A comprehensive study of catch-'em-all measures for evaluating consumer attitudes regarding product features

ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Bachelor Thesis [Economie en Bedrijfseconomie]

A comprehensive study of catch-'em-all measures for evaluating consumer attitudes regarding product features

Name student: Dennis Brouwer
Student ID number: 387572

Supervisor: Dr. J.P.M. (Jan) Heufer
Co-reader: Drs. B. (Benjamin) Tereick

Date final version: 26/07/2018

Abstract

The purpose of this paper is to examine the effects of rating scale design on the results produced by different designs. The main instigator of this thesis was the result of past research, which demonstrated an apparent bias in the results, which was potentially the result of the applied rating scale system.

Therefore, two different rating scale systems were compared, the first being a replication of the rating scale system from the previous research (a combination of two seperate rating scales) and the second being a specifically designed potential improvement over the first. Additionally, the second scale was augmented with a mechanism called Choice-Matching, that aims to induce (more) truthful responses from participants, forming a third and final variant.

To test whether the first variant is an improvement over the other *and* whether Choice-Matching augmentation proves to be an improvement, several different statistical and visual measures where applied, including the Shapiro-Wilk test on normality, Q-Q plots, histograms, kernel density plots, resampling, Pearson's Chi-Squared test on Goodness of Fit and the independent samples test on proportions.

Consequentially, the second rating scale was found to be an improvement over the first. The Choice-Matching augmented scale was burdened with a high drop-off rating, resulting in a significantly lower sample size of finished respondents, compared to the other two variants. Additionally, the drop-off was found to be selective, leading to a significant sample selection bias and a base of respondents that is not representative of the potential consumer base. Therefore, the second rating scale system, unaugmented by the Choice-Matching mechanism, was found to be the preferred variant among the three that were evaluated in this study.

Acknowledgements

I would like to use this page to thank some of the people that supported me during my work on this thesis and throughout my bachelor.

First, I want to thank Dana, Harmanan and Justus, who worked together with me on the group project that instigated the inception of this thesis. Discussions that I had with them during that project and during the time that I spent working on this thesis were very helpful for establishing this thesis.

Next, I want to express my gratitude to my bachelor thesis supervisor, Dr. Jan Heufer, for supervising my thesis over the past few months. I especially want to thank him for being very receptive to any questions, feedback or points of discussion I had during the research period. He was always very quick to adequately respond to any of my comments, which greatly helped to keep up the pace, solve any issues as soon as possible and thoroughly discuss and decide regarding any points of discussion or unclarity. Additionally, I want to thank Drs. Benjamin Tereick for agreeing to be the co-reader for my bachelor thesis.

I also want to thank all those thousands who participated in the exploratory research and the experiment itself. As most of the participants were contacted through local Pokémon GO Whatsapp groups, I would hereby like to thank all those Pokémon enthusiasts and the Pokémon GO community. In particular, I would like to thank Jamie, Kim and Maureen. I could always reach out to them to discuss anything related to the Pokémon games. As this was very relevant to the research conducted for this thesis, these discussions were often essential for making important decisions that involved specific game mechanics.

Finally, I want to thank my parents, Ron and Conny. They have always supported me, both during my thesis and through the whole bachelor. I would undoubtedly never have been able to achieve what I have over the last few years and in this research without their support, encouragement and advice and I would like to sincerely thank them for that.

Dennis Brouwer, 26/07/2018, Spijkenisse

Table of contents

1 Introduction

One of the most important concepts within marketing is customer satisfaction. For the evaluation of customer response to certain characteristics of products, market research is essential. A lot of vital marketing questions can be answered using marketing analytics and more and more companies are adopting it (Moorman, 2018). For example, let us consider The Pokémon Company, which takes care of the marketing of the Pokémon brand and its games. A lot of potential questions may come up when discussing which features should be included in a new game. During game development, The Pokémon Company might want to send out a survey to potential customers of the new game to analyze their preferences. Furthermore, when the game is close to being released, the company might want to start a global marketing campaign to raise awareness about the game coming out soon.

When the game is almost finished, they have video materials to show to their audience and need to know which features resonate best with them. In that way, they know which features they need to emphasize in their marketing ads. The main question that remains for marketing analysts is how exactly to measure the opinions of their potenial customers. How do they evaluate if features are met with great enthusiasm or if they are generally disliked? What is the feature that their fans care about most and should definitely be in the new game? To answer these important questions, it is essential to make the right decision regarding the to-be-used measurement tool. Choosing an unsuited or faulty system might lead the company into including a feature that their fans do not like, which could have major consequences for the sales of this game and future installments.

To evaluate and predict how people judge certain characteristics of objects or matters, several types of rating scales can be applied. The aggregated scores for the different characteristics that are produced by the rating scale can be compared. They may give an indication of the relative preferences of the research subjects (those participating in the survey or experiment) regarding these characteristics. Comparing the scores for the different characteristics within one rating scale system might not be that controversial. However, it is much more difficult to argue for basing actual decision-making solely on these aggregate ratings. This is not only the case because aggregate ratings might potentially dilute highly varying opinions, but even more so because the different rating scales produce non-linearly comparable outcomes. One rating scale might be better at representing aggregate opinions, compared to another. For example, by reflecting general trends, while also representing minority opinions. This holds both within (between individuals or groups) and between the rating scales themselves. Avoiding this all along and using one and the same rating scale seems like a relatively simple way to avoid this. However, market researchers simply do not always have the data available to them in one scale format, let alone one that is easily comparable to other available scaling formats and is inherently accurate and efficient.

As an example, let us discuss a situation in which a subject rates a certain characteristic of a service. The subject might rate a certain characteristic of a service with a 7 out of 10 on a 1-to-10 rating scale. At the same time, the same subject might have rated that exact same characteristic with either 3 or 4 out of 5 stars in a 5-star rating system. Moreover, what if we compare two different rating scale systems and the aggregate ratings for characteristics A and B and $A_{System1} \succ B_{System1}$ 1 $\wedge$ $B_{System2} \succ A_{System2}$, a preference reversal occurs? What is now the *real* preference of the subject: A or B?

As mentioned before, aggregate ratings can also cause problems when comparing them while applying the same rating scale. For example, it might not be clear how to judge absolute and relative differences. Moreover, behavioral biases can lead to aggregate scores that are clearly not linearly comparable among groups, thus making them unsuited for decision-making.

I stumbled upon this myself during a study on consumer preferences (Brouwer, Ibragimova, Katerberg, & Singh, 2018) for a yet to be released Pokémon game for the Nintendo Switch videogame console (Frank, 2017). For this game, the attitude of participants towards the potential inclusion of five different features for the game was measured. In this study, a clear bias in the results was encountered. In spite of phrasing that tried to prevent this bias from occurring, features that were evaluated more positively by the respondents were given significantly more importance points than features which they awarded with more negative scores. This subsequently led to positive evaluations of features being overvalued compared to negative evaluations. This ultimately led to a positive bias in overall feature evaluation and an inefficiency of the prior-set decision rule. The main aim of this paper will be to investigate one of the potential factors that might lead to this result, namely:

The used measurement scale and phrasing has a significant effect on the respondents' relative evaluations of the features.

In this paper, I provide a framework that might pose a solution for the positive overweighing bias.

To alleviate the bias, research must be done to determine which measurement design would be optimal for representing people's preferences regarding these and similar decisions. This new design would then allow for more optimal decision-making that is in both corporate and societal interest.

The following research question will be answered in this thesis:

'Which measurement scale represents people's preferences best and leads to optimal decision-making?'

The remainder of this paper comprises of a few distinct parts, starting with a theoretical framework. The theoretical framework will discuss previous literature on the topic. Furthermore, it presents and explains the sub-hypotheses and the underlying theoretical concepts that are applied to test these (and that can be used for future research into the topic). Formulating this framework will ultimately support the resolving of the main research question.

Next, there will be a section for Data & Methodology. This section will first discuss the types of data that will be collected and compared. Additionally, it will describe the way in which these data will be collected, the experimental design and how inference is drawn from these data conducive of answering the sub-questions and subsequently the main research question.

