Research regarding these topics focuses on either decreasing the difference between stated and revealed preference, such as the new types of conjoint analysis as Build-your-own exercises; on the other hand, research focuses on translating the conjoint results into good market shares, such as better calibration. This research focuses on the primary, as this study investigated the effects of various variables on the difference between market and preference shares.

The first way to address the difference in preference and market shares is the design of surveys used for conjoint analysis, as well as the respondents. Factors that influence the design are the intensity of the survey and simplification of the products, as well as the number of respondents who have filled out the survey. Increasing the number of choice tasks or the number of choices per task could yield two outcomes: the tasks become too complex, and respondents get exhausted, resulting in a worse outcome (Allenby et al., 2005; Louviere and Timmermans, 1992); or the results of the conjoint analysis improve as there is more data available.

Research conducted by Bansak et al. (2018) and Johnson and Orme (1996) show empirical evidence that the number of tasks does not have a deteriorating effect on the performance of respondents during the survey, leading to more useful information and a better estimation of real-life purchase behaviour (Malhotra et al., 2017).

The factors mentioned above derive the following hypotheses, which test the effect of design variables on the difference between the share of preference and market share:

> **H1**: An increase in the number of tasks decreases the overall difference between preference and market share, as there is more information.

H2: An increase in the number of choices per task results in more information and a better real-life decision-making experience, which decreases the difference between preference shares and market shares.

H3: More respondents increase the representation of the real world, leading to a smaller overall difference.

Secondly, the gap between choice and purchase behaviour is context-dependent. Allenby et al. (2005) suggest that experiments should replicate a specific market context, as this is important for the validity. Hainmueller et al. (2014) support this by stating that external validity is expected to be different based on context and field of study. Moreover, Sichtmann et al. (2011) find that the degree of (in)accuracy of a conjoint model is dependent on product categories. Context-dependency might also have to do with the size of the markets in which the categories operate. Some markets are saturated, hold many competitors and products, while others only have a few products. The impact of context on the difference between preference and market shares is investigated on both study- and product level, using preference shares and product categories.

Purchase frequency is connected to the level of involvement, which is connected to whether or not the product is utilitarian or hedonic. Products that are utilitarian and have a lower level of involvement are bought without much research. These products include products such as a carton of milk or toothpaste. When asking respondents about this kind of product in a survey, respondents take more time thinking about a decision than they would in real life. On the other end of the spectrum, there are products with a hedonic need and a very high level of involvement, such as a photo camera. Respondents might take days, weeks or even longer deciding what product they would buy in real life (Szmigin and Piacentini, 2018). However, in conjoint analysis, respondents are making multiple of these decisions in a very short time frame.

Even though purchase frequency is connected to the actual product, it is assumed that purchase frequency is similar for all products within a study, disregarding packaging volume for the sake of this study. The influence of frequency is assessed throughout this study, testing the following hypothesis:

H4: A high purchase frequency negatively influences the total magnitude of the absolute difference in a study.

As described above, the influence of context is also addressed by looking at product categories and their effect on product-dependent variables. Consumers might focus more on specific attributes in the conjoint than in the real world, such as pricing or packaging volume. This could lead to a bad reflection of reality. Focusing on particular attributes could be a result of choice heuristics (Bremer et al., 2017) or level of involvement. More expensive products often require a higher level of involvement, leading to a more thought-through decision and a minor difference (Sichtmann et al., 2011). A general assumption is that consumers prefer a lower price for the same product.

When the problem of bad reflection occurs, the *hypothetical bias* is at play (Beck et al., 2016). This bias and thus lack of external validity have been researched in a meta-analysis by Murphy et al. (2005), yielding the result that bias seems to be driven by the hypothetical values in willingness-to-pay questions. Higher hypothetical values yield a higher bias; thus, higher monetary values result in lower external validity.

External invalidity might also occur when the customer would prefer combinations of attribute levels that are not available in the actual market and thus has to deal with unmet demand (Chandukala et al., 2011). This problem corresponds to differences in product distribution. Consumers may prefer products that are only available in a small section of the total market, i.e. a small distribution. If a big proportion of the respondents prefer a product that is not widely available, the preference share would be high, while in reality, the product only holds a small fraction of the market (Natter and Feurstein, 2002).

H5: Relatively expensive products are seen as products that have a higher level of involvement, making choosing these a more thought-out choice, leading to a smaller difference between preference and market share for expensive products.

H6: Bigger packages will be chosen less in real life than in the conjoint, as people purchase more cheap in the conjoint.

H7: Products with a lower distribution will have a relatively low market share compared to preference share, meaning that they are overrepresented in the conjoint analysis

H8: Price, distribution, and volume have different effects on the difference between preference and market share on a product level, depending on the product category.

The relations between the variables and the tested hypotheses are schematically shown in the conceptual framework in Figure 1. The figure shows the level of measurement of the variables, as well as the connections between variables.

3 Methodology

This section addresses an explanation of the data and the necessary steps for preparing the data, as well as explaining the different techniques. Multiple statistical tests and machine learning techniques were applied to test all given hypotheses. These techniques and analyses were applied in Python, with the use of SciPy (Virtanen et al., 2020), scikit-learn (Pedregosa et al., 2011), Keras (Chollet et al., 2015), and statsmodels (Seabold and Perktold, 2010), as these are the most advanced and prominent Python modules for machine learning and deep learning[1].

Hypotheses 1 through 4 are tested by fitting a regression model to predict the total magnitude of the absolute differences between preference and market shares. The effects that are to be tested are the same within each study, and therefore there is one observation per study used. If variables have little impact on the magnitude of difference, the quality of the conjoint analysis is not widely affected by these factors. Thus, the model would be externally valid for that product.

Hypotheses 5, 6, 7, and 8 are also addressed using regression analysis, where the effect of Price, Distribution, and Volume on the difference between preference share and market share per product is assessed. These variables, as different per product, are of high impact on purchase in general, as discussed in Section 2. The difference used in this analysis is the difference on an individual product level: Preference share minus Market share. The explanatory variables used from Hypotheses 1 through 4 are included as control variables.

The regression is also expanded with interaction effects between product categories and price, volume, and distribution, addressing Hypothesis 8 to find similarities between studies.

[1]The implementation of this thesis can be found at https://github.com/l-schipper/MSc-Thesis-final

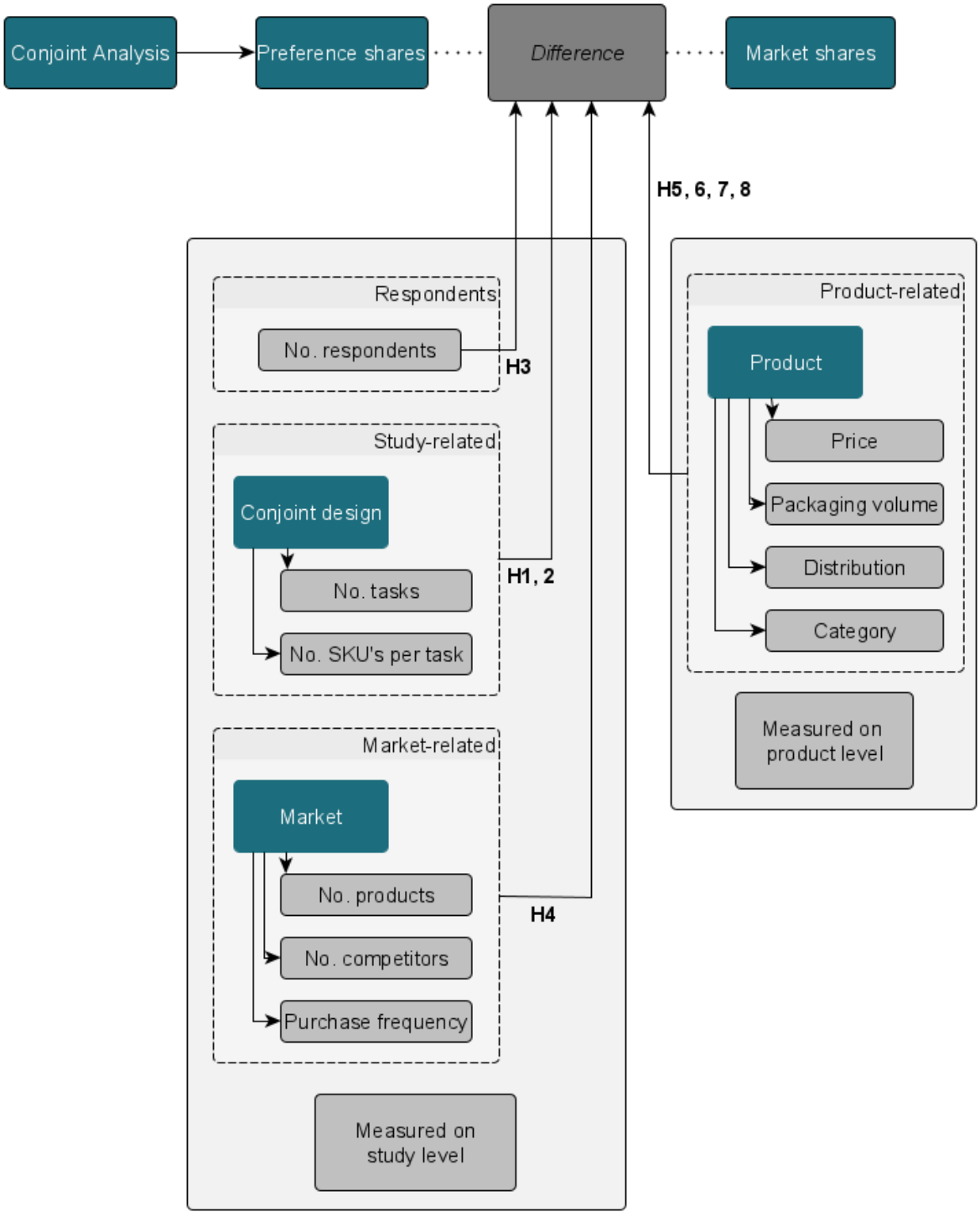

Figure 1: Conceptual framework of variables and their relations.

3.1 Data description

For this research, 33 simulators of previously executed conjoint studies within SKIM, where the focus has been on pricing, are used. Each simulator contains a various number of products or SKUs. All products in a simulator, or study, are seen as all products in that market. The number of SKUs in the conjoint analysis are thus the same as the number of SKUs in the market. All shares are scaled such that they make up 100% exactly. Combining all products of all studies leads to a total of 1,876 individual products.

Hypotheses 1 through 4 use 33 observations, one for each study, because all explanatory variables are the same within a study. The dependent variable for the models created for these hypotheses is as follows;

$$Total\,magnitude\,of\,absolute\,difference_i = \sum_{j=1}^{J} |preference\,share_{ij} - market\,share_{ij}|, \tag{4}$$

where i represents a study, j represent a product in study i and J equals the number of products in that study. The preference share of a product j in a study i is thus denoted as $preference\,share_{ij}$. The total magnitude of absolute difference thus is the sum of the absolute values of the differences per product per study.

Hypotheses 5 through 8 use all 1,876 individual products, as these hypotheses assess differences on product level. The dependent variable used in the analyses regarding those hypotheses is

$$Difference\,on\,product\,level_j = preference\,share_j - market\,share_j, \tag{5}$$

where j represents an individual product. This equation is not dependent on study; however, the sum of all differences on product level within a study equals 0. This results as both the sum of preference share and the sum of market share within a study equal 100%.

The simulators used for this study hold information about the actual price, volume, distribution, and market share of a product, at the moment of surveying. Using the results of the surveys, given as part-worth utilities, Preference shares of the different products are computed following Equation 2, where the input is the result of the survey. Products that are not requested in this analysis by the company are not considered when computing the preference shares. Therefore, not all possible attribute (level) combinations

are represented in the preference shares. However, this study focuses on the translation from the conjoint to the market, and the gaps found there. The market shares of the products researched are scaled to have a total of 100%, such that the preference and market shares can be compared without accounting for other products.

The actual market shares are acquired through Nielsen. The shares are aggregated on the channel level, which is study-dependent, e.g. a specific supermarket chain or nationwide. The shares are then scaled such that the products in the scenarios make up a hundred per cent, as mentioned before.

The numbers for distribution are also acquired through Nielsen. These distributions are weighted based on the selling volume of the stores. A store with a higher sales volume is accounted for with a more considerable weight, meaning that a big store that does not sell a specific product has more impact than a smaller store in the eventual Distribution number.

The variables for price, distribution, and volume were scaled. These values of these variables are often category dependent, e.g. a carton of milk is less expensive than a bottle of perfume and could very well be hugely divergent between studies. Therefore the variables are scaled into the relative price, distribution and volume by dividing by the study's average. A relative Price of 1 means that this price is the average of that study. A relative Price of 1.5 thus means that this price is 50% higher than the average price. Standardizing the values of these variables is necessary, as the parameters could very well be misleading without standardizing first. A relatively expensive product in study A could have the same price as an incredibly cheap product in study B. Not using relative numbers could lead to wrong conclusions of the importance of price, distribution, and volume. Besides, the parameter and significance in a regression could lead to wrong insights. Therefore scaling is applied to ensure that the study dependent values of price, volume or distribution do not cause issues.