The subsequent section will discuss the results of the experiment, focussing mainly on the possible implications for the to-be-answered research sub-questions. Moreover, it will involve multiple statistical tests to robustly check whether the sub-hypotheses hold or not.

The next-to-last section will draw conclusions based on the results of the experiment and answer the sub-questions, which will collectively and subsequently answer the main research question. Finally, the Discussion will consider possible shortcomings of the research and improvements that could be made. Additionally, this section will provide recommendations for potential future research on the topic.

The main result of this study are as follows. Replicating the methodology of the Pokémon Study by Brouwer et al. yields similar results compared to that study, demonstrating the same bias as previously described. Using a newly designed rating scale system proves to alleviate a substantial part of the bias that is observed using the original rating scale system. Additionally, the new rating scale is also tested separately as a third alternative, being augmented with the honesty-inducing Choice-Matching mechanism. Because of a significant and selective drop-off in this third variant, the results produced by this variant are not to be deemed representative of the total population of potential consumers. Even if we would correct/weigh for demographic and/or behavioural variables, some groups of people would still not be represented (sufficiently) for robust statistical inference about the preferences of people in these groups. Consequentially, the new rating scale system, unaugmented by Choice-Matching, is found to be the preferred method for evaluating consumer attitudes regarding product features.

2 Theoretical Framework

As previously discussed in the introduction, the results of the Pokémon Study by Brouwer et al. were the main impulse behind the inception of this thesis. In this section, the methodology and results of this study will be discussed more comprehensively. Moreover, possible solutions and recommendations will be presented based on both the findings of this study and literature that is related to the topic of measurement scales.

2.1 The Pokémon Study

The main aim of this study was to measure the attitude of participants (all Pokémon enthusiasts) towards the potential inclusion of five different features into the game. Close to 200 Pokémon enthusiasts participated in a survey and significant efforts were undertaken to ensure that the participating group was as representative of the target demographic as possible. To measure the preference score for each of the five features, a Likert scale (Likert, 1932) scoring was used. It had a five-point scale (-2 to +2 type) that went from 'fully disagree that feature X should be included' to 'fully agree that feature X should be included'. Additionally, we decided to let our participants allocate units in a 100 points-Constant Sum to express their relative importance of the possible to-be-or-not-to-be included features of the game. We had a thorough discussion about how to phrase the survey questions, with the main aim of preventing any (behavioural) biases from affecting the participants choices. Furthermore, we wanted to have a robust threshold of whether the feature should or should not be included in the game. After that, we made some decisions to tackle biases that could potentially cause the stated preferences of our participants to be a misrepresentation of their actual preferences regarding the features. To start off, we decided that it would be beneficial for our analysis to multiply the individual Likert-scale scores and Constant Sum scores and compute an *Importance-Weighted Preference Score* or *IWPS*.

$$IWPS_f = \frac{L_{1f}CS_{1f} + L_{2f}CS_{2f} + \ldots + L_{nf}CS_{nf}}{n} \quad (1)$$

Where $IWPS_f$ is the importance-weighted preference score for feature f, L_{xf} is the Likert scale preference score of person x for feature f, CS_{xf} is the Constant-Sum importance score and n is the number of participants (Brouwer et al., 2018).

This system, that combined preference and importance ratings, allowed us to tackle a potential bias in the scores that would have caused us to neglect an important part of people's relative evaluation of the features. Specifically, while subjects might indicate that they are either very positive or very negative about a certain feature, they might not feel that it is very important to them whether it is included in the game. This is, compared to other features to which they express a similar degree of preference through the Likert Scale-ratings. So, for example, while a person can feel that they are 'very positive' about two features, they do not necessarily have to feel

that these are equally important to them to be included in the game. We felt that the best way to combine the attitudes towards and importance scores for different features was to multiply them on the individual level. This in turn provided us with individual importance-weighted preference scores that could be compared between different (demographic, behavioural and attitudinal) groups.

Considering that the importance-component provided additional information about the subjects' relative preferences for the features, this system was an improvement over a simple Likert Scale-preference rating scale. However, we acknowledged that there might be another bias. Specifically, a bias in subjects' expression of their preferences, which might be even harder to counter. Essentially, we expected that people might focus more on the features that they wanted to be included in the game (the features for which they expressed a positive preference score), rather than the features they did not want to be included in the game. As a result, they might give the attributes they evaluate more positively more importance points compared to more negatively evaluated features. In our case, this bias might cause people to focus more on features which they awarded more positive preference scores and giving these more importance points than features which they awarded more negative/less positive scores. We made a serious attempt to counter this bias, by including a clear instruction that stated explicitly (in a more informal way) that importance points should be distributed in a certain way. We made it clear that when a feature is equally disliked to how another feature is liked, and the exclusion of the former is as important to the individual as the inclusion of the latter, the same importance score should be given. The main goal of this instruction and the way in which the IWPS was intended, was to make sure that positive opinions of features would not overshadow negative opinions. If there was to be an imbalance between the two, the set decision rule (which assumes neutrality of magnitude for positive and negative evaluations) would be inefficient.

The most interesting finding of the study, however, was that, regardless of our efforts to prevent this bias from happening, the importance-weighted preference scores were seriously right-skewed. More than 73% of the negative scores was between -30 and 0, while only 48% of the positive scores was between 0 and 30 (Brouwer et al., 2018) as can also be seen in Appendix A. Essentially, subjects tended to assign more of their 'importance points' to attributes they are positive about and less to attributes they are negative about. It seems that they are more focussed on the inclusion of features they like than on the exclusion of features they do not like. This skew makes the zero threshold for the importance-weighted preference score an inaccurate and inefficient decision rule.

The remaining parts of this theoretical framework will present possible explanations for this result and will subsequently aim to come up with solutions for the problems that this result causes.

2.2 Measurement scales and phrasing

As discussed in the previous section, the results in the Pokémon Study are quite remarkable. The design of the survey questions might have had a significant effect on how participants responded in the study. Therefore, this following section will focus on a practical and in-depth analysis of the methodology used in the Pokémon Study, adding to paragraph 2.1. Furthermore, this section will discuss in what way the structure of the survey might have influenced the results and, finally, what techniques and measurement scale designs might aid in the debiasing of the results.

2.2.1 The methodology used in the Pokémon Study

Reiterating on 2.1, the main aim of the Pokémon Study was to determine which potentially included features would appeal to the different customer segments that make up the Pokémon fanbase. Ultimately, we wanted to know which of these should (not) be included in the announced Nintendo Switch game, according to potential customers. The to-be-used features in the survey would have to be features that had a realistic, although not certain, chance of being included in the game. Therefore, they should not be features that are already always included in a Pokémon game, but might be features that are included in some of the Pokémon games, but not in all. Exploratory research was conducted among a small sample of Pokémon fanatics, in the form of both interviews and focus groups. These interviews and focus groups applied a laddering procedure that aimed to uncover the meaning that an attribute had to the individual (Reynolds & Gutman, 1988). The main goal of the exploratory research was to discover which features they would like to be included in future Pokémon games (See also Appendix B).

As for the survey itself, it started out with some questions regarding general demographics, familiarity with the Pokémon franchise, playtime regarding different generations [1] of Pokémon games and console ownership. As a result of the exploratory research, the five features that generated the most discussion and interest were selected to be the main variables for the survey. As mentioned before,

[1] Pokémon games are usually released within a three- to four-year cycle, with every new cycle providing new features for the games, such as new Pokémon, updated graphics and a new region to explore. Each generation includes multiple games, but they are all built around the framework that was set-out by the first game of the generation (Bulbapedia, 2018).

a combination of a Likert Scale and Constant Sum was employed to evaluate participants' preferences for the five features.

To limit task complexity, the Likert Scale consisted of a relatively small number of five items, which translated to a -2 to +2 rating-scale for IWPS-calculations. According to a paper on Likert Scale items (Matell & Jacoby, 1971), the number of possible ratings is unlikely to harm validity and reliability, even more so if the sample size is sufficient. Finally, the relative importance of (the inclusion/exclusion of) a feature was measured by the Constant Sum, where participants were asked to allocate 100 points among the five features.

As a result, the general research design for the preference- and importance-evaluating questions for the five features was as follows (for a description of the five features, as phrased in the survey, see Appendix B):

> Preference-evaluation
> Feature X, which means *description of feature,* should be a core feature in the new Pokémon game.
>
> Importance-evaluation
> Below are five potential features of the new Pokémon game, please allocate 100 points among the features, according to how important the features are to you, regardless of whether you like or dislike them.

Clearly, this design already attempted to make clear that, for the importance scoring in particular, participants should not lead themselves to be biased towards awarding importance points towards features they liked, rather than to features that they disliked. But despite this clarifying statement, participants still expressed a bias towards awarding more importance points towards more positively evaluated features.

2.2.2 Potential methodological explanations for the results found in the Pokémon Study and techniques to improve the rating system

Reiterating, the Pokémon Study clearly showed a bias for awarding more importance points to features that were evaluated positively (See also Appendix A), compared to features that were evaluated negatively. There are various factors that might have influenced and caused this result.

First, the *framing* of the questions and the *question order* within the survey should be examined closely. It seems that the questions for the preference ratings were phrased in a relatively straightforward way. Only a short description of each feature was given, but even if a description would cause any unclarities for participants, it

would have more likely affected the preference-score (and subsequently, the IWPS) for that feature, rather than the importance score solely. Let alone the importance score dependent on the preference score. Given that importance scores were only awarded after the preference scores had already been allocated, there might have been order effects. Furthermore, the respondents might have been biased in their importance ratings as a result of the ratings they awarded in the Likert Scale-preference scoring and inadequate adjustment afterwards. This is even more likely, since rating the preference and relative importance of the five features are clearly very much related, even though they are not the same (Gehlbach & Barge, 2012).

Let us then evaluate the rating scales itself. One of the biases that might have occurred, specifically for the Likert Scale preference-rating, is the primacy effect. As described in a paper by Chan on Response-Order Effects, the order in which the value labels are organized might have a significant effect on the amount of times it is selected. This is due to the primacy effect, that causes participants to be more likely to choose the first option that is seen as satisfactory. This means that the order in which the Likert-Scale is constructed (from negative to positive or vice versa) has an influence on the responses that participants give (Chan, 1991). Since it is desirable to put the value labels in a logical order (there are only two of these on a linear scale), we should consider randomizing this order. However, this is not a measure that is to be applied as an experimental variable. This is especially so given the focus of this research with as the main target not to debias the preference scores, but rather the debiasing of the final score (because of the previous bias in importance scores that was dependent on preference scores). Nonetheless, it is still a factor we should consider and therefore, it is important to make sure that all the applied measurement scales have the same order, either from positive to negative or the other way around.

One a side note, one of the techniques that is commonly used for similar analyses, conjoint analysis, can also be applied to determine the attitudes of subjects towards different combinations of attributes/features that are (not) to be included in a prospect. However, since this type of analysis focusses mainly on the result of the in-/exclusion of a certain combination of features, it has two downsides (in the context of our main research topic). Firstly, it focusses less on the individuals' evaluation of the individual features. Secondly, because of the inclusion of combinations (where there are usually *a lot* if one wants to include them all), it requires a significantly higher number of respondents (compared to when a measure would be used that focusses on individuals attitudes towards the in-/exclusion of individual features). This is even more important when one wants to make a segmentation analysis where the attitudes of different groups (segmented based on demographics, behavioural variables, etc.) are analysed, per group or comparatively. Since this is one of the most important analyses that are done using

data that result from this type of analyses, conjoint analysis will not be considered as one of the potential improvements over the IWPS-system.

As a final remark, it might have been the case that the combination of preference- and importance ordering was not clear to the participants. Therefore, it might have been better to use a combined preference- and importance scale. A scale that provides clear endpoints of what is most and least desirable, could possibly make it easier for participants to express their opinion. This is especially important regarding relative comparison among the features since every feature is rated on the same single scale. Furthermore, it might be beneficial to include labels that directly state what the decision of the participant indicates towards the interviewer or the (external client) of the organization taking care of the questionnaires. In this particular case, literally stating the value labels 'this feature should definitely (not) be included in the new game', similarly to a semantic scale, on the two extreme points of the rating axis might make it a lot clearer for the participants. Moreover, this semantic format has been shown to be superior compared to Likert scaling, also in model fit (Friborg, Martinussen, & Rosenvinge, 2006).

As we will, in more depth, discuss in section 3.1.2, I propose to use a new scale that combines preference and importance. The scale will mention the meaning of the extreme points, making it clearer to the participants that their preference signals whether (on the extremes) a feature should definitely be included or not. Moreover, the scale will include an explanation on the middle point. This semantic scale will be shown as a semi-continuous slider to participants in an experiment and will contain 81 points. To recall, we observed in the Pokémon Study that participants gave less importance points to features they expressed a negative opinion towards, compared to features to which expressed a positive opinion. Participants even gave more importance points on average that they were more slightly positive about on the Likert rating scale (+1), compared to features they were strongly negative about (-2). This new scale will be used to test whether a different scale design might lead to an alleviation of this bias. To compare the different rating scale systems, we will also replicate the rating scale from the Pokémon Study and compare it to the newly designed semantic rating scale.

In this paragraph, an alternative for the IWPS-rating scale system was formulated. The analysis in this paragraph leads to the following sub-question:

Sub-question 1: "Is the proposed method of debiasing the IWPS effective in producing a more symmetric distribution of preferences?"

The next section will discuss one additional potential improvement for the measurement scale.

2.2.3 The Bayesian Truth Serum and Choice-Matching

2.2.3a The Bayesian Truth Serum

One additional method of improving the methodology that we will consider is the Bayesian Truth Serum (BTS), which was first described in a paper by Prelec (Prelec, 2004). The main aim of the method is to improve the reliability of subjective data, by making participants think more seriously about their answers. Furthermore, it prevents them from making decisions they feel might be strategically in their favor, but do not accurately present their actual preferences.

The method does this by not only asking for the individuals preferences, but also for his prediction of the preferences of others. By rewarding participants for making better, more accurate predictions (compared to actual preferences of all respondents), they are stimulated to give truthful answers. By incorporating this factor into the model, BTS takes advantage of individuals' perception that there is a correlation between their opinion and the opinions of others.

However, in stark contrast to the models that were introduced before it, BTS does not make any assumptions about this correlation. As a result, consensus is no longer one of the main drivers behind the decision-making process of individuals. This is because more common answers are no longer rewarded, essentially taking away the argument for biasing your answer toward the average of the participant population. Subsequently, the proposed framework ensures that even views that are held by a (small) minority, will be represented. Correspondingly, it provides more accurate and honest responses, as these answers will maximize their expected information score and giving these truthful answers is therefore in the own best interest of individuals (Prelec, 2004).

Essentially, BTS works as follows. It attaches numerical scores to the responses that individuals provide within a rating scale and compares them with the average predicted frequency of these responses, resulting in the information score or iscore (Weaver & Prelec, 2013).

As a result, untruthful answers have lower scores than truthful answers and therefore, truth telling is a strict Bayesian Nash equilibrium for an individual who expects that others provide truthful answers and provide perfect Bayesian predictions of the distribution of the answers (Prelec, 2004). This is regardless of their belief of the relative commonness of their own opinion. After the information scores for the different possible responses are calculated, one or multiple participants are (randomly) selected and paid in form (chosen by the experimenter), according to their information score, rewarding truthful answers.

One possible downside of using the BTS is that it might take more time and effort from your respondents to finish the survey, which could lead to lower completion

rates due to survey fatigue. Fortunately, especially for longer surveys, it is sufficient to select a smaller sample within the group of participants to extract predictions from and calculate iscores, as long as these are randomly selected (Weaver & Prelec, 2013).

2.2.3b Choice-Matching

Recently, a more simple, robust and practical theory for incentive-compatible research has been developed, called Choice-Matching. Considering the standard assumptions of risk-neutral individuals that maximize their expected score (and subsequently maximize utility), according to Bayesian principles, this mechanism promotes truth-telling behaviour in participants, given that everyone else tells the truth (Cvitanić, Prelec, Riley, & Tereick, 2017).