Interpreting their coefficients in the models is then done as follows: An increase of 1 in price means that the difference increases with its parameter value. An increase of 1 in the standardized price means an increase of 100% relative to the mean price for that study. Thus, the product's actual price has increased with the average value for price in its study. A more concrete example is as follow: when a product has a price of $60 and the average price in that market is $50, an increase of 1 in the standardized price means

that the product now would cost $110.

The variables mentioned above all hold information on product level. Six variables are included to get a better idea of the influence of the market- and study-related variables. These variables give insight into the market and the conjoint study itself. Three variables are included to look at the market: (1) the number of products in the market, (2) the number of competitors, and (3) the purchase frequency of the product. The market here is defined by the products used in the simulator. The number of products thus equals the total number of products used in the simulator; the number of competitors equals the number of brands in the simulator.

Purchase frequency is a self-defined variable where all products are divided into one of the following categories: daily purchase, weekly, monthly, quarterly, less frequent than quarterly. A new photo camera or mobile phone would then be categorized as 'less frequent', while a milk bottle is bought weekly.

Study-related variables consider the design of the surveys used for conjoint analysis and respondents. The resulting variables are (1) the number of respondents for this survey, (2) the number of choice tasks a respondent had to do, and (3) the number of options per choice task. These variables are taken into account as the second and third variables might deal with the cognitive issue, as discussed in Section 2.

Lastly, product categories were added using dummy variables. All studies are about FMCG products; however, product categories are addressed as well. A product category is added as a dummy when there are at least two studies within that category. The categories included in the analyses are three studies about beauty products; eight studies about cigarettes; two involving snacks; seven studies regarding various dairy products; six about batteries; three studies about drinks; and two regarding home products. The two remaining studies are categorized as 'other' and will function as the base level in the regressions.

Table 1 shows the summary statistics of the study and market-related variables as found by the 1,876 different products found in all studies combined. The number of respondents differs considerably, where the maximum is twenty times larger than the minimum. Its standard deviation is also massive. The median is on the left of the mean, showing that this variable is positively skewed. A positive skew also occurs for the number of SKUs per task. The number of choice tasks per study shows a small range, where the

minimum and the first quantile have the same value, and the median and third quantile also have the same value. It can be deduced that the number of tasks does not vary a lot. The number of competitors seems to be somewhat normally distributed.

The two variables at the bottom rows show the variables of interest: the differences between preference and market share. The first one shows the difference between Preference and Market share per product, where the second is subtracted from the first, as shown in Equation 5. A negative difference thus means that the market share is greater than the preference share. There is a wide range with a negative skew.

The magnitude variable is the sum of the absolute differences between preference and market shares per study, as seen in Equation 4. A lower magnitude means that overall the Market shares are predicted well, whereas a high Magnitude shows a bad prediction. There is a slight positive skew in the magnitude of differences, showing that there might be outliers regarding this variable.

Scatterplots are included to get an idea of how the variables impact the dependent variables. Figures 2 and 3 hold scatterplots of the aforementioned market-related and design-related variables, with the dependent variables on the y-axes. The difference on product level is used in Figures a through c; the magnitude of total absolute difference per study is used in Figures d through f. Figure 4 also consists of different scatterplots. Only the difference on product level is used, since price, volume, and distribution are on

Table 1: Summary statistics

	mean	std	min	25%	50%	75%	max
No. respondents	1398.57	1035.73	202	728	869	2044	4076
No. tasks	10.532	3.478	6	6	12	12	16
No. SKU's per task	24.866	18.380	2	14	14	31	69
No. SKU's in market	128.564	77.854	8	55	100	230	230
No. competitors	15.020	5.139	2	12	16	18	27
Difference	-0.0005	0.02531	-0.4518	-0.0037	0.0007	0.0048	0.1740
Magnitude of total absolute difference	0.654	0.199	0.206	0.514	0.685	0.741	1.507
N = 1,876							

product level while the magnitude is on study level.

Figure 2 shows scatterplots for the number of tasks (Figures a and d), the number of SKUs per task (Figures b and e), and the number of SKUs in the market (Figures c and f). The number of tasks seems to have a wider range of differences for every value other than 6 and 15. On the magnitude, it appears that an increase of Tasks increases the magnitude non-linearly. For the number of SKUs per task, the range decreases the difference per product and the magnitude of difference when there are more SKUs per task. The number of products per market also affects the difference, such that more Products result in a smaller difference between preference and market share. For magnitude, the number of products seems to incline slightly if the outlier is neglected.

Looking at Figures 3a and d, it is visible that an increase in Respondents leads to a smaller Magnitude of Difference and a smaller range of difference on product level. The number of competitors (Figure 3b and e) also decreases the difference on a product level,

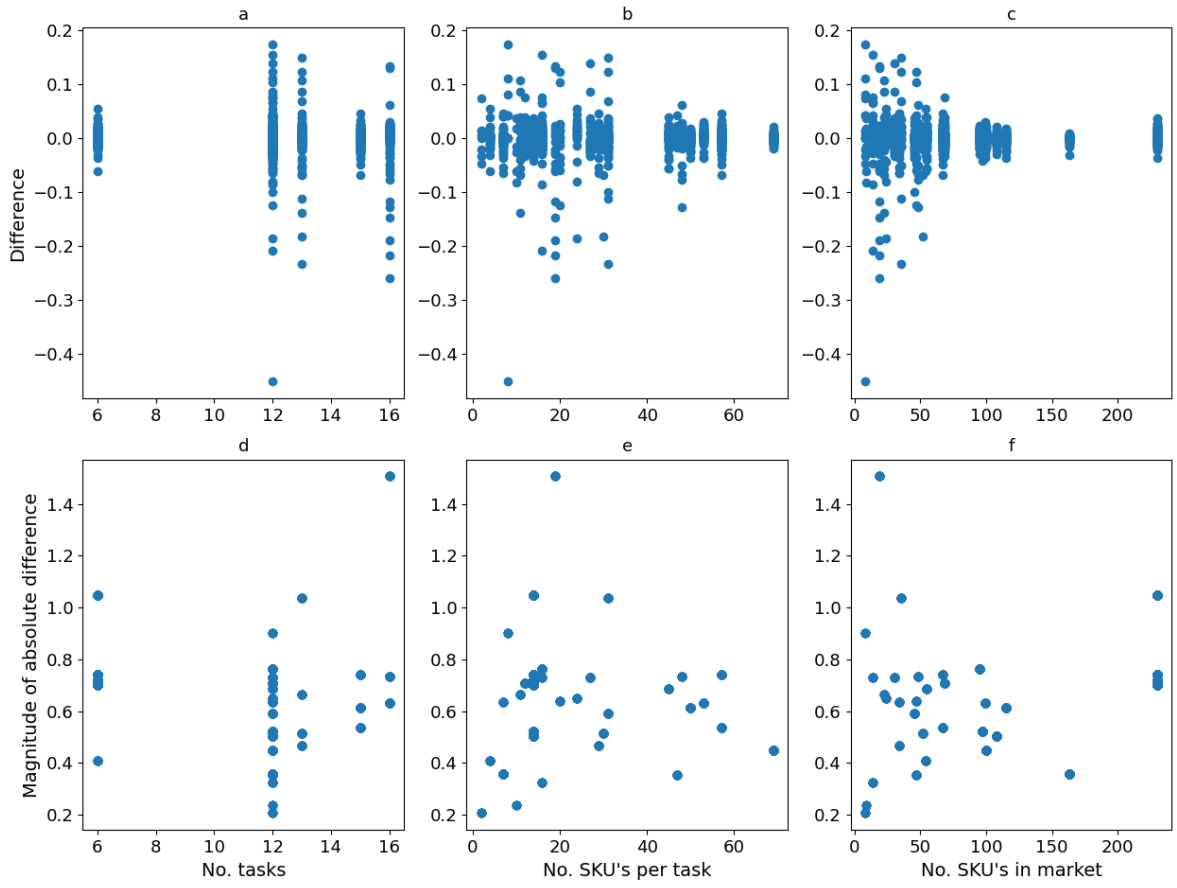

Figure 2: Scatterplot of the relation between Difference and Magnitude of total absolute Difference, and Number of tasks, Number of SKUs per tasks and Number of SKUs per market.

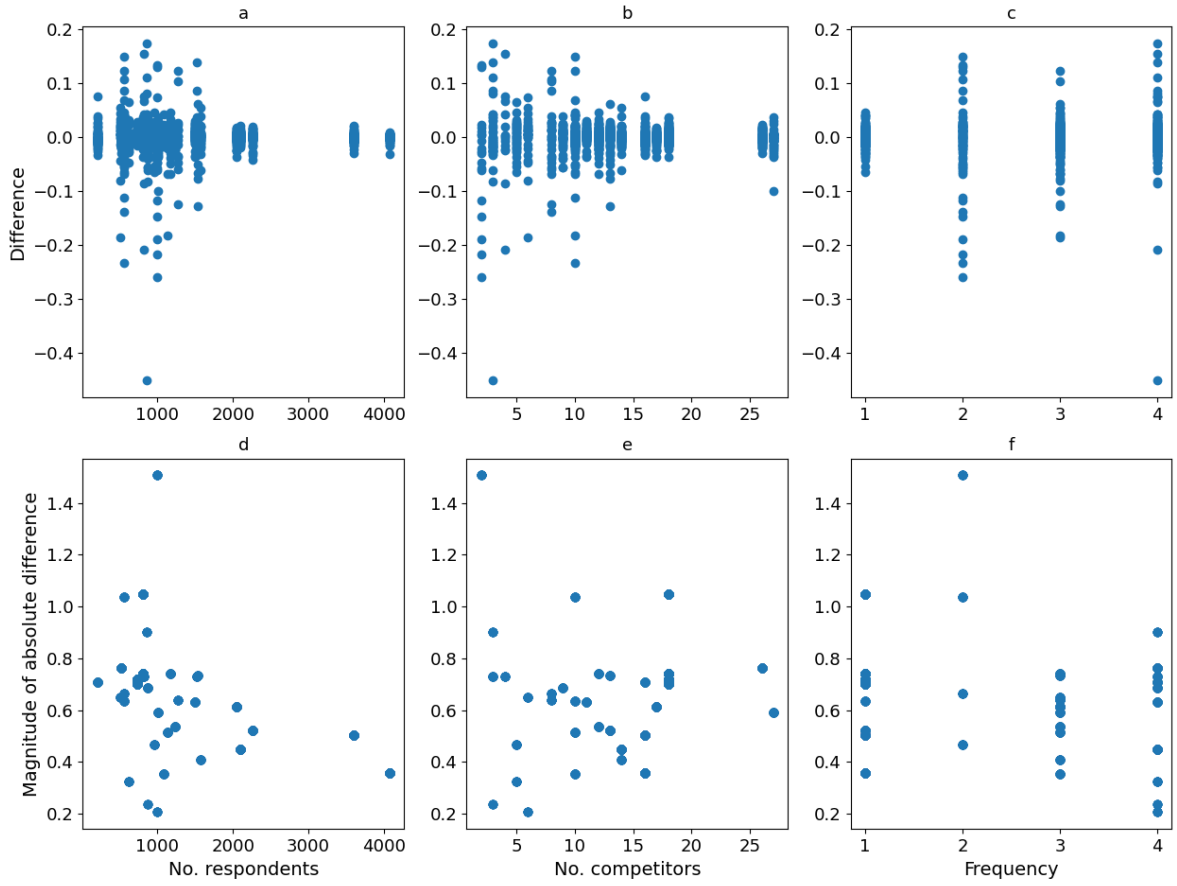

Figure 3: Scatterplot of the relation between Difference and Magnitude of total absolute Difference, and Number of respondents, Number of competitors and Purchase frequency.

but for the magnitude, there is no clear relation. When looking at frequency, the first level (daily purchase) has a much smaller range of difference than the other levels. There seems to be no clear relation with the magnitude of difference; however, level 2 has a possible outlier.

Finally, Figure 4 shows the relations of Price, Volume, and Distribution, to difference. Right away, it is evident that a higher standardized price results in a minor Difference. The same appears to happen for volume, but this is less clear and might result from outliers. For distribution, the opposite occurs: a higher Distribution leads to a greater Difference. This is in line with theory as previously discussed. However, as the relation between the explanatory variables and the dependent variables is not very clear in the scatterplots, models are needed to investigate any relationship with a significant impact on the difference. Only when a significant impact can be found can it be said that there is a variable influencing the external validity.

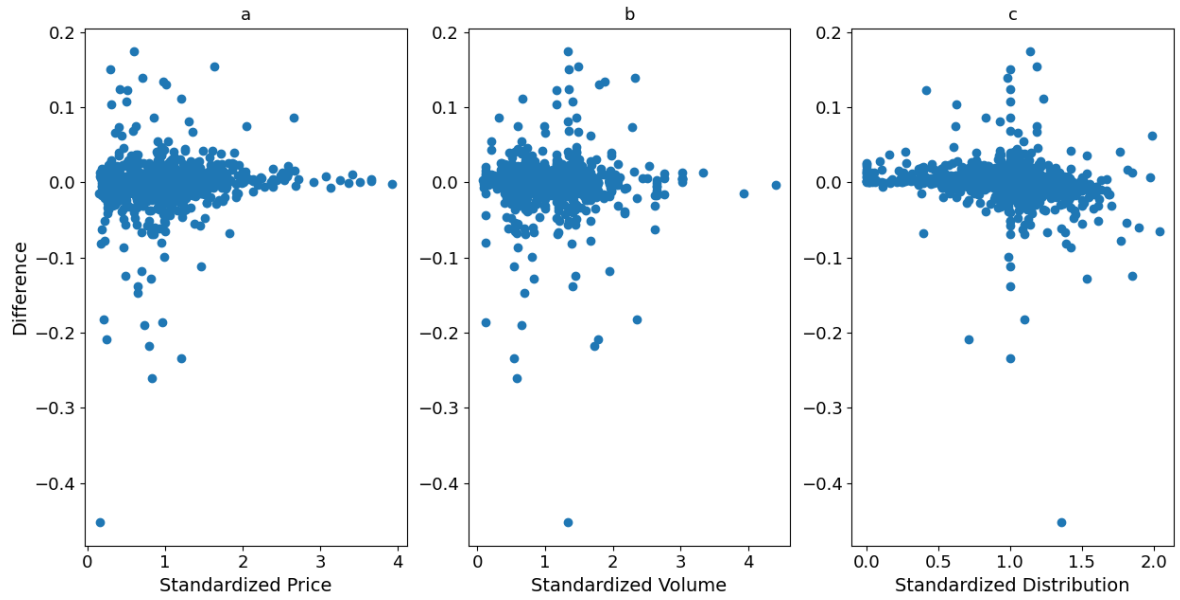

Figure 4: Scatterplot of the relation between Difference, and standardized Price, standardized Volume, and standardized Distribution.

The data is randomly split into two portions for all analyses, creating a training and validation set. Splitting is done by randomly assigning 70% to the in-sample set, whereas the rest is considered out of the training sample, thus out-of-sample set.

3.2 Regression analysis

As stated before, most of the hypotheses are addressed using regression analysis. The most ordinary regression analysis, Ordinary Least Square (OLS)-regression or linear regression, fits a model while minimizing the residual sum of squares (RSS). However, OLS works under assumptions. One of these assumptions is the absence of (multi-)collinearity, as (multi-)collinearity causes problems with interpreting the coefficients. The problem with collinearity is that there are predictor variables closely related, meaning there is a strong correlation between them. The correlation makes it difficult to separate the individual effects of the variables. This problem is expected in this data set, as the number of products and the number of competitors on the market might be correlated, as well as the number of products and the variables regarding the design of the conjoint analysis. A study for a market with many different products and different attribute levels might result in more tasks or more options per task. Collinearity might slightly reduce the accuracy of a model, which then causes the standard error of the affected coefficients

to grow. When two variables in the model are correlated, it is called collinearity and can be checked by looking at correlations of the variables. An easy solution would be to discard either of the variables (Stock and Watson, 2014).

It gets more difficult when more than two variables are correlated as such, resulting in *multicollinearity*. This is not picked up by single correlations but can be investigated using the variance inflation factor (VIF) of the regression. The VIF is a ratio of the variance of the coefficient $\hat{\beta}_j$ in the total model, divided by the variance of the coefficient of $\hat{\beta}_j$ when fitting the model using only this variable:

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}, \tag{6}$$

where the denominator holds $R_{X_j|X_{-j}}$, the R^2 from a regression of X_j on all the other predictors. The smallest value of VIF is 1, meaning no multicollinearity. A value above 5 indicates a problematic amount of multicollinearity, according to the rule of thumb. When $R_{X_j|X_{-j}}$ is big (close to one), the VIF will be large, indicating that there is collinearity is present (James et al., 2013).

Hypotheses 1 through 4 are addressed using a regression with the magnitude of the absolute difference as dependent variable:

$$\begin{aligned}\text{Magnitude of absolute difference}_i = \beta_0 &+ \beta_1(\text{Number of respondents}_i) \\ &+ \beta_2(\text{Number of choice tasks}_i) + \beta_3(\text{Number of SKUs per task}_i) \\ &+ \beta_4(\text{Number of SKUs in market}_i) + \beta_5(\text{Number of competitors}_i) \\ &+ \beta_6(\text{Purchase frequency}_i),\end{aligned} \tag{7}$$

where i represents a study. To make sure the assumption no multicollinearity is not violated, Table 2 shows the VIF for the variables that are used in this model. As can be seen here, there is some multicollinearity regarding both the Number of choice tasks and the Number of products in the market, as both VIF's are above 5. Therefore, shrinkage methods are applied, as to divert the consequences of correlation.

3.3 Shrinkage methods

Shrinkage methods are based on the multiple regression as given above but penalize, or regularize, the model's coefficients such that the correlation is not affecting the coefficients. The two most used shrinkage methods are *Ridge regression*, which shrinks

Table 2: Variance Inflation Factor for variables in first model

Variable	VIF
Purchase frequency	3.79
Number of respondents	1.73
Number of choice tasks	5.85
Number of SKU's per task	2.41
Number of SKU's in market	6.30
Number of competitors	1.65

All variables, except Purchase frequency, are scaled using min-max normalization.

the parameters of all independent variables; and *Lasso*, which uses feature selection by shrinking selected parameters to exactly 0. The *Elastic Net regression* combines the two models and shrinks all parameters as in the Ridge regression while also using the feature selection from the Lasso method to shrink parameters to 0 (James et al., 2013).

As the three methods work slightly differently, selecting one to use is not done beforehand. After computing all, the best-fitting model is selected by looking at the Mean-Squared Error (MSE). The MSE is an often-used performance measure and is computed as;

$$MSE = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}, \tag{8}$$

where y_i holds the actual value, $\hat{y}_i$ the predicted value and N the number of observations (Ahmad et al., 2017).

3.3.1 Ridge regression

As stated, Ridge regression is an adaptation of (multiple) linear regression that regularizes coefficients. The model is found by minimizing the loss function in either Equation 9a or 9b, which both are based on the RSS as is used in linear regression. The first part of the loss function is called the regression term, while the second term of the function is the term specific to shrinkage methods: the penalty term. The Ridge loss-function that

is to be minimized, is as follows (James et al., 2013; Tibshirani, 1996);

$$L_{ridge1}(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2; \tag{9a}$$

$$L_{ridge2}(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq t \text{ for } t > 0; \tag{9b}$$

where L_{ridge1} and L_{ridge2} are two mathematical expressions for minimizing the loss-function; both yield the same result. The regression weights for the variable j are represented by the β_j's, x_{ij} represent predictor variables, and y_i equals the value of the dependent variable (James et al., 2013). The penalty term, $\lambda \sum_{j=1}^{p} \beta_j^2$, holds the penalty parameter λ, which gives the level of regularization. A higher λ means a higher level of shrinkage of the β's. The best value for this parameter is to be estimated using cross-validation. The variable t as presented in the second mathematical expression, has a relation with λ (Tibshirani, 1996).

3.3.2 Lasso regression

Another option for a shrinkage model is a Lasso regression. As mentioned, this method does not shrink the coefficients of important variables as much as Ridge does; rather, it eliminates variables that are not of great importance by shrinking their parameters to 0. The remaining parameters are slightly regularized after the variables selection. This variable selection is thus the key component of this method. Similar to the Ridge regression, Lasso builds on the idea of an OLS regression by feature selection. Equations 10a and 10b show how the calculations are done for this method. The same regression term is present as in the Ridge regression formulas, but the penalty term deviates. One of the formulas below is to be minimized in order to find the Lasso coefficient parameters:

$$L_{lasso1}(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|; \tag{10a}$$

$$L_{lasso2}(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t \text{ for } t > 0; \tag{10b}$$

where the variables have the same meaning as in the Ridge regression. Similar to Equations 9a and 9b, Equations 10a and 10b yield the same outcome; however their mathematical expression differs (James et al., 2013; Tibshirani, 1996).

3.3.3 Combining into elastic net

Both of the methods have their shortcomings. Ridge keeps all variables, even though some might be uninteresting, while the feature selection of the Lasso might be too dependent on the data sample and can be unstable. Combining the two and keeping their strengths can be a solid option. This combination is done by building on the OLS and combining both penalty terms into one whilst adding with an extra parameter that shows to what extent either penalty term is used. The loss-function for the elastic net is;

$$L(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right), \qquad (11)$$

where β's are still regression coefficients, or weights, λ represents the penalty parameter and where α is a mixing parameter between 0 and 1, regulating the mix between Ridge and Lasso. A high α indicates more Lasso, while a small α indicates more Ridge. When the mixing parameter equals 0.5, it holds as much of the characteristics of Lasso as of Ridge (James et al., 2013).

All three shrinkage methods were used to find the best fitting model as described in Equation 7. The regularization parameter for each method is selected using Leave-one-out cross-validation (LOOCV). This method is a more extreme version of the commonly-used k-fold cross-validation (CV), where the total sample is divided into k-folds or partitions. The main objective of CV is to make sure the same observations are not drawn multiple times while other observations are not drawn at all. LOOCV performs k-fold CV with $k = n$ observations. For every fold, a single observation is left out, which can never be the same observation. As there are much more folds needed to perform LOOCV, it can be slow to compute but is beneficial for smaller data sets. LOOCV is, in this case, used to find the optimal regularization parameter for each method, thus λ. The optimal regularization parameter ensures the best-performing model, thus where the MSE is lowest (Lantz, 2013).

3.4 Influence on product level

After addressing difference between preference and market share on study level, the influence of variables on product level is investigated, addressing Hypotheses 5 through 7. To test the impact of price, distribution, and volume, a linear regression is used. The

variables used in the previous regressions are included as control variables. The resulting linear regression thus has the form;

$$\begin{aligned} Difference\,per\,product = {} & Price + Distribution + Volume \\ & + Number\,of\,respondents + Number\,of\,choice\,tasks \\ & + Number\,of\,SKUs\,per\,task + Number\,of\,SKUs\,in\,market \\ & + Number\,of\,competitors + Purchase\,frequency \\ & + Product\,category, \end{aligned} \tag{12}$$

where purchase frequency and product category are represented using dummy variables. The difference per product is computed as in Equation 5. The regression is also conducted with the absolute difference per product as dependent variable, to identify the size of the difference, which is the absolute value of the Equation 5.

3.5 Similarities within product categories

Hypothesis 8 looks at the similarities across studies, more specifically it addresses significant effects of product categories and significant differences between categories. To find these similarities, Equation 12 is expanded. Interaction effects between product categories and price, volume, and distribution, are added. Significant interaction effects show a significant difference in the slope of the regression. The individual variables for price, distribution and volume are excluded as they are present in the interaction effects. The dummy variables for product categories are not excluded, as significant effect hold information on the constant. Significant differences indicate that the product categories always have a higher or lower difference, on average. The regression that is used is as follows:

$$\begin{aligned} Difference\,per\,product = {} & Price \times product\,category + Distribution \times product\,category \\ & + Volume \times product\,category + Number\,of\,respondents \\ & + Number\,of\,choice\,tasks + Number\,of\,SKUs\,per\,task \\ & + Number\,of\,SKUs\,in\,market + Number\,of\,competitors \\ & + Purchase\,frequency + Product\,category. \end{aligned} \tag{13}$$

3.6 Predicting market shares with a black box

The variables included in this research might not be sufficient in the prediction of the actual market shares. Several variables influence customer decisions, such as placement, discounts, and seasonal changes, which are sometimes hard or even impossible to include in conjoint analysis. These variables are not included in the used data sets and thus might interfere in providing the best results. Therefore, it is expected that there is still variance unexplained when predicting the difference between preference and market shares or when trying to predict market shares using the preference shares. Actual Market shares are predicted using black-box models, with preference shares and all other variables as predictor variables. Predictions are made using these methods, not to understand why the results from a conjoint study work well; but to improve the conjoint study's predictions.

Black-box models, such as Support Vector Machines (SVM) and Neural Networks (NN), as well as Random Forests (RF), have a low level of interpretability of the algorithm that yields the predictions or results. However, the accuracy of the fitted models can be high, as is their predictive power (Lantz, 2013). Extensive explanations of the Random Forest, Support Vector Machine, and Neural Network are given in Appendix A.

The models are compared using different performance measures, and the models are used for predicting the Market shares of a completely new data set. The measures that are used are the MSE and Root Mean-Squared Error (RMSE), Mean Absolute Deviation (MAD), and the Mean Absolute Percentage Error (MAPE). These often-used performance measures can be calculated for each of the models and are reliable measures. They are calculated by the equations below (Ahmad et al., 2017);

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}}; \tag{14}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i - \hat{y}_i}{y_i} \times 100; \tag{15}$$

$$MAD = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{16}$$

where N represent sample size, $\hat{y}$ are predicted values and y are actual values.