As a scoring rule, the quadratic scoring rule and logarithmic scoring rule can be applied. These scoring rules are an indication of how truthful a respondent has answered and works in a similar way as described for the Bayesian Truth Serum.

Let us now consider the logarithmic scoring rule. In this case, K stands for the fixed payment, $\mathrm{p_{ie}}$ denotes the true relative frequency of an event within the event set *e* for individual *i*, $\mathrm{q_{ie}}$ is *i's* honest estimate of the relative frequency of the events within *e* and $\tilde{\mathrm{q}}_{ie}$ is *i's* claimed estimate of the relative frequency of the events within *e*[2], we observe the following.

Considering these variables, the scoring rule takes the following form (in the case of *n* events):

$$\mathrm{K} + \sum_{e=1}^{n} \mathrm{p_{ie}}\ log\ (\tilde{\mathrm{q}}_{ie}) \quad (2)$$

If individual *i* believes that their honest estimate $\mathrm{q_{ie}}$ is equal to $\mathrm{p_{ie}}$, then the scoring rule for individual *i* will look like this:

$$\mathrm{K} + \sum_{e=1}^{n} \mathrm{q_{ie}}\ log\ (\tilde{\mathrm{q}}_{ie}) \quad (3)$$

In that case, given that individual *i* is a Bayesian expected score maximizer:

$$\mathrm{K} + \sum_{e=1}^{n} \mathrm{q_{ie}}\ log\ (\mathrm{q}_{ie}) > \mathrm{K} + \sum_{e=1}^{n} \mathrm{q_{ie}}\ log\ (\mathrm{q'}_{ie}) \quad (4)$$
$$\forall\ \mathrm{q'}_{ie} \neq \mathrm{q}_{ie}$$

And therefore, any Bayesian expected-score maximizer will be incentivized, through the truth telling mechanism with the logarithmic scoring rule, to answer truthfully. This is the case, since, according to their expectation that their estimate of the true

[2] With $\sum_{e=1}^{n} p_{ie} = 1$, $\sum_{e=1}^{n} q_{ie} = 1$ and $\sum_{e=1}^{n} \tilde{\mathrm{q}}_{ie} = 1$

relative frequency is the best estimate, deviating from their honest estimate is suboptimal (Cvitanić, Prelec, Riley, & Tereick, 2017).

Choice-Matching also contains a truth-inducing mechanism, called Matching. This means that, for each attribute, you are matched with those that expressed a similar opinion, regarding that attribute. Within the Choice-Matching framework, Choice-Matching assigns you a score that is the weighted average of your own prediction score and the prediction score of other participants that expressed the same choice regarding a certain feature. While we are not interested in the actual predictions of participants, the Matching mechanism is a method to elicit truth-telling by participants by rewarding them for it. This is based on the following concept:

You would expect others, who are more similar to you, to also give better predictions, since you also expect that your own prediction is the best (otherwise you would deviate from it). As a result, you would expect this to provide participants with an incentive to speak the truth. This is because one would expect that people want to be matched with people with similar preferences (and a, by this individual, higher perceived prediction score), to maximize their Subjective Expected Utility.

In this paragraph, an augmentation for the newly designed evaluation scale of section 2.2.2 was presented in the form of the truth-inducing and incentive-compatible Choice-Matching procedure. Given that it is simpler to implement, we prefer to use this mechanism over the BTS. To test whether this procedure improves the newly designed scale, the second and final sub-question has been formulated as follows:

Sub-question 2: "Is the Choice-Matching augmented method of debiasing the rating scale effective in producing a more symmetric distribution of preferences?"

The upcoming Data & Methodology-section will provide more details on how this system will be implemented for the experiment, Furthermore, it will discuss how it will be applied to test for the normality and symmetry of preferences for features of consumer goods and finally, how the incentive-compatibility will be ensured (how the participants will be paid out and how this will stimulate participants to answer truthfully).

3 Data & Methodology

This section will discuss the methodology that will be applied to test the two sub-hypotheses, that will subsequently be used to answer the main research question.

3.1 Experimental Design

To test the hypotheses, an experiment will be conducted in the form of a market research survey, in a similar context as the Pokémon Study by Brouwer et al. Five different features will be evaluated by the participants, who will all be Pokémon enthusiasts, that are well-acquainted with the series and will likely know what to expect and what they want out of a Pokémon game. To check eligibility for entering the experiment, participants first must answer some questions about their purchasing behaviour regarding Pokémon games in the past. If they are sufficiently acquainted with the series to partake in the experiment, they advance to the experiment itself.

In an addition to the experimental design of the Pokémon Study, a trailer of a different upcoming Pokémon game, 'Let's Go! Pikachu & Eevee' (also known as LGP/LGE), will be showed to the participants. LGP/LGE introduces some new features that are never seen before or are returning to the series after a long period. Let's Go! is, however, not a 'regular' Pokémon game (Radulovic, 2018). Therefore, it should be regarded as a separate game. The Pokémon game that is referred to in the Pokémon Study is still being worked on by the developer Game Freak and will be released in 2019 (Frank, Pokémon's next core RPG out in 2019, 2018).

Consequently, the video materials that have been released in anticipation of the release of the Let's Go games later this year, equip us with an interesting illustration of how several features look like when implemented in a Pokémon game. A video that shows how the actual features would work in a game, is likely to help the participants to get a better view of whether they would like or dislike them, than a relatively abstract textual explanation. As discussed in a paper by Wakker, experiments should preferably mimic real-world situations as much as possible (Wakker, 2010). Showing a video to participants that shows gameplay of a Pokémon game with several features (that might also be included in the 2019 game), helps contribute to better consonance to real world consumer choice. Since the Let's Go! Games are separate from the 'regular' Pokémon games, there is uncertainty about whether these new or returning features will also be included in the 2019 game. As a result, they provide a very good benchmark to test the preferences of our participants for the 2019 'regular' Pokémon game. Participants will not feel like there is either no chance that they are included in the 2019 game (since it is included in this game, the subjective expected probability is likely higher than 0 for all

participants) or whether it is certain that it is also included in the 2019 game (since the Let's Go! games are not 'regular' games).

The trailer that will be shown is the reveal trailer of the Let's Go! Games (The Official Pokémon Youtube Channel, 2018). The trailer shows several new features of the Let's Go! games, of which five have been selected for this experiment. Additionally, a short textual explanation is provided, to make sure that the participants understand which part of the video is referred to. Table 1 shows the five features that have been selected, including a short textual explanation that was also included in the survey.

① **Wild Pokémon catching (instead of wild Pokémon battles)** Instead of having wild Pokémon battles, Pokémon can be caught using motion controls
② **Pokémon encounters in overworld (instead of random encounter)** Rather than through random encounters (where you randomly encounter Pokémon, for example when walking through grass), Pokémon are encountered in the overworld where you see a specific Pokémon move on the map and can walk into them to start an encounter)
③ **Play together** You can play together with a friend in a local co-op mode
④ **Pokémon GO integration** You can transfer Pokémon from Pokémon GO to your Nintendo Switch game
⑤ **Following Pokémon** Your favorite Pokémon can follow you around and you can ride on them

Table 1: An overview of the five features that will be evaluated by the participants

These features will be evaluated by the participants, who will be distributed randomly among three different experimental settings (that are summarized in figure 3), which will first be discussed one-by-one now (full overview of the survey design can be found in Appendix C).

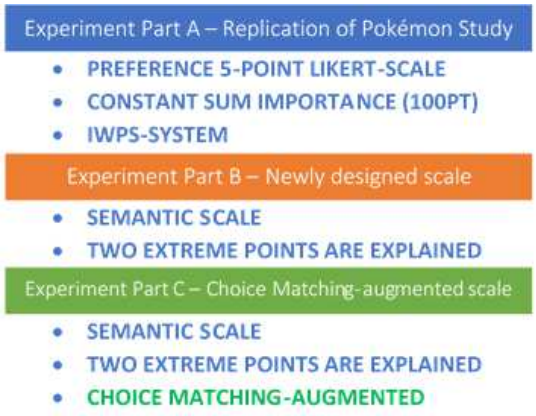

Figure 3: A schematic overview of the experimental design with the three different variants

3.1.1 Experimental variant A – Replication of the Pokémon Study

The first variant includes a rating system that is similar to that used in the Pokémon Study, as described in sections 2.1 and 2.2.1. This means that participants will be able to rate the features according to their preference on a 5-point (-2 to +2) Likert-scale and will be able to express the relative importance of the features with a 100-point Constant Sum. Afterwards, a combination of the individual preference scores (L_{xf}) and importance scores (CS_{xf}) for each future will be multiplied, aggregated and averaged. The result of this is the Importance-weighted preference scores for each of the five features (1), providing a score for each feature on a scale from -200 to +200.