Using the black box-methods aims to improve the market share prediction such that future predictions are improved. The black box-methods correct for the small variance explained that is expected to occur when predicting market share using only preference

shares. Preference shares hold the information on utilities of consumers and are used to predict market share.

4 Results

In this section, the results of the analyses are given. This is done in order of the hypotheses, meaning that first, the magnitude of absolute differences is assessed, followed by analysing the influence of a product's price, volume, and Distribution in Section 4.2. Next, the categorical effects are addressed. Finally, the results of the black-box models are given in Section 4.4.

4.1 Magnitude using study and market variables

The magnitude of the absolute difference per study is researched using study and market-related variables. More specifically, this analysis looks to find if the number of tasks (H1), Choices per task (H2), number of respondents (H3), and Purchase frequency (H4) have a negative influence on the magnitude of difference. As described in Section 3.3, this is done using shrinkage models, as there is multicollinearity present (Table 2). To find the best fit for each model, LOOCV is used when estimating the best parameters. This leads to the models as given in Table 3.

All variables, except purchase frequency, as it is a dummy variable, it does not need to be scaled, are scaled using min-max normalisation; the interpretation of these variables is: an increase of 1%, relative to the range of the variable, would lead to an increase in the magnitude with 1% of the size of the found coefficient, ceteris paribus[2]. This results in the following example when looking at the Ridge model. An increase of 1% in the number of respondents would lead to a decrease of 0.00156 in the magnitude of the absolute difference. The minimum number of respondents is 202, and the maximum is 4076. An increase in Respondents of 1 per cent, relative to the range equals $(4,076 - 202) * 0.01 = 38.74$ respondents: an increase of 38.74 respondents thus leads to a decrease of, on average, 0.00156 on the magnitude of the difference, ceteris paribus.

Table 3 shows the outcome for the best fitted parameters for all three methods,

[2]An increase of 1 of any of the coefficient, except frequency, would mean an increase of 100% of the range, as 1 states the maximum of the range.

Table 3: Shrinkage regressions on magnitude of absolute difference

Variable	Ridge	Lasso	Elastic Net
Dependent: magnitude of absolute difference			
Intercept	0.609	0.575	0.524
No. respondents	-0.156	-0.163	-0.277
No. tasks	0.051	0	0.112
No. SKU's per task	0.040	0	0
No. SKU's in market	0.104	0.184	0.284
No. competitors	0.025	0	0
Frequency: weekly	0.178	0.297	0.288
Frequency: monthly	-0.078	0	-0.012
Frequency: quarterly	-0.087	-0.006	-0.033
Regularization parameter	2.009	0.008	0.008
R^2	0.319	0.333	0.394
MSE - in sample	0.002	0.076	0.076
MSE - out of sample	0.040	0.021	0.045

Notes: The mixing parameter is set to 0.5 for the elastic net function.

where the regularization parameter is represented by λ in Equations 9, 10 and 11. For the Elastic Net, the mixing parameter is set to 0.5. Given the performance measures, it would seem that the Ridge regression has the lowest, thus best, Mean-Squared Error (MSE) in-sample. However, the out-of-sample MSE is much higher (0.040), indicating overfitting. For both the Lasso and Elastic Net-regressions, the out-of-sample MSE are lower than those in-sample, which might indicate underfitting of the sample (James et al., 2013). This shows that all three models do not predict new data well, indicating difficulty in generalising the model to new data.

As magnitude ranges from 0.206 to 1.507 (Table 1), the MSE's for all models are small enough to indicate a well-performing model. However, when selecting one of the models, looking at the out-of-sample MSE would give the best information (James et al., 2013);

therefore, the Lasso is seen as outperforming the others. Figure 5 shows the weights of the explanatory variables for every lambda, where the dotted line shows the best-performing lambda. Immediately it can be seen that there are four variables with very high absolute weights, compared to the other four. Weekly frequency and the number of tasks, Respondents, and SKUs in the market all have very high absolute weights. In contrast, the other dummy variables for frequency, number of competitors, and the number of options per choice task have very low weights. This last group also contains the variables that are deselected first, meaning they are less significant in estimating the magnitude of the absolute difference.

Interestingly, increasing the number of tasks seems to result in a worse external validity. As discussed in Section 2.3, the increase in the number of tasks could lead to two possible outcomes: (1) there is more data, therefore results improve, and the difference decreases (Bansak et al., 2018; Johnson and Orme, 1996), or (2) respondents get exhausted after a while, due to information overload and limited cognitive processing capabilities,

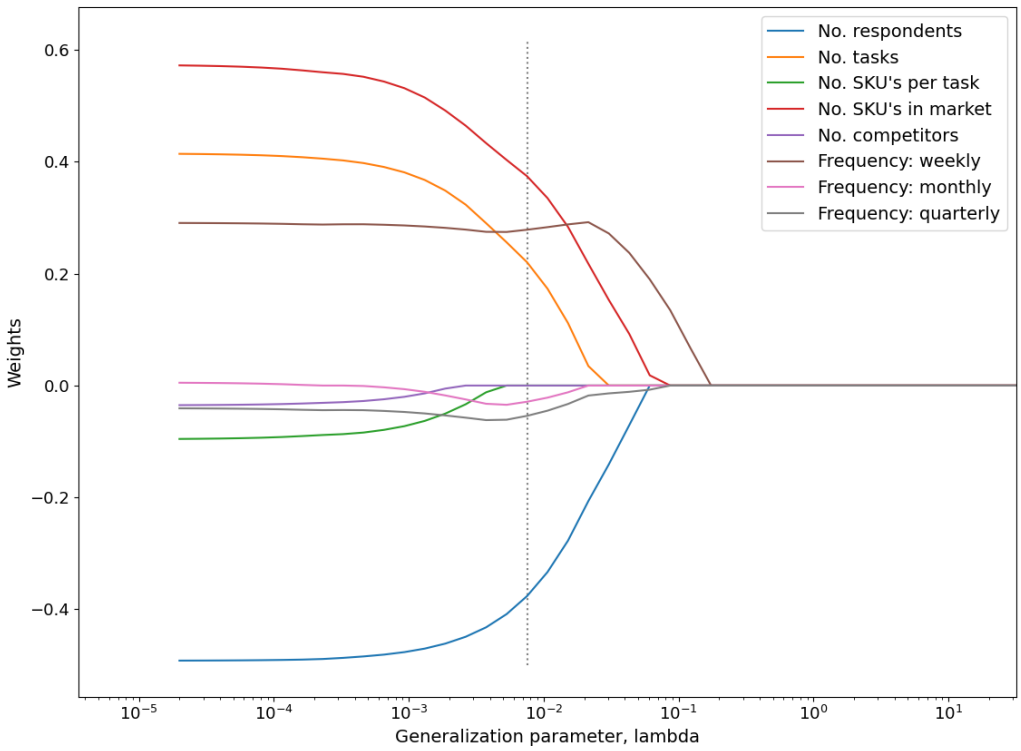

Figure 5: Weights per generalization parameter for chosen Lasso model.

leading to an increase in difference (Allenby et al., 2005; Louviere and Timmermans, 1992). The results as given in Table 3 would suggest the second outcome. When looking into this variable more closely, Figure 6 results. The magnitude of the difference of each study is set against the number of tasks; however, there is no control for other variables. Besides, an outlier has been identified in the magnitude of differences. Excluding the outlier results in a different linear slope: where it is increasing in Figure 6, Figure 12 in Appendix B shows a declining slope. The outlier is not excluded for the analyses in this study, as the interpretation of the original data appears valid. There are no obvious disruptive products and differences on a product level in the original data set; the resulting magnitude might appear as an outlier. It is not necessary to exclude it based on interpretation.

As stated, there is a positive linear slope in Figure 6, meaning that an increase in Tasks would lead to an increase in the magnitude of difference. However, both the median, minimum, and maximum per number of tasks seem to follow a curve that slopes down from 6 to 10 tasks. It increases and dips around fifteen tasks before increasing at sixteen again. These relations are created by interpolating the values of observations using Python's SciPy (Virtanen et al., 2020). Quadratic interpolating results in the four other functions given in Figure 6, where the values used for interpolation were either the mean, minimum, maximum or median of the number of tasks. The functions that make use of the mean, minimum, and maximum values show a somewhat similar relation; however, the size of the minimum and maximum in this plot differ. Nonetheless, either of these relations shows a minimum at nine tasks; a dip between fourteen and fifteen tasks; and a maximum at six, thirteen, and sixteen tasks.

Looking at the median shows a different image. There it can be seen that an increase to fourteen tasks results in a minimum average magnitude of difference, after which increasing the number of tasks results in an immediate increase of the magnitude of difference. Given the optimum number of tasks and Lasso parameters, Hypothesis 1 cannot be accepted. It cannot be said that increasing the number of tasks always leads to a decrease in magnitude, as the Lasso model shows an increase in Magnitude and the interpolated functions show no clear decrease in magnitude.

The number of products-variable is present in all of the models and holds positive coefficients. This shows that increasing the number of products researched in a conjoint study increases the magnitude of the difference.

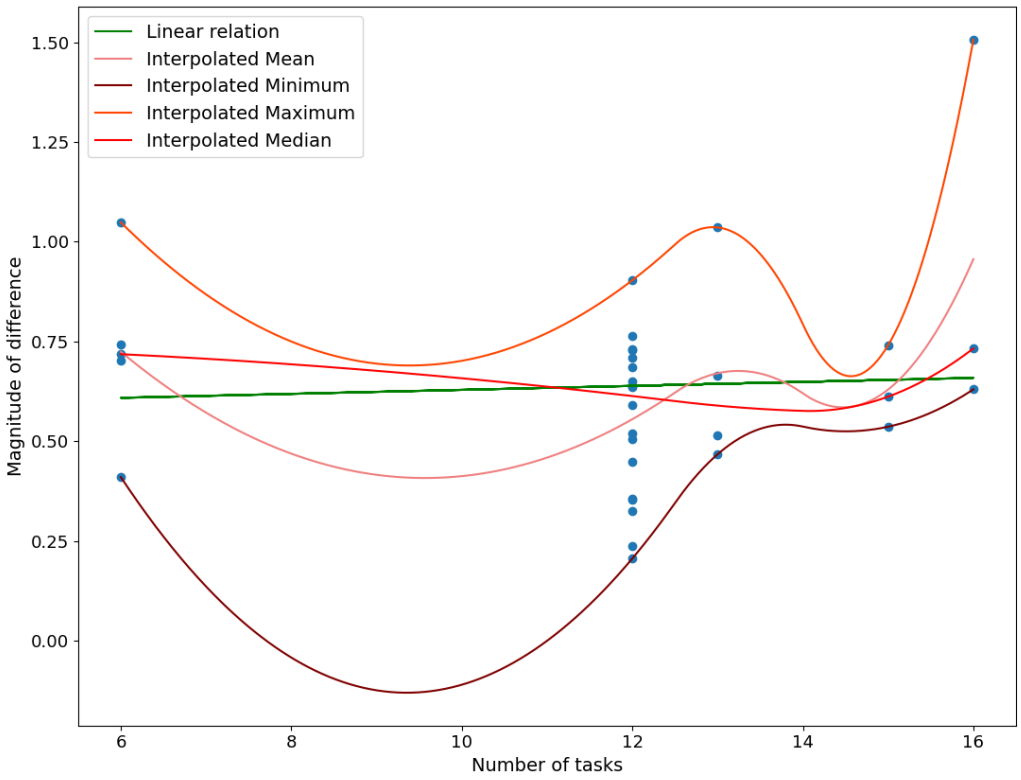

Figure 6: Relationship between Number of tasks and Magnitude

The coefficient for the number of competitors is minimal for the Ridge regression and 0 for both Lasso and Elastic Net. It can thus be viewed as not interesting enough when determining the magnitude of difference. Similar results can be seen when looking at the SKUs per task. Looking at the Ridge model, an increase of SKUs per task leads to an increase of the magnitude, decreasing external validity. However, this variable is not selected in Lasso and Elastic Net, indicating that this variable is not that significant for determining the magnitude of the difference. Consequently, Hypothesis 2 cannot be accepted.

A relatively high number of respondents leads to a lower magnitude of differences for all three models, thus a better external validity. Given the chosen model, Hypothesis 3 is accepted.

Lastly, Purchase frequency has also been added to the model. A dummy variable is added for all categories, where 'daily' functions as a base. This means that the coefficients for frequency are relative to 'daily'. Immediately it is visible that a lower Frequency,

thus monthly or quarterly, results in a lower Magnitude compared to daily products, where quarterly also has a more negative impact on the magnitude compared to monthly. This does not go to say for weekly products, as they increase magnitude compared to daily products. Given the Lasso mode, the variable for monthly is 0, meaning it is not significantly different from the daily products. Overall, it can be said that products with a higher Purchase frequency do have a lower Magnitude, except for the increase in magnitude when looking at weekly products compared to daily products. Therefore, Hypothesis 4 cannot be accepted with complete certainty. The positive parameter for weekly could result from the number of observations per frequency level and the outlier that also interfered with the number of tasks. The outlier, visible in Figure 3f, causes a much wider range for the 'weekly' level compared to the other levels.