3.1.2 Experimental variant B – Newly designed scale

The second variant will include a newly designed scale that applies potential improvements, in line with the discussion in the theoretical framework. Several designs were considered before choosing what eventually became variant B. These included a design that would only be compared to the Likert-scale, but failed to comprise the importance-part that is fundamental to this thesis research. Another design that was considered used the same 401 point-scaling as the IWPS-system but did not allow for multiple 'extreme' responses, because of the resource-constrainted Constant Sum approach of the IWPS-system.

In the end, the newly designed scale that was chosen is what will be referred to as variant B in this Study. This semantic scale mentions the meaning of the extreme points, making it clearer to the participants that their preference signals whether (on the extremes) a feature should definitely be included or not. Moreover, it will include an explanation on the middle point. This semantic scale will be shown as a semi-continuous slider to the participants and will contain 81 points. Ultimately, those 81 points translate to -40 to 0 to +40 scores. This scale has been chosen, because of its comparability to the IWPS system. This is because, the Constant Sum system under IWPS allows for 100/5 = 20 importance points on average per attribute. Combining this with the Likert-Scale component of -2 to +2, yields the -40 to + 40 scale.

In this way, the new scale is carefully designed to prevent participants from assigning less importance to features they (really) dislike, compared to features they like and alleviate the bias found in the Pokémon Study. Consequently, this design both provides a framework that is easily comparable with the IWPS-framework and is designed to alleviate the biases found in the results produced under IWPS.

Exploratory research, in the form of interviews and focus groups, was conducted to help determine whether this new variant was clear to the respondents. During the exploratory research, no significant problems were found.

3.1.3 Experimental variant C – Choice Matching-augmented scale

The final variant uses the same scale as in variant B, but is augmented with the concept of Choice-Matching. This means that, for each feature, an extra question is included, where participants have to predict how other participants rated the features. They will be matched with participants that have similar preferences compared to them.

Their score regarding one of the features will be calculated according to (7), which can be 'translated' (opposite to generalized) to the situation in the experiment as follows:

$$\text{Participant i's score} = K + p_1 \log(\tilde{q}_{i1}) + p_2 \log(\tilde{q}_{i2}) + p_3 \log(\tilde{q}_{i3}) + p_4 \log(\tilde{q}_{i4}) + p_5 \log(\tilde{q}_{i5}) \quad (9)$$

Where K is the fixed maximum payment, p_q is the actual relative frequency of participants that 'assigned score in q-th quantile for this attribute' and $\tilde{q}_{iq}$ is the individuals claimed prediction of the relative frequency of quantile q for this attribute, among the participants, which we expect to be equal their real estimate, q_{iq} as a result of (8).

As discussed before, Choice-Matching does not only depend on the score of the individual itself, but also on the scores of those that gave similar responses as this particular individual. This is the result of the Matching procedure where the individual is matched with other participants that are similar to them.

Like variant B, exploratory research (interviews and focus groups) was also conducted for variant C. During the exploratory research, some concerns were stressed over the (textual) length of the variant and it not being clear for respondents what they had to do. Nevertheless, the (length of the) text of this variant was not changed, because the effect of the Choice-Matching procedure depends on a number of critical concepts that have to be explained to the respondents. As a result, these could not be left out of the description, since we want to test the effect of Choice-Matching and need to be sure that the explanation for this mechanism is complete.

3.2 Rewarding the participants

After the participants finish with answering the questions of the survey, their answers will be recorded and they get the chance to enter their e-mail address to enter the draw for the price. This is primarily done to make the survey incentive-compatible for the Choice-Matching augmented variant C.

To ensure that the *opportunity* to earn money does not bias the response between the variants, a *fixed* payment is offered (after the survey/experiment has been

conducted) to one ramdomly selected participant for both variant A and B. For variant C, a *variable* amount is offered to one randomly selected participant. How these amounts are determined, depends on the experimental variant (see Table 2).

For experimental **variant A**, a fixed amount of **7.50 Euros** is offered to one respondent.
For experimental **variant B**, a fixed amount of **7.50 Euros** is offered to one respondent.
For experimental **variant C**, the amount depends on their individual score (*i*) for 50% and the score of those they have been matched (*M*) with (50%), for a certain feature. It will be calculated according to this formula[3]: $\text{Payout} = 10 + (20 * (0.5(1 + \sum_{e=1}^{5} \text{p}_{\text{ie}}\ log\ (\tilde{\text{q}}_{ie})) + (0.5(1 + \sum_{e=1}^{5} \text{p}_{\text{Me}}\ log\ (\tilde{\text{q}}_{Me})))))$

Table 2: The payment structures for the three variants

15-20 Euros is here approximately (depending on the distribution of the participants' opinions) the maximum fixed payment amount paid out to the selected participant. This means that this is (dependent on the actual distribution of the opinions) approximately the amount they will earn if they correctly predict everything without the smallest error. This amount will decrease the further the participants' (and those matched with the participant) predictions are from the actual findings in this variant, but will never drop below 10. This amount is higher than the fixed amount of 7.50 for variants A and B. This is not only because 20 is (approximately) the highest possible amount to be paid out, but also because experimental variant C requires more time and effort from the participants. Therefore, it seems fair to also reward them at least a bit more compared to the other variants, regardless of the accuracy their predictions.

3.3 Testing the hypotheses

For testing both hypotheses, we will use statistical methods that evaluate the shape, and specifically the normality and symmetry, of the distributions that result from the three experimental variants. Replicating the experimental design from the Pokémon Study in variant A, the expectation there is that the distribution of the results will be clearly non-normal. For variant B, we apply the new semantic scaling, which has been specifically designed to be clearer for respondents. As a result of this new design, the expectation is that the distribution of the results will be more symmetric. Therefore, we expect it to be (closer to) normally distributed (which is in line with the first hypothesis). Finally, we have variant C, which adds the Choice-Matching mechanism to the experiment. Since this is a truth-inducing mechanism, we expect it to produce results that even more accurately reflect participants' actual opinions.

[3] The set of events (e) contains all five intervals for the preference of the feature, that are used to let the participants make predictions.

As a result, we would expect that the resulting data show even stronger evidence for following a normal distribution (in line with the second hypothesis).

For testing the symmetry and normality of a distribution, several statistical tests can be applied. A number of papers have discussed which of these is best suited for general tests of normality. One of them is a paper by Ghasemi and Zahediasl, which discusses most of these different tests. It puts most emphasis on the Kolmogorov-Smirov (K-S) test, which is the most widely used test for normality, and the Shapiro-Wilk test (S-W). These normality tests compare the results (for example those produced by an experiment), with a set of numbers that is normally distributed and has the same mean and standard deviation as the results. The most important factor in deciding which test to use is its power, which is determined by how capable it is of detecting non-normality in a sample of data. According to Ghasemi and Zahediasl, the widely used K-S test (even if it contains a Lilliefors correction, which makes it less conservative), should not be used, in favor of the S-W test (Ghasemi & Zahediasl, 2012). This finding is in line with the findings from papers by Ahad et al. and Razali and Wah, which both test four widely used normality tests for their power, both again resulting in the Shapiro-Wilk test scoring the highest power (Ahad, Yin, Othman, & Yaacob, 2011) (Razali & Wah, 2011).

The general recommendation that is expressed by these papers is to use the Shapiro-Wilk test, in combination with visual evaluation methods like the Q-Q plot (which is recommended particularly for larger sample sizes) and histograms. Therefore, these methods will be the main framework for the analysis of the results. Additionally, numerical indicators, in the form of skewness and kurtosis indicators, will be used to evaluate the results in the experiment and compare the normality and symmetry among the three variants. Kernel density plots are included in Appendix D as a supplement to the histograms. The following paragraph will give short descriptions of the tools that will be applied.