4.2 Effect of product-specific variables

The remaining hypotheses all revolve around the product-related variables. Their individual effects on the difference between preference and market share are tested using linear regression. This is done to test whether Price decreases difference (H5) and Volume and Distribution increase difference (H6 and H7).

The results of the different linear regressions are summarised in Table 4. Model 1 follows the form of Equation 12. Model 2 uses the same formula but has the absolute values of difference as its dependent variable. Model 2 makes it possible to address the magnitude of difference on product level.

Looking at Model 1, it is evident that price has a significant positive influence on the difference, while distribution and volume negatively influence the difference. Of the latter two, only distribution holds a significant impact. As the dependent variable is preference share minus market share, this regression shows the influence of variables on the over-and underestimation of a product. As price has a positive parameter, the interpretation of this parameter is that an increase in the standardised price of 1 results in an increase of difference by 0.0046, as discussed in Section 3.1. Preference share increases by 0.0046, on average, compared to market shares when keeping all other factors constant. A higher price thus means an overestimation of preference share compared to market share.

The opposite occurs for distribution and volume, as their parameters are negative. An increase of these variables would, on average, lead to a decrease in the difference of

Table 4: Linear regressions on difference between preference shares (P) and market shares (M)

	Model 1		Model 2	
Dependent variable	$P - M$		$\|P - M\|$	
Constant	0.0015	(0.005)	0.0173***	(0.004)
Price[a]	0.0046*	(0.002)	-0.0037**	(0.002)
Distribution[a]	-0.0133***	(0.002)	0.0066***	(0.002)
Volume[a]	-0.0013	(0.002)	0.0036*	(0.001)
Frequency - weekly	-0.0057	(0.005)	0.0204***	(0.004)
Frequency - monthly	-0.0024	(0.003)	-0.0031	(0.003)
Frequency - quarterly	0.0034	(0.002)	0.0031	(0.002)
Category - Batteries	-0.0049	(0.003)	0.0183***	(0.003)
Category - Beauty	0.0048	(0.003)	-0.0100***	(0.003)
Category - Dairy	0.0079	(0.005)	0.0029	(0.004)
Category - Drinks	0.0114	(0.007)	-0.008	(0.006)
Category - Home	0.0035	(0.003)	-0.0052**	(0.002)
Category - Cigarettes	0.0062	(0.004)	-0.003	(0.003)
Category - Snacks	0.0086	(0.006)	0.007	(0.005)
No. respondents[b]	-0.0014	(0.004)	-0.0094**	(0.003)
No. tasks[b]	0.0054	(0.005)	0.0006	(0.004)
No. SKU's per task[b]	-0.0013	(0.005)	-0.0063	(0.004)
No. SKU's in market[b]	0.0079	(0.006)	-0.0152**	(0.005)
No. competitors[b]	-0.0074	(0.008)	0.0015	(0.007)
R^2	0.042		0.225	
N	1313		1313	

Notes: [a] Price, Distribution and Volume are standardized by dividing value by average value within studies. [b] Variables are normalized using MixMax-normalization.

[*] = .05, [**] = .01, and [***] = .001.

0.0133 and 0.0013 percentage points, respectively, keeping all other variables constant. Therefore, preference share is less overestimated or even understated compared to market shares.

Model 2 has the absolute values of the difference as the dependent variable, addressing the magnitude of difference. There are many more variables of significant influence compared to Model 1. Price, distribution, and volume are of significant influence, where the first has a negative parameter, and the other two have a positive parameter. All three are of significant influence. Increasing price decreases the absolute difference, while increasing distribution or volume increases the absolute difference.

Besides price, distribution, and volume, other variables have a significant influence. The number of respondents also has a negative influence on the magnitude of difference. This was also evident in Table 3. The number of products in the market also has a negative parameter on product level; however, it has a positive parameter when addressing the total magnitude difference in Table 3. This indicates that many SKUs in the market result in a greater total absolute difference, while it also indicates a smaller absolute difference on product level. It might be possible that more products in the market result in a larger number of smaller absolute differences, resulting in a larger sum.

Frequency is included using dummy variables, where daily purchases function as the base level. In the second model, 'weekly' has a significant positive influence, stating that weekly purchases increase the absolute difference compared to daily purchases. This was also concluded in Table 3.

Lastly, product categories were added using dummy variables. Products that do not fall in any of these categories are categorised as 'other' and function as the base level. There are three categories with significant influence on the absolute difference compared to the 'other' category: batteries, beauty products, and home products. Falling in the batteries category results in an increase of absolute difference compared to not falling in a category. Falling in the beauty or home products category decreases the absolute difference compared to not falling in a category.

Hypothesis 5 states that price decreases difference. This hypothesis does not involve the sign of the difference, just the size of the difference. Model 2 shows a negative parameter for price, indicating that price decreases absolute difference, therefore accepting Hypothesis 5. Model 1 shows a positive parameter which means an overestimation of

preference share compared to market share. Combining these findings, it is clear that products with relatively high prices in the market are overestimated by the conjoint; while the largest absolute errors are made for products with relative low prices.

Hypothesis 6 states that a higher Volume leads to a higher preference share compared to market share. Volume has a significant influence in Model 2; however, there is no significant influence in Model 1, leading to a rejection of Hypothesis 6. A higher volume results in a higher absolute difference between preference and market share; however, as its parameter is negative, the value for preference share minus market share decreases. Preference share would thus decrease compared to market share and a higher volume likely results in an underestimation of preference share.

Lastly, Table 4 addresses Hypothesis 7. This hypothesis states that market shares are expected to be lower than the preference shares for products with a lower distribution. Model 2 shows that the absolute difference is significantly increasing for a higher Distribution. Model 1 states a significant decrease caused by either a smaller preference share or greater market share. This means that there is either a greater market share or a smaller preference share. Increasing distribution thus leads to a higher market share compared to preference share. Decreasing distribution thus would result in the opposite. Hypothesis 7 is to be accepted. There is significant evidence that an increase in distribution results in a relatively smaller preference share than market share. Similar to volume, products with a relatively high distribution are likely underestimated in a conjoint analysis, while an increase in distribution leads to a higher absolute error.

4.3 Category effects

The only remaining hypothesis is addressed with interaction effects between the product-specific variables and product categories. The new regression follows Equation 13. It is similar to the linear regression in Table 4, but includes interaction effect and excludes the individual effects of price, volume, and Distribution, as their coefficients are present in all interaction effects. The dummy variables for the products categories remain as explanatory variables to see if these hold a significant difference compared to not having a product category. The resulting regression coefficients are found in Table 5. The R^2 increased compared to Model 1 in Table 4; however, this might be due to the inclusion of many variables. Besides, for both models, the R^2 is very low. This means that there is

only a small percentage of the variance of the difference between preference and market shares explained by either of these models.

Of all control variables, only the dummy variables for purchase frequency are significant; the number of respondents, tasks, SKUs per task, SKUs in the market, and competitors are insignificant. The parameters for frequency show that weekly and monthly purchases significantly increase the difference between preference share and market share, indicating an overestimation of products with a weekly or monthly purchase frequency. The magnitude of this overestimation is, on average, the same for both weekly and monthly purchases, compared to daily purchases. Products that are purchased less frequent have a lower difference compared to daily purchases. This indicates that quarterly purchases are less overestimated compared to daily purchases. As the parameters for weekly and monthly are much greater than quarterly, this also indicates that for quarterly purchases, the preference shares are less overestimated compared to weekly and monthly purchases. This coincides with the expectation that higher frequencies result in a better representation of the conjoint analysis.

The variables of interest, product categories, show a significant effect on the difference for only two product categories when looking at the individual effects. On average, all product categories decrease the difference compared to not having a product category; however, only batteries and dairy products have a significant effect. As these categories are dummy variables, the interpretation is that products in either of these categories have a lower constant of difference of, on average, 0.0568 for batteries and 0.0875 for dairy products.

There is only one significant interaction effect for price and only two for volume. For distribution, almost all of the interaction effects have a significant effect. The interaction effects influence the slope of the linear regression, where the slope is dependent on the product category. Products that fall in the category 'batteries' only use the interactions effects that correspond to that product category: an increase of 1 in price results in an increase of difference of 0.0557. The product category 'batteries' is the only category from which all interaction effects are significant, where price and volume increase and distribution decrease difference. The fact that only 'batteries' has all significant effects is contradicting the results of Model 1 in Table 4 which indicates that product categories might affect difference. The rest of the distribution parameters, except for the interac-

Table 5: Fixed Effects of Product Categories

	Coefficient	(Std Error)
Dependent: preference share - market share		
Intercept	0.0448**	(0.017)
Frequency - weekly	0.0562*	(0.023)
Frequency - monthly	0.0558*	(0.023)
Frequency - quarterly	-0.0404**	(0.013)
Category - Batteries	-0.0568***	(0.012)
Category - Beauty products	-0.0018	(0.009)
Category - Dairy products	-0.0875*	(0.04)
Category - Drinks	-0.0586	(0.044)
Category - Home	0.0182	(0.012)
Category - Cigarettes	-0.0268	(0.017)
Category - Snacks	-0.0317	(0.043)
No. Respondents[a]	-0.0009	(0.003)
No. Tasks[a]	0.0019	(0.004)
No. SKU's per task[a]	-0.0006	(0.004)
No. SKU's in market[a]	0.0043	(0.005)
No. Competitors[a]	-0.004	(0.006)
Category - Other:Price	-0.0071	(0.006)
Category - Batteries:Price	0.0557***	(0.006)
Category - Beauty:Price	0.0025	(0.003)
Category - Dairy:Price	-0.0005	(0.003)
Category - Drinks:Price	-0.0017	(0.011)
Category - Home:Price	0.003	(0.004)
Category - Cigarettes:Price	-0.0036	(0.006)
Category - Snacks:Price	-0.0108	(0.006)
Category - Other:Distribution	-0.0789*	(0.033)
Category - Batteries:Distribution	-0.0375***	(0.01)
Category - Beauty:Distribution	-0.0071	(0.005)
Category - Dairy:Distribution	-0.0096*	(0.004)
Category - Drinks:Distribution	-0.0285***	(0.008)
Category - Home:Distribution	-0.0247*	(0.013)
Category - Cigarettes:Distribution	-0.0101***	(0.002)
Category - Snacks:Distribution	-0.0658***	(0.009)
Category - Other:Volume	-0.0213***	(0.006)
Category - Batteries:Volume	0.032***	(0.006)
Category - Beauty:Volume	0.0023	(0.003)
Category - Dairy:Volume	-0.0031	(0.003)
Category - Drinks:Volume	-0.0141	(0.009)
Category - Home:Volume	0.0007	(0.004)
Category - Cigarettes:Volume	-0.0047	(0.003)
Category - Snacks:Volume	0.0076	(0.013)
R^2	0.124	
Adjusted R^2	0.124	
N	1876	

Notes: [a] Variables are normalized using MixMax-normalization. [b] Price, Distribution and Volume are standardized by dividing value by average value within studies. [*] = .05, [**] = .01, and [***] = .001.

tion with beauty, significantly affect the difference. All parameters are negative, which complies with Model 1 in Table 4; the magnitude is category dependent. The difference between preference share and market share for dairy products are least impacted by distribution, while snacks and products without a distinctive category are impacted the most.

The parameters for the interaction effects with volume with a significant effect on difference are those for the 'other' category and batteries. The sign of the parameters is different, indicating that for batteries, an increase in volume leads to an increase in difference; the opposite occurs for the 'other' category.

Overall it is visible that most categories have a different slope for price, distribution, and volume, compared to Model 1 in Table 4. This indicates that products in product categories have a diverse effect on the difference between preference and market shares. However, only for the interaction effects with distribution are almost all parameters significant; for price and volume, only one or two are significant. As so few of the parameters are significant, Hypothesis 8 cannot be accepted. Thus, there seems to be an effect caused by product categories, but further research is needed to find evidence to accept the hypothesis.

4.4 Predicting market shares

4.4.1 Fitting the predictive models

As previously discussed, the preference shares from a conjoint analysis hold information on the utilities and preferences of consumers. After the computation of the utilities acquired from the survey, the preference shares are computed using Equation 3. These thus hold information that cannot directly be derived from the other variables in the conjoint analysis, such as preference based on appearance, scent, or flavour. Hence, the preference shares can be used as a predictor variable when looking at the actual market shares. Following the methodology, as given in Section 3.6, different black-box models are fitted to predict market shares. The results of the used models are given in Table 6.