3.4 Applied methods

First, we will have a look at the *histogram*, which shows the relative distribution of the results. If the shape of the graph resembles that of the typical bell shape of the normal distribution, this might give us an initial indication whether the results are normally distributed or not. Another graph, similar to the histogram, that we will have a look at, is the *kernel density plot*. The kernel density plot is a visualization of the estimates for the probability density function for the evaluated results of an experimental variant. Since this graph is a bit smoother, it makes it easier to evaluate normality and symmetry and to spot signs of non-normality compared to using only the histogram.

The *Q-Q* or *Quantile-Quantile plot* is a graph that compares the (distribution of the) collected data with, in the case of this research, a (theoretical) normal distribution.

The Q-Q plot includes a diagonal line, which indicates whether the distribution of the collected data is (approximately) normally distributed. The closer the datapoints are to the diagonal line, the closer the distribution of the data is to the normal distribution. We will use two variants: one which only shows the data compared to the diagonal line and one that also includes a 95%-confidence interval for normality.

We will also use *skewness and kurtosis indicators*, which are indicators of respectively the skew and shape of the distribution of the data. The closer the skewness is to 0, the less skewed (and the more symmetric) the data are. A negative value for skewness indicates a negative or left-skewed distribution and a positive value for skewness indicates a positive or right-skewed distribution. The value for kurtosis (also known as Pearson's measure of kurtosis) indicates the shape and specifically the tailedness and peakedness of a distribution (compared to the normal distribution). Distributions with a positive kurtosis (>3), also called leptokurtic, have more pronounced tails and peaks and distributions with a negative kurtosis (<3), also called platykurtic, have less pronounced tails and are relatively flatter. The normal distribution has a kurtosis of 3 (also called mesokurtic) and will therefore be the benchmark to which we will compare the results of the three experimental variants in this research (DeCarlo, 1977). Figure 4 includes an illustration of the three types of kurtosis.

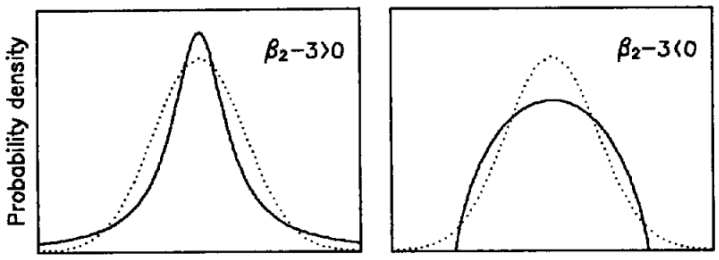

Figure 4: An illustration of kurtosis. The graph on the left shows negative kurtosis/leptokurtic and the graph on the right shows positive kurtosis/platykurtic. The dotted lines show (mesokurtic) normal distributions (DeCarlo, 1977).

Finally, we will evaluate the collected data on normality with the *Shapiro-Wilk test*. The Shapiro-Wilk test evaluates the variance of a data sample and compares the distribution of the data to a normally distributed set of data with a similar mean and standard deviation. The test produces a p-value that can be evaluated using the following framework:

$$H_0 = \textit{the data are normally distributed} \text{ and } H_a = \textit{the data are not normally distributed}$$

In this research, we will apply an α of 0.05. Therefore, if the p-value of the Shapiro-Wilk test is higher than 0.05, then the null-hypothesis can not be rejected and we do not have (sufficient) evidence for the distribution of the data being non-normal. If the p-value of the Shapiro-Wilk test is below 0.05, then the null-hypothesis can be rejected and we can conclude that there is sufficient evidence that the distribution of the data is non-normal (Shapiro & Wilk, 1965). One important note

on the Shapiro-Wilk test is that it is known for more quickly rejecting the null-hypothesis for larger sample sizes, even for small deviations from normality. This is a result of the fact that larger samples include more datapoints and therefore potentially provide more statistical evidence for concluding that the distribution of a data sample is non-normally distributed. Therefore, it is very important that we do not only consider this statistical normality test, but also use the visual evaluation methods and skewness and kurtosis measures (Chan Y. H., 2003).

For the statistical analysis of the data, we will be using the statistical software package R.

3.5 Collected data

In total, 2990 people took part in the experiment (will be referred to as 'participants' from now on). They were mainly approached through local Pokémon GO Whatsapp groups and were stimulated to spread the survey among friends and family that were also adept Pokémon players. In the initial phase of the survey, data were collected on their Pokémon playing behaviour to assess whether they were acquainted enough with the so-called mainline Pokémon games (Bulbapedia, 2018). Figure 4 displays a flowchart that shows the number of participants that were acquainted enough with Pokémon to enter the experiment. Furthermore, it shows how they were subsequently randomly distributed among the variants and how many of those finished the survey, for each variant.

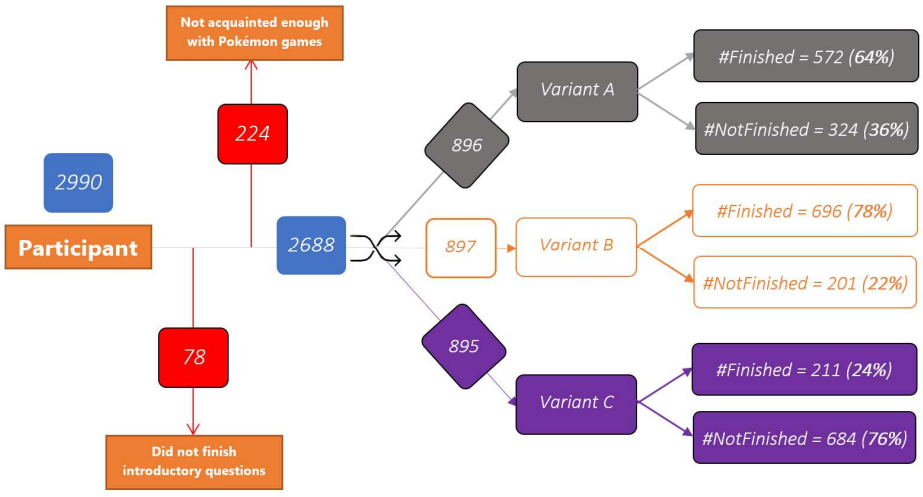

Figure 5: A schematic overview of how a participant went through the experiment

Now, the results of the experiment will be discussed. First, we will discuss the results of the three variants individually, evaluate them using the tools described in the methodology-section and then subsequently compare the outcomes among them.

4 Results

4.1 Results – variant A – Replication of the Pokémon Study methodology

4.1.1 The results within the IWPS-framework

For variant A, we have collected data from 572 participants (64% out of a total of 896 that were selected for this variant) that finished the experiment/survey. We will discuss the Likert-scale preference ratings and Constant Sum importance-scores both individually and as being combined in the Importance-Weighted Preference Score (IWPS). First, we will discuss the results for the five features on the Likert-scale preference scale.

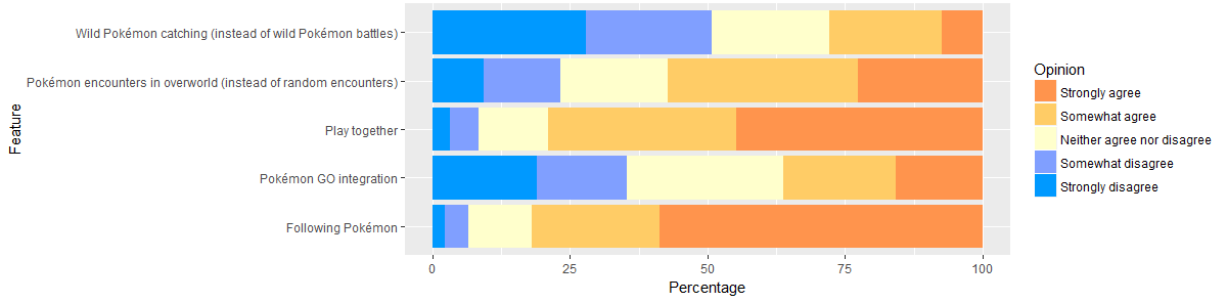

Figure 6: Distribution of the Likert-scale preference scores for the five features

Summarizing this figure and looking at the Likert-scale preference scores, the feature "Wild Pokémon catching" seems to receive generally negative ratings from the respondents. Furthermore, the features "Pokémon GO integration" and "Pokémon encounters in overworld" seem to be quite divisive among our respondents. The features "Play Together" and "Following Pokémon" seem to be generally liked by our respondents.