For each of the black-box methods, the best model is presented. The Random Forest model was built using 600 individual trees and without a maximum depth, leading to trees with depths of more than 20 nodes in a single branch. The resulting forest scores promising on the different performance measures. The MSE and RMSE have low values,

Table 6: Performance measures of the models on predicting market share

		MSE	RMSE	MAD	MAPE
Random Forest	In sample	0.0001	0.0099	0.0031	0.60
	Out of sample	0.0002	0.0157	0.0072	2.27
Support vector machine	In sample	0.0018	0.0419	0.0356	19.10
	Out of sample	0.0033	0.0577	0.0405	16.34
Neural Network	In sample	0.0011	0.0324	0.0114	1.11
	Out of sample	0.0004	0.0190	0.0092	1.29
Baseline	In sample	0.0006	0.0240	0.0098	2.20
	Out of samplet	0.0003	0.0172	0.0091	3.46

indicating that the model fits the data well. The values for the out-of-sample MSE and RMSE are slightly higher. The out-of-sample MAPE also has a higher value than the in-sample MAPE; however, still low. The MAD is also low, stating that the in-sample predictions are, on average, less than a half percentage point deviating from the actual values. Overall, the performance measures are very low, indicating a good fit and high accuracy. As the out-of-sample measures show a higher value than the in-sample measures, the model might be slightly overfitted.

The Support Vector Regression was fit using different kernels. The model with the most promising results is shown in the table, which has been fit using a polynomial kernel with three degrees. The model performs reasonably well; however, worse than the Random Forest and is therefore neglected for the remainder of this study.

Coming to the last of the black box predictors, the Neural Network, shows promising results. This model is fitted with five hidden layers and a combination of different activation functions. Further, 50 epochs are used, and the used batch size is 150. The RMSE of both samples is low; however, slightly higher compared to the Random Forest. Interestingly, the in-sample RMSE shows worse performance than the out-of-sample RMSE, whereas the MAPE shows the opposite. The MAD is also slightly higher in the out-of-sample set. The MAD is also slightly worse compared to the Random Forest for both in-sample and out-of-sample sets. The MAD, MSE, and RMSE might indicate a

form of underfitting, whereas the MAPE indicates overfitting.

Besides the black box predictors, a baseline is added. This baseline shows the performance of predicting the market share by just looking at the preference shares. The performance measures for the baseline are relatively low; however, the Random Forest and Neural Network perform even better.

The Random Forest is selected as best-performing and used to predict with a new data set considering all performance measures. Random Forest is preferred over the NN here, as the RMSE and MAD are slightly lower, while the MAPE is very good.

Figure 7 shows the variable importance of all variables when predicting market share, using the Random Forest. As expected, Preference Share has the most impact on the prediction. Market variables and study-related variables have the most negligible impact. However, these also impact the Preference Share, which might make for (multi)collinearity within the model. A benefit of the RF is the decorrelation technique that takes (multi)collinearity into account. This technique could very well be why the market and study-related variables have none or little importance in the prediction. Multicollinearity is therefore not seen as a problem in this situation.

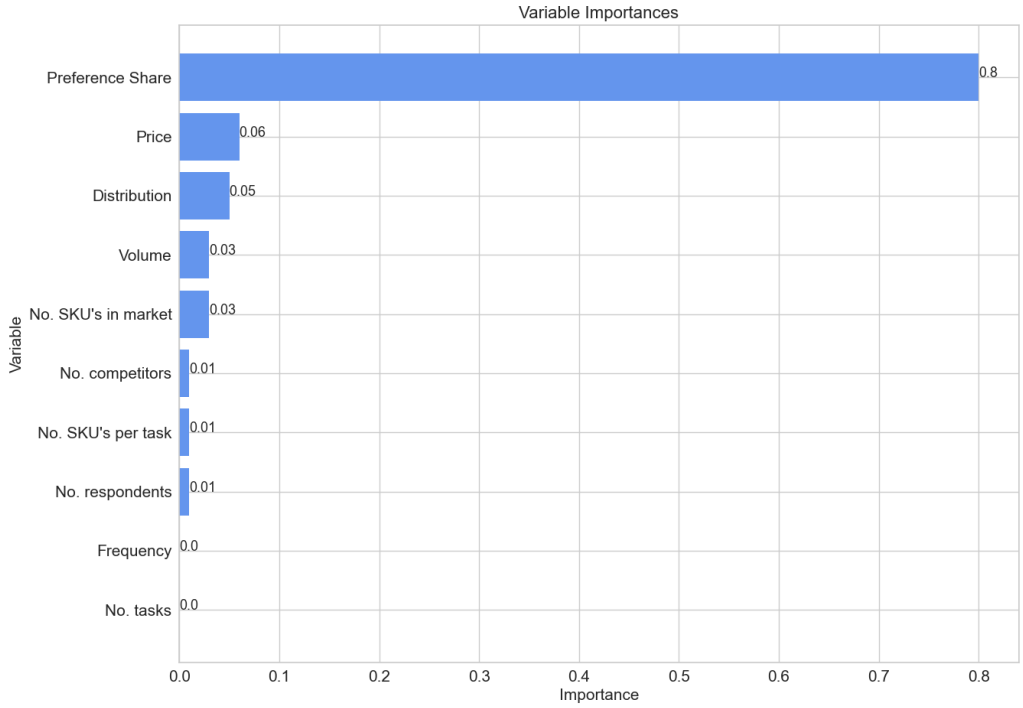

Figure 7: Variable importance of predictor variables on prediction of Market shares, using Random Forest model

4.4.2 New data introduction

New data is introduced to evaluate the performance of the black-box model. The new observations are acquired through new studies. In total, there are 481 new observations introduced, yielding the results in Table 7. In calculating market shares' prediction, all three of the black box models were heavily influenced by a single prediction with an enormous absolute percentage error. The predictions for all observations had low values; one observation had an absolute percentage error of over 6000 in all of the models. This single outlier has been neglected. The performance measures including this outlier are shown in Table 8, Appendix C.

Table 7: Performance measures on new data

	Baseline	Random Forest	SVM	Neural Network
MSE	0.0006	0.0007	0.0427	0.0009
RMSE	0.0241	0.0262	0.2066	0.0303
MAD	0.0077	0.0102	0.1647	0.0180
MAPE (in %)	3.49	2.85	48.71	7.17

Immediately it can be seen that the SVM models perform worse on the new data. All performance measures are worse for the SVM, but the MAPE increased tremendously. The Random Forest has slightly higher values for all performance measures, compared to the out of sample performance in Table 6. For the Neural Network, the performance measures do show a decline in performance, but not extensively. The baseline, where preference share is used as a predictor of market share, performs similar to Table 6. Even though the black box predictors perform slightly worse, the accuracy is still very high, making it a handy tool in predicting the actual market shares based on the preference shares obtained through conjoint analysis.

5 Conclusion and discussion

Conjoint analysis is used widely to give information about the market and its respondents. By estimating the preference of respondents, and subsequently, their utilities, the share of products on the market can be estimated. Even though this method should give

a good representation of the real world, more often than not, the gap between conjoint analysis and market is more prominent than expected. This study focuses on the external validity of this method by looking at the influence of study, market, and product-related variables on the difference between preference and market share. Throughout this study, the effects of different variables on the external validity of conjoint analysis are assessed. A high external validity is essential as it ensures good generalization of the results from the conjoint analysis to or across different people, timing, or settings. A low external validity results in a difference between preference shares and market shares, leading to marketing decisions that are not optimal. Therefore, conducting research on this subject is of importance. Preference shares that result from the conjoint analysis are often not too similar to actual market shares, which makes generalization difficult, leading to sub-optimal situations (Feit et al., 2010).

The difference between actual market shares and preference shares is investigated in this thesis by looking at product, study, and market-related variables. The data used for the conducted research includes 1,894 different products from 33 different conjoint studies in the FMCG branch. First, the influence of study and market-related variables are investigated using shrinkage regression techniques as multicollinearity is present. With the magnitude of the absolute differences as the dependent variable, the resulting models give insights into the influence of the number of respondents, tasks, options per task, products on the market, competition, and frequency. The number of options per task and the number of competitors do not significantly influence the magnitude of absolute difference, rejecting Hypothesis 2. The other variables do have an impact. The number of respondents decreases the magnitude of difference, which leads to accepting Hypothesis 3. This supports the general idea that states that a bigger sample size is desired (Malhotra et al., 2017).

The number of tasks was expected to decrease the magnitude of difference, as the performance of the respondents does not necessarily decrease by increasing the number of tasks (Bansak et al., 2018 Johnson and Orme, 1996). The resulting Lasso models suggest the opposite, supporting theory that suggests that there is an information overload when increasing the number of tasks (Louviere and Timmermans, 1992). When excluding an outlier found in the data, a linear regression shows a negative slope in Figure 12. However, this observation appears relevant and is therefore not excluded from the data. Including

the outlier, Figure 6 shows a slightly increasing slope. This complies with the positive parameter found in the Lasso regression.

More in-depth analysis, Figure 6, showed that this is the result of an optimum number of tasks of fourteen when interpolating using median; or roughly ten tasks when interpolating using mean, minimum or maximum. After this optimum, an increase in magnitude occurs. The interpolated relations for mean, minimum, and maximum also show a decrease in magnitude between twelve and fifteen tasks, where the median shows a minimum at fourteen tasks. Hypothesis 1 is not accepted, given the findings mentioned above. Even though the hypothesis is not accepted, it is advised to show respondents fourteen tasks in a survey.

A limitation at play in this situation is that the conclusions are based on a small sample. As there are only 33 studies, there are only 29 observations that can be used in training this model while having only four observations as the out-of-sample group. It can also be seen in Figure 6 that there are relatively many studies with twelve tasks. Thus, the distribution of the number of tasks could also be interfering with the results. Future research should look into this relation using bigger sample size, with more variation in the number of tasks used.

After addressing the study-related variables, the context-dependency of difference was addressed. First, Hypothesis 4 used purchase frequency to investigate its impact on the difference. Hypothesis 8 addresses the effect of product categories on difference but uses product specific information as price, volume, and distribution, and is discussed after the conclusions regarding the product-specific variables.

Products that have a higher purchase frequency were expected to negatively influence the magnitude of difference due to the level of involvement and the theory of hedonism (Szmigin and Piacentini, 2018). Table 3 shows a decrease for monthly and quarterly products, compared to daily-bought products; however, it shows much higher positive values for the parameters of the weekly-bought products, leading to the rejecting of Hypothesis 4.

Besides having the limitation of a small sample size, the frequencies used are based on purchase frequency. An assumption is made that the frequencies are equal throughout the whole study, neglecting the impact of volume size on purchase frequency, which might hold relevant information. Moreover, all products are in the FMCG branch, making various

products having the same frequencies. Recommendations for future research would be to go over the specification of this variable and try to perform the same analysis where usage frequency would be used as the indicator for frequency instead of purchase frequency.

Based on this regression and the significant outcomes, market researchers should increase the number of respondents to balance the number of products on the market. Given the Lasso model, which performs best, the coefficients for respondents and products on the market almost balance each other out. Using the interpretation as given in Section 4, for approximately every two products in a market, at least 39 respondents should be included to balance out the influence on the magnitude, ceteris paribus. Even though Hypothesis 4 was not accepted, increasing the number of respondents for more frequently purchased products would be more necessary than for less-frequently purchased products.

Hypotheses 5 through 7 focus on the individual product's variables, price, distribution, and volume. Conducting research for these hypotheses has been done with linear regressions, as can be seen in Section 4.2. Two models are fitted with difference per product is used as the dependent variable. Price, distribution, and volume are included in these regressions as predictor variables. All other variables are included as control variables. This includes the variables regarding the design of the conjoint survey, variables regarding market information, purchase frequency, and product category. As discussed multiple times, it is likely that there are other variables not included in this model that explain the variance in the difference between preference and market shares, leading to a low R^2. Even though there is not much variance explained, there is a significant influence of the predictor variables on the differences. For Models 1 and 2, price and distribution are significant; volume is only significant in Model 2.

Similar to the results of Sichtmann et al. (2011), price has a significant influence in decreasing the difference between preference share and market share, resulting in the acceptance of Hypothesis 5. This contradicts the results found in the meta-analysis conducted by Murphy et al. (2005), where external validity in willingness-to-pay decreases with the increase of monetary value. A higher price thus results in an overestimation of the preference shares, whereas the greatest absolute errors are found for products with relatively low prices.

It was expected that volume would increase the difference, as packages with a greater volume are chosen more often in surveys than in the actual market. However, volume only

gives a significant coefficient in estimating the magnitude of difference, which is not robust enough to accept Hypothesis 6. The lack of significance could be a result of the small sample size.

The results regarding Hypothesis 7 are in line with expectation and theory. Products that are not widely available, thus having a low distribution, are expected to be over-represented in conjoint analysis results: preference shares are relatively higher than market shares. Alternatively, products with high distribution are expected to be under-represented (Chandukala et al., 2011; Natter and Feurstein, 2002). As can be seen in the table, the coefficient for Model 1 is negative and positive for Model 2. An increase in Distribution thus leads to a smaller positive or larger negative difference. This means that the preference share is getting smaller, or the market share is increasing. As market shares are not affected by conjoint analysis, preference shares are getting smaller than the market share: an under-representation of this product in the eventual preference shares. Alternatively, a decrease in Distribution leads to an increase in difference; hence an increase in preference share compared to market share, meaning an over-representation of this product. Hypothesis 7 is thereby accepted.