For the Constant Sum importance-scores, once again we see that participants attribute much more importance points to features they like, rather than features they do not like. Furthermore, they tend to give more importance points, even for features they do not care about, compared to features they do not like at all. Figure 7 shows the results of the CS importance-scores for one of the more divisive features in the preference rating, Pokémon GO integration, based on the preference ratings of the participants for this feature.

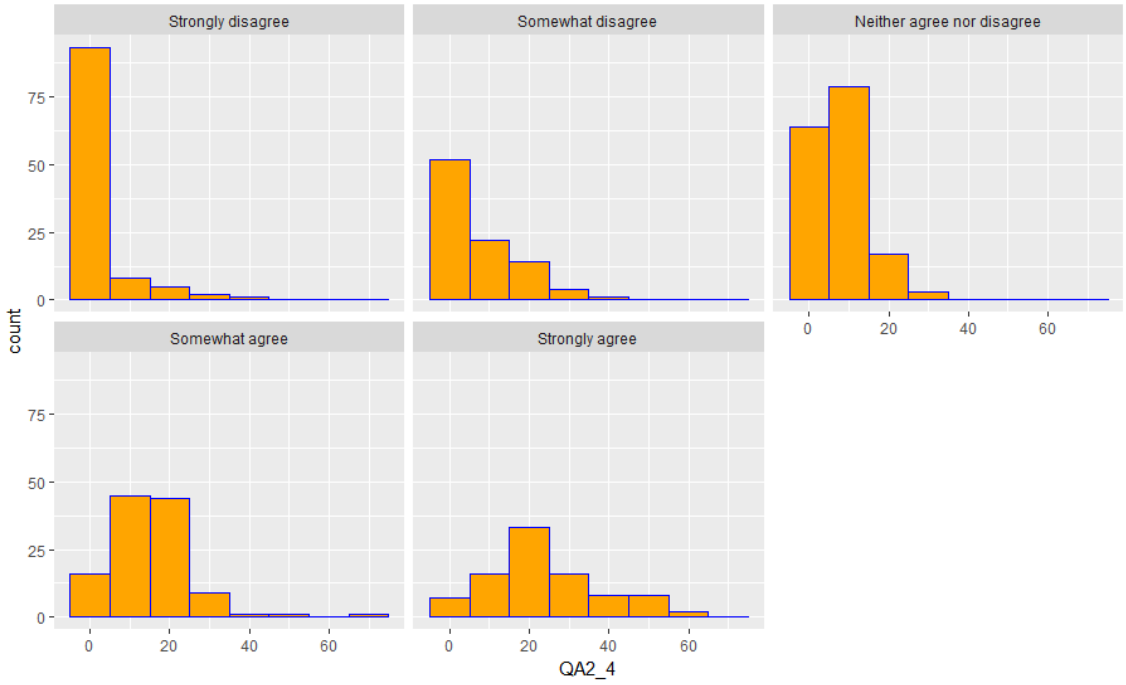

Figure 7: Distribution of the importance scores for each of the five Likert-scale preference categories for the feature 'Pokémon GO Integration'

This is in line with what was to be expected, given the results of the Pokémon Study[4]. When we combine the scores of the preference and importance ratings, we get the following results for the IWPS-score, as shown in Figure 8.

Wild Pokémon catching	Pokémon in overworld	Play Together	Pokémon GO integration	Following Pokémon
-4.21678	**16.86801**	**34.27622**	**8.442308**	**43.70804**

Figure 8: The average IWPS-scores (IWPS = preference x importance) for the five features

According to the IWPS, only the feature "Wild Pokémon Catching" is (by a small margin) rated negatively on average. Pokémon GO integration is rated positively, also by a small margin and the other three features are rated positively overall. The individual results of the preference rating showed us that "Pokémon in overworld" was also a divisive feature for the community. This is however not represented well by the IWPS framework of Variant A, as a result of the participants giving more importance points to features they like, compared to features they do not like.

For the results for the other four features, see Appendix D, which features the full results of the survey/experiment.

[4] For the results for the other four features, see Appendix D, which features the full results of the survey/experiment.

4.1.2 Testing the results for normality and symmetry

We will now have a look at the distribution of the aggregate scores for all five features. They will be evaluated through the methodological framework that was laid out in the methodology-section.

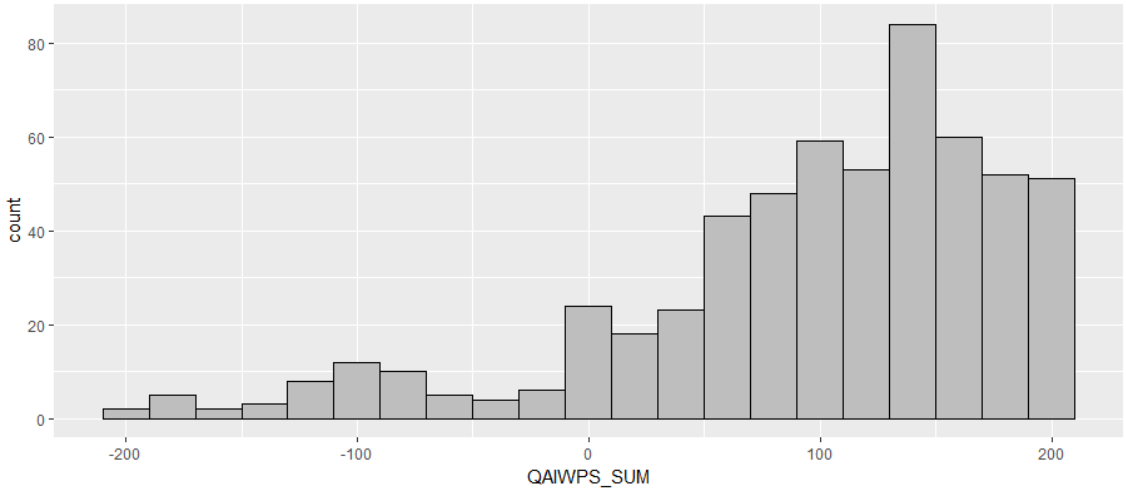

Figure 9: Histogram for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the five features in variant A

When we look at the histogram (figure 9), we see quite a lot of results in the positive half of the domain of [-200,200] of the IWPS framework and almost no (very) negative responses. This could mainly be due to two factors, namely:

- Participants giving (on average) more positive than negative opinions in the preference scoring
- Participants giving (on average) more importance points to features they like compared to features that they do not like

Considering the analysis of the results within the IWPS framework in part 4.1.1, we can conclude that both factors played a role in the realization of this outcome. The histogram clearly shows that the distribution of the results in variant A is neither symmetric nor normally distributed. The data are clearly left-skewed, resulting in a very long tail on the left side of the density plot.

We now take a look at the Q-Q plots (figure 10), which also clearly show that the data are non-normally distributed.

The first (10a) plots the data from variant A against what would result from a perfectly normal distribution with the same mean and standard deviation. The data do not fit the diagonal line well. The second plot (10b) adds an 95%-confidence interval to the graph, which makes it easier to assess whether the data are approximately normally distributed. A large number of datapoints are not within the 95%-confidence interval, so this again provides evidence for the data not being normally distributed.