A recommendation for future research would be to investigate this phenomenon more closely to find a pattern in this over-and under-representation. A deeper investigation could lead to a better correction for the calculation of preference shares. Another recommendation is to use the regression results from Model 1 to predict market shares.

Following the intuition of several academics, context-dependent behaviour has been addressed in Hypothesis 8 (Allenby et al., 2005; Hainmueller et al., 2014; Sichtmann et al., 2011). Reactions to variables are captured by the parameter of interaction effects between price, distribution, and volume, and product categories. Only one category holds significant effects on the difference for all interaction effects and its individual effect of the seven product categories. The parameters of the interaction effect differ between product categories; however, fourteen of the 24 parameters are not significant. Almost all of the interaction effects with distributions are significantly decreasing difference, as was also found in Table 4. Due to the lack of significant evidence, Hypothesis 8 cannot be accepted. There appears to be an influence of product category which should be investigated further, by looking at more studies.

As proposed before, performing a similar analysis should be performed using a bigger

sample size. Also, looking into different categories could give insights. Products that fall in the FMCG category might behave more similarly to each other than technological or healthcare products.

Lastly, the black box predictors are used to estimate market shares given all other variables. The Random Forest and Neural Network perform very well on the training and validating set; however, they seem to overfit the new data. Predicting the market shares of new data showed increased MSE, RMSE, MAD, and MAPE. This increase was minimal for the Random Forest, while it was more remarkable for the Support Vector Machine and Neural Network. This indicates that the Random Forest model is better suited to predicting the market shares than the other models. Even though the Random Forest works well, it is only slightly better compared to the baseline. This could result from the fact that the Random Forest predicts the market share primarily using preference shares. Including more control variables could help to improve the predictions.

Fitting the model on more data would be a recommendation for the future, as well as including studies from different branches. However, the black-box models show significant room for improvement when it comes to predicting actual market shares, as the performance measures are similar to the baseline. Given the variable importance plot in Figure 7, it can be seen that the preference shares do not give enough insides of the actual market. Interestingly, frequency and the number of tasks, as addressed in Hypotheses 1 and 4, do not influence market share prediction when preference shares are included. This might be the result of the decorrelation technique of the random forest.

An overall limitation in the previously addressed analyses is the sample size and the studies' original (sub)categories. It would be interesting for future research to focus on different product categories, such as more hedonic products as televisions or mobile phones, and assess the performance of conjoint analysis for these products. As found in this study, relatively expensive products in a study result in smaller differences between preference and market shares. Conducting research on the generalization of this founding to product pricing, in general, would possibly make for compelling insights. It would also be interesting to see if other findings could be generalized to other categories as well.

Another substantial limitation that needs to be addressed is the following: the data set is randomly divided into two. However, this split should be reconsidered in future research. The main objective of this study was to find the most influential predictors

and investigate their impact. However, when a study has the main objective for training the best predictor model, this split causes issues. The single observations all are part of different simulators, where the market shares sum up to 100% and all differences sum up to 0. The observations are thus dependent on the values of the predictor variables and the other products in that study. By splitting randomly, this is neglected, and the training and validation set are somewhat influenced by each other. It could very well happen that half of a study is present in the training set, whereas the other half is present in the validation set, while their predictions should sum up to 1.

Neural Networks can be created with the SoftMax activation function in the last layer. This layer creates a distribution of the predictions, such that the sum of predictions results in 1. However, ways for modelling this layer so that all observations of one study sum up to 1 were not found. Future, more in-depth and econometric research should focus on this problem, such that the different markets are to be mapped more accurately.

References

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs neurons: Comparison between random forest and ann for high-resolution prediction of building energy consumption. *Energy and Buildings*, *147*, 77–89.

Allenby, G., Fennell, G., Huber, J., Eagle, T., Gilbride, T., Horsky, D., Kim, J., Lenk, P., Johnson, R., Ofek, E., Orme, B., Otter, T., & Walker, J. (2005). Adjusting choice models to better predict market behavior. *Marketing Letters*, *16*(3-4), 197–208. https://doi.org/10.1007/s11002-005-5885-1

Awad, M., & Khanna, R. (2015). *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. Springer nature.

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2018). The number of choice tasks and survey satisficing in conjoint experiments. *Political Analysis*, *26*(1), 112–119.

Beck, M. J., Fifer, S., & Rose, J. M. (2016). Can you ever be certain? reducing hypothetical bias in stated choice experiments via respondent reported choice certainty. *Transportation Research Part B: Methodological*, *89*, 149–167.

Bengfort, B., Bilbro, R., Danielsen, N., Gray, L., McIntyre, K., Roman, P., Poh, Z., et al. (2018, November 14). *Yellowbrick* (Version 0.9.1). https://doi.org/10.5281/zenodo.1206264

Boesch, I., Schwaninger, M., Weber, M., & Scholz, R. W. (2013). Enhancing validity and reliability through feedback-driven exploration: A study in the context of conjoint analysis. *Systemic Practice and Action Research*, *26*(3), 217–238.

Bonhomme, S., & Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, *83*(3), 1147–1184.

Bordt, S., Farbmacher, H., & Kögel, H. (2019). *Estimating grouped patterns of heterogeneity in repeated public goods experiments* (tech. rep.). Working paper.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Bremer, L., Heitmann, M., & Schreiner, T. F. (2017). When and how to infer heuristic consideration set rules of consumers. *International Journal of Research in Marketing*, *34*(2), 516–535. https://doi.org/10.1016/j.ijresmar.2016.10.001

Chandukala, S. R., Edwards, Y. D., & Allenby, G. M. (2011). Identifying Unmet Demand. *Marketing Science*, *30*(1), 61–73. https://doi.org/10.1287/mksc

Chollet, F. et al. (2015). Keras. https://github.com/keras-team/keras

Cunningham, C. E., Deal, K., & Chen, Y. (2010). Adaptive choice-based conjoint analysis. *The Patient: Patient-Centered Outcomes Research*, *3*(4), 257–273.

Darmon, R. Y., & Rouziès, D. (1999). Internal validity of conjoint analysis under alternative measurement procedures. *Journal of Business Research*, *46*(1), 67–81.

Feit, E. M., Beltramo, M. A., & Feinberg, F. (2010). Reality Check: Combining Survey and Market Data to Estimate Choice Models. *Management Science*, *56*(5), 785–800. https://doi.org/10.2139/ssrn.1154222

Fifer, S., Rose, J., & Greaves, S. (2014). Hypothetical bias in stated choice experiments: Is it a problem? and if so, how do we deal with it? *Transportation research part A: policy and practice*, *61*, 164–177.

Fitzner, K. (2007). Reliability and Validity A Quick Review. *The Diabetes Educator*, *33*(5), 775–780. https://doi.org/10.1177/0145721707308172

Green, P. E., Krieger, A. M., & Wind, Y. (2001). Thirty Years of Conjoint Analysis: Reflections and Prospects. *Interfaces*, *31*(3), 56–73. https://about.jstor.org/terms

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of consumer research*, *5*(2), 103–123.

Guadagni, P. M., & Little, J. D. (2008). A logit model of brand choice calibrated on scanner data. *Marketing Science*, *27*(1), 29–48.

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, *22*(1), 1–30. https://doi.org/10.1093/pan/mpt024

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/

Johnson, R. M., & Orme, B. K. (1996). How many questions should you ask in choice-based conjoint studies. *Art Forum, Beaver Creek*, 1–23.

Kroes, E. P., & Sheldon, R. J. (1988). Stated preference methods: An introduction. *Journal of transport economics and policy*, 11–25.

La Cuesta, B., Egami, N., & Imai, K. (2019). Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution. *Political Analysis*, 1–27.

Lantz, B. (2013). *Machine learning with r*. Packt publishing ltd.

Laurent, G. (2000). Improving the external validity of marketing models: A plea for more qualitative input. *International Journal of Research in Marketing*, *17*(2-3), 177–182. https://doi.org/10.1016/s0167-8116(00)00020-3

Liu, Q., & Tang, Y. (2015). Construction of Heterogeneous Conjoint Choice Designs: A New Approach. *Marketing Science*, *34*(3), 346–366. https://doi.org/10.1287/mksc.2014.0897

Louviere, J. (1988). Conjoint analysis modelling of stated preferences: a review of theory, methods, recent developments and external validity. *Journal of Transport Economics and Policy*, *22*(1), 93–119. https://www.researchgate.net/publication/235356400

Louviere, J., & Timmermans, H. J. (1992). Testing the external validity of hierarchical conjoint analysis models of recreational destination choice. *Leisure Sciences*, *14*(3), 179–194. https://doi.org/10.1080/01490409209513167

Lucas, J. W. (2003). Theory-testing, generalization, and the problem of external validity. *Sociological Theory*, *21*(3), 236–253.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27. https://doi.org/10.1016/0022-2496(64)90015-X

Malhotra, N. K., Nunan, D., & Birks, D. F. (2017). *Marketing research: An applied approach*. Pearson Education Limited.

Mas-Colell, A., Whinston, M. D., Green, J. R., et al. (1995). *Microeconomic theory* (Vol. 1). Oxford university press New York.

Mingers, J. (1989). An empirical comparison of selection measures for decision-tree induction. *Machine learning*, *3*(4), 319–342.

Moore, W. L. (1980). Levels of aggregation in conjoint analysis: An empirical comparison. *Journal of Marketing Research*, *17*(4), 516–523.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement*, *7*(2), 191–205.

Murphy, J. J., Allen, P. G., Stevens, T. H., & Weatherhead, D. (2005). A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics*, *30*(3), 313–325.

Natter, M., & Feurstein, M. (2002). Real world performance of choice-based conjoint models. *European Journal of Operational Research*, *137*(2), 448–458. https://doi.org/10.1016/S0377-2217(01)00147-3

Netzer, O., & Srinivasan, V. (2011). Adaptive self-explication of multiattribute preferences. *Journal of Marketing Research*, *48*(1), 140–156. https://doi.org/10.1509/jmkr.48.1.140

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement*, *22*(4), 287–293.

Orme, B. (2004). Managerial overview of conjoint analysis. *Getting Started with Conjoint Analysis*.

Orme, B. (2009). Maxdiff analysis: Simple counting, individual-level logit, and hb. *Sawtooth Software*.

Orme, B. K., & Heft, M. (1999). Predicting actual sales with CBC: How capturing heterogeneity improves results. *Sawtooth Software Conference Proceedings*, 183–199.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rogers, G., & Soopramanien, D. (2009). The Truth is Out There! How External Validity Can Lead to Better Marketing Decisions. *International Journal of Market Research*, *51*(2), 1–14. https://doi.org/10.1177/147078530905100216

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Sichtmann, C., Wilken, R., & Diamantopoulos, A. (2011). Estimating Willingness-to-pay with Choice-based Conjoint Analysis - Can Consumer Characteristics Explain Variations in Accuracy? *British Journal of Management*, *22*(4), 628–645. https://doi.org/10.1111/j.1467-8551.2010.00696.x

Stock, J., & Watson, M. (2014). *Introduction to Econometrics, Global Edition*. Pearson Education Limited.

Szmigin, I., & Piacentini, M. (2018). *Consumer behaviour*. Oxford University Press.

Theil, H. (1969). A multinomial extension of the linear logit model. *International economic review*, *10*(3), 251–259.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Vieira, S., Pinaya, W., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, *74*. https://doi.org/10.1016/j.neubiorev.2017.01.002

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Yang, L., Toubia, O., & de Jong, M. G. (2018). Attention, Information Processing, and Choice in Incentive-Aligned Choice Experiments. *Journal of Marketing Research*, *55*(6), 783–800. https://doi.org/10.1177/0022243718817004

A Explanation of Black Box-methods

A.1 Random Forests

Random Forests are computed by a machine-learning algorithm that builds on combining decision trees. The resulting forest is an ensemble C of $T_1(X), T_2(X), ..., T_C(X)$ trees (Ahmad et al., 2017). To understand a random forest, first, the basics of decision trees are given. A decision tree is built by splitting the data into partitions or branches, by making logical and binary decisions on features, looking similar to a flowchart. Starting in the root node, data flows through the tree, following the branches that it is pushed into by the decision nodes. Eventually, the branches end in leaf nodes, giving the classification or value of the observation. The splits hold decisions on a feature and a certain level. The split that is to be included in a tree is selected by trying to find the purest outcome after splitting. This means that the algorithm looks for a split that has the least number of classes per branch after the split; the smaller the number of classes in a branch, the purer the branch (Lantz, 2013). This purity has different measures, from which the Gini index is used often, and by default in scikit-learn (Pedregosa et al., 2011). The Gini index is a measure of overall impurity after a split. It uses the probabilities of each class (p_i) and the probability of not being in that class (p_j), then sums the product of the two for each resulting partition. The general function for the Gini index is then

$$\text{Gini} = \sum \sum_{j \neq i} p_i p_j = 1 - \sum p_i^2. \tag{17}$$

A higher Gini index means a higher level of impurity. A split resulting in a lower level of impurity, thus lower Gini, is then preferred (Mingers, 1989).

A random forest is created such that there is as little correlation between the individual trees, as possible. Using this *decorrelating* technique ensures a more reliable and robust model compared to just combining multiple trees that use all variables (Breiman, 2001). To do so, a tree uses a subset m of all n explanatory variables. Often $m = p/3$ is used in a regression, where $m = \sqrt{p}$ is used by default for classification problems. Besides using m, not all N training samples are to be used in every tree, to ensure no bias to outlier data. Another hyperparameter to tune is the number of trees used for the growing of the forest (James et al., 2013). Scikit-learn uses a hundred trees by default, which can be altered manually. To have a broad number of trees, in this study 600 trees are created

(Pedregosa et al., 2011).

The interpretation of a random forest is more difficult compared to a single decision tree, as there is not just one tree used. The Gini index helps with a part of the interpretation of the random forest, as it is possible to compute the variable importance, using this index. As discussed, Gini shows how well a split performs. By looking at the Gini index of splits, the importance of variables can be conducted. Variables that cause a lot of purity in multiple trees are more important in prediction, compared to variables that do not purify the partitions much. This way, variables can be ranked to determine which variables have the most impact.

A.2 Support Vector Machines

Another method that is used is the Support Vector Machine (SVM). It can be used for both classification and regression, where the second one behaves as a generalized version of the classifier, meaning that a Support Vector Classifier (SVC) determines a finite set of classes, while the Support Vector Regressor (SVR) uses the SVC technique, but estimates a continuous-valued output. To give a better view of how the SVR works, the SVC is explained shortly.

This method creates separating hyperplanes in a p-dimensional space, to divide observations into classes. In p dimensions, a hyperplane is a subspace of dimension $p-1$. In a two-dimensional space, a hyperplane thus is a line. The optimal hyperplane is determined by finding the *Maximum Margin Hyperplane*: a separating plane where the distance between observations of the different classes, called the margin, is as big as possible. When this distance is at its maximum, the different classes are most distinct. The observations that determine the margin are called the support vectors and these observations are used for future prediction. Observations that lie between the support vectors while training, are treated as outliers. To minimize this occurrence, a cost parameter is included that penalizes these outliers (James et al., 2013, Lantz, 2013).

To go from SVC to SVR, an ε(error)-insensitive region around the found function is introduced, called the ε-tube or error-tube. By doing so, the optimization problem has slightly changed. While an SVC just splits the data into partitions, the SVR tries to find the smallest tube possible with as many, preferably all, observations inside the tube. Optimizing this indicates that many of the observations are as close to the found function

as possible, making for a lower error. Observations that lie very far from the function, result in a high error. A schematic of a SVR is given in Figure 8, where the ε-tube is shown around the resulting function in a two-dimensional space. This results in the following functions, where Equation 18 is used in a one-dimensional space and Equation 19 for multidimensional data:

$$f(x) = \langle w, x \rangle + b = \sum_{j=1}^{M} w_j, x_j + b, \tag{18}$$

$$f(x) = \begin{bmatrix} w \\ b \end{bmatrix}^T \begin{bmatrix} x \\ 1 \end{bmatrix} = w^T x + b. \tag{19}$$

Here w holds the weights given to observations x, and b equals a bias (Awad and Khanna, 2015).

Lastly, the SVM makes use of the so-called *kernel-trick*, which makes it possible to split data when there seems no split possible in the number of dimensions used. This is a trick that is applied when, for instance, the split between two groups is circular-shaped in a flat space. The kernel $K(x_i x_{i'})$, which can be defined using multiple functions, is a generalization of the inner product that is used in the computation of the SVM, as seen in Equation 18. The function holds similarities between observations. Using a linear kernel is essentially the same as using the distance, described above. By using a non-linear kernel, the SVM can be approached in an enlarged feature space, without having to do

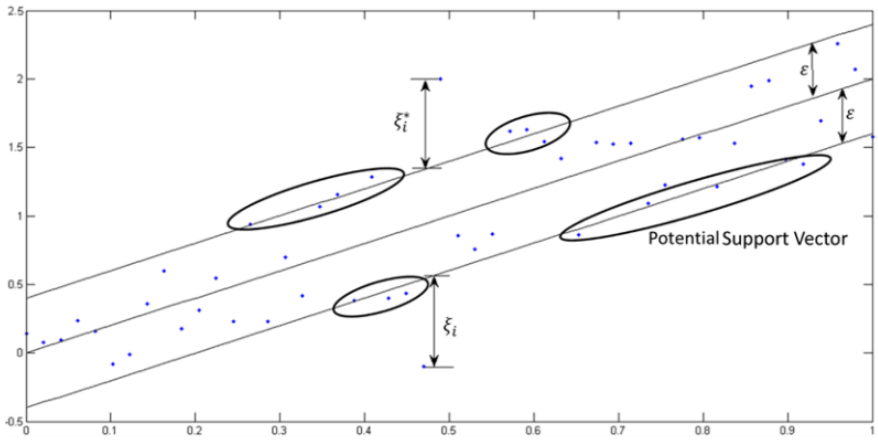

Figure 8: Schematic of a Support Vector Regression

Note: This figure was created to show what a Support Vector Regression model looks like and what the error-tube and support vector look like, surrounding the model. From Awad, M., & Khanna, R. (2015). *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. Springer nature

the computations explicitly in that space. The polynomial kernel is used often, which uses the distance between observations to a pre-determined polynomial degree (James et al., 2013).

A.3 Neural Networks

Lastly, Neural Networks (NN) are addressed. This black-box model works similar to the working of neurons inside the body, imitating its networks. The basic working of a neural network is that explanatory variables are injected in the input layer, pass through (multiple) hidden layers, to eventually end up in the output layer, resulting in a value for y. The different neurons that the information passes through, are all connected with different weights, similar to dendrites. The neurons all have activation functions, which make it possible for these neurons to process the input information and let it pass through to the next neuron (Awad and Khanna, 2015).

Figure 10 shows a schematic of the working of a Neural Network with two hidden layers. The top part of the figure shows the weighted connections and activation functions. The lower part of the figure shows how the neurons are connected with the weighted dendrites. For each of the layers, an activation function is set, as shown in y_j, y_k and y_l. Multiple activation functions can be used and all work slightly different in processing the information. First, there is the **unit step**, which has an output of 1 if the input exceeds a certain threshold, or 0 if it does not. Then there is the commonly used **Sigmoid** activation function, which has an output of $f(x) = 1/(1 + e^{-x})$. Further, there are also a linear, saturated linear, (hyperbolic) tangent and Gaussian function (Lantz, 2013).

Creating the Neural Network is done using the Keras-package in Python, which works using a sequential model and often uses the **ReLU** activation function; the Rectified Linear Unit activation layer. The ReLU-function returns

$$f(x) = \begin{cases} z & \text{if } x \geq z; \\ x & \text{if threshold} \leq x < z; \\ r \times (x - \text{threshold}) & \text{otherwise,} \end{cases} \tag{20}$$

where z is a pre-determined maximum value to be returned, 'threshold' a pre-determined threshold and r a pre-determined negative slope. By default, the first is set to no maximum and the other two are set to 0 (Chollet et al., 2015).

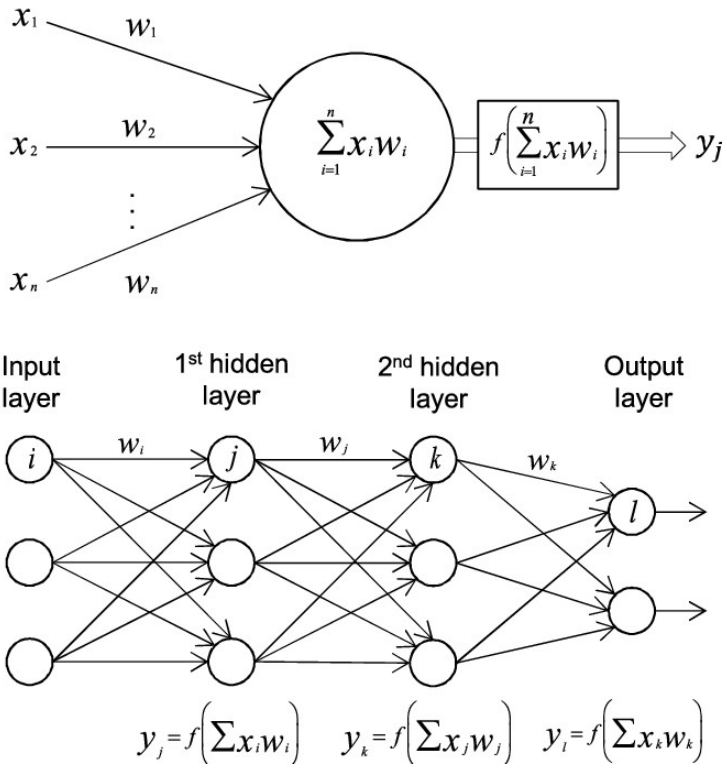

Figure 10: Schematic of an Artificial Neural Network

Note: This figure was created to show a schematic of the workings of an Neural Network. The top figure shows that the weights and activation functions that are assigned to each neuron, before passing through to the next layer, either hidden or output. From Vieira, S., Pinaya, W., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, *74*. https://doi.org/10.1016/j.neubiorev.2017.01.002.

Beforehand, there are two hyper parameters to be determined for the creation of the NN. First, there is the number of *epochs*. This is the number of times all the observations in the sample are passed through the Neural Network when training. The weights as shown in Figure 10 are set randomly at first, and are adjusted during training. The higher the number of epochs, the more the model will be trained and the less random the weight will be. This process is done using Gradient Descent, where the weight slowly changes until it has found a local optimum. This optimum is reached when the highest accuracy or the lowest error is established. Here, MSE is used as loss-function, meaning the model searches for the smallest MSE possible.

Secondly, the batch size is determined. Batch size holds a pre-determined value that gives the number of observations used in one run through the Network. It uses smaller subsets of the data to train the model, which has a faster computational time compared to training with the total sample. Training a Neural Network with a sample of 1000

observations, a batch size of 100 and 50 epochs, means that the first 100 observations are passed through the Neural Network, then the second 100 observations pass through hand so on. This process is then repeated 50 times over (Chollet et al., 2015).

B Relationship between tasks and magnitude

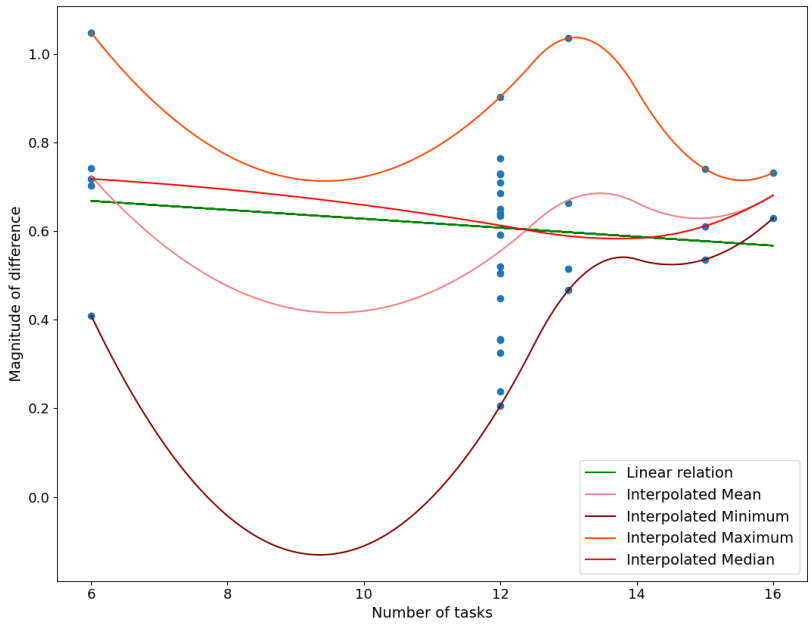

Figure 12: Relationship between number of tasks and magnitude - excluding outlier

C Performance measure on new data - including outlier

Table 8: Performance measures on new data - including outlier

	Baseline	Random Forest	SVM	Neural Network
MSE	0.0006	0.0007	0.0036	0.0009
RMSE	0.0241	0.0268	0.0598	0.0303
MAD	0.0077	0.0126	0.0373	0.0180
MAPE (in %)	3.49	21.08	64.20	12.72

D Neural Network layers

Table 9: Neural Network Design

Layer	Output	No. parameters	Activation
Dense	8	96	ReLu
Dense	2066	18594	ReLu
Dense	1200	2480400	Sigmoid
Dense	500	600500	ReLu
Dense	1	501	Linear