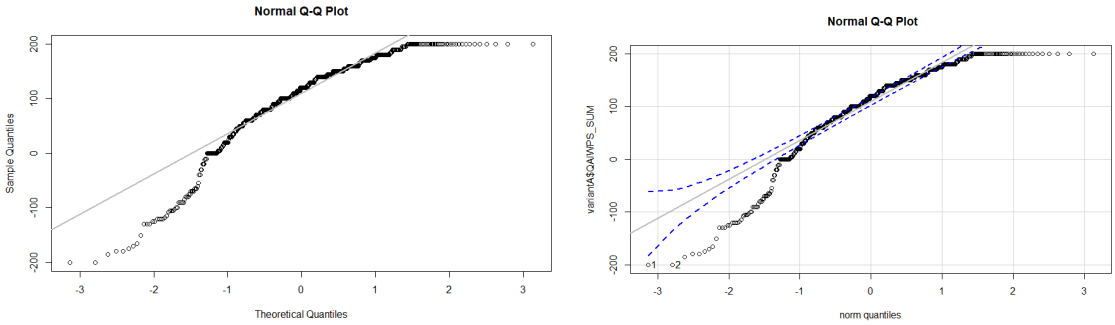

Figure 10: Q-Q plot for the sum of the IWPS results in variant A. On the left (10a) we see the data plotted against the diagonal line and on the right (10b) we see the data plotted against both the diagonal line and a 95%-confidence interval

The skewness statistic for the results in variant A is -1.2538, which is in line with the negative-/left skewed distribution that was observed previously. The kurtosis is 4.35676, which is higher than 3 and therefore the distribution is platykurtic.

Finally, we take a look at the Shapiro-Wilk test for normality. The R-output for the Shapiro-Wilk test is as follows:

Shapiro-Wilk normality test

data: variantA$QAIWPS_SUM
W = 0.89029, p-value < 2.2e-16

According to the results of the test, which show a p-value that is much smaller than 0.05, we can conclude that the data provide enough evidence to reject the null-hypothesis that the data are normally distributed.

As has been shown in this section, the results produced using the methodology of the Pokémon Study by Brouwer et al. and under the IWPS are consistent in showing a left-skewed distribution of Importance-Weighted Preference Scores as expressed by the participants. The results of the replication in this thesis show the same tendency of respondents to overweigh features they evaluate positively compared to features they evaluate negatively that was observed in the Pokémon Study. This holds, even if they are more negative about a feature than they are positive about another.

4.2 Results – variant B – Newly designed scale

4.2.1 Initial analysis of the results

For variant B, we have collected data from 696 participants (78% out of a total of 897 that were selected for this variant) that finished the experiment/survey. We will now first discuss the distribution of the scores for the individual features and subsequently describe the aggregated scores.

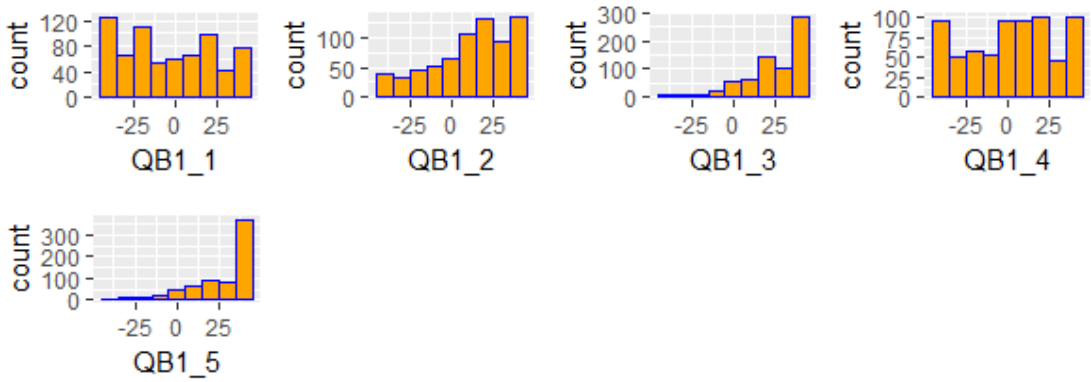

Figure 11: The distribution of the scores for each of the five features (QB1_1 = Pokémon Catching instead of wild battles, QB1_2 = Pokémon encounters in the overworld, QB1_3 = Play Together, QB1_4 = Pokémon GO Integration, QB1_5 = Pokémon following)

In these graphs, we can clearly see that the opinion trends of variant A for the five features are upheld. The major difference with the results under the IWPS-system of variant A is, that in variant B we do see quite a lot of very negative responses among the features. This is in stark contrast with what was observed under IWPS in A, where features that were much disliked were awarded very low importance scores. This resulted in a very small number of very negative responses in the final IWPS-scoring.

4.2.2 Testing the results for normality and symmetry

We will now have a look at the distribution of the aggregate scores for all the five features in variant B.

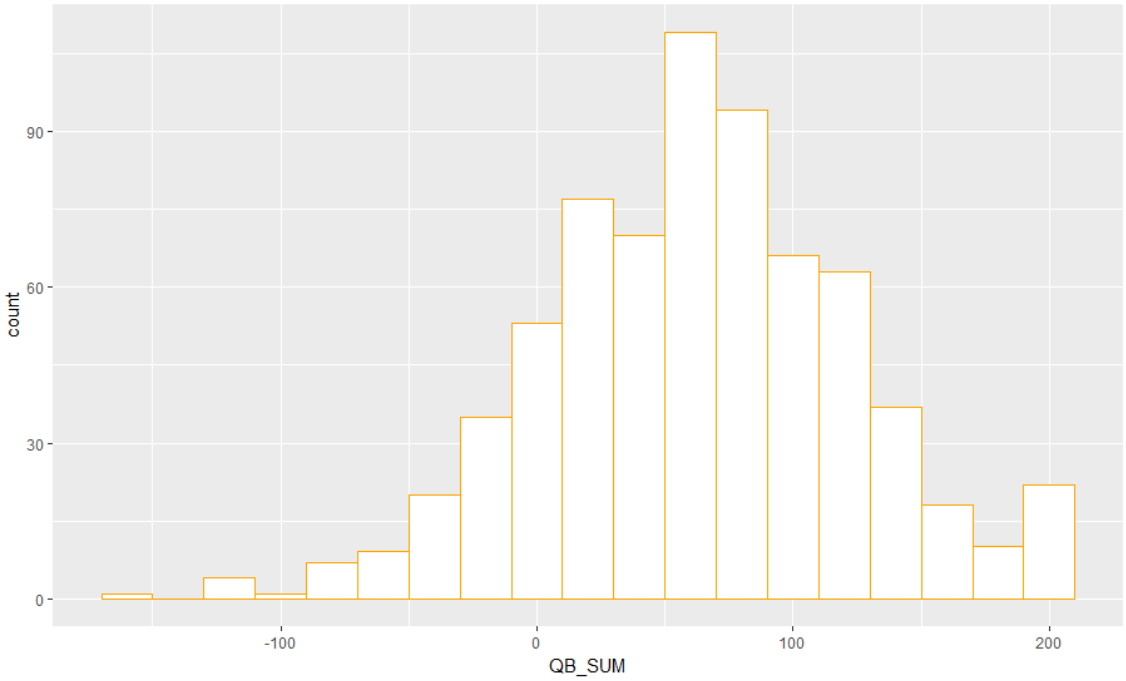

Figure 12: Histogram for the distribution of the participants' sum of the scores for the five features in variant B

As shown in the histogram in figure 12, the scores seem to be relatively (especially compared to variant A's results) symmetrically distributed among a value that is somewhere between 50 and 70 (the exact mean of the data is 63.11). This confirms the previous finding that most of the five features are generally received positively by the respondents, which is in line with what you would expect for the new Let's Go games. Even though the game is quite different from what Pokémon fans are used to and the game developers will try out some new things that might generate mixed opinions, they would of course not want to add (too many) features that are disliked by a significant proportion of the Pokémon fanbase.

One other thing that stands out is that a serious number of participants allocated the maximum positive score of 40 to all five attributes. This is to the detriment of the relatively symmetrical distribution of the data.

The Q-Q plots (figure 13), give some evidence for advocating that the data are approximately normally distributed. The plot on the left (13a) shows that most datapoints lie almost perfectly on the diagonal line, which indicates that they follow the normal distribution. Additionally, the plot on the right (13b) shows that almost all

data points lie within a 95%-confidence interval among the diagonal line. The main expectation here are the datapoints at the end of the tails. The fact that these are partly out of the confidence interval mainly seems to be the result of the significant number of participants that indicated a maximum positive attitude towards all five features.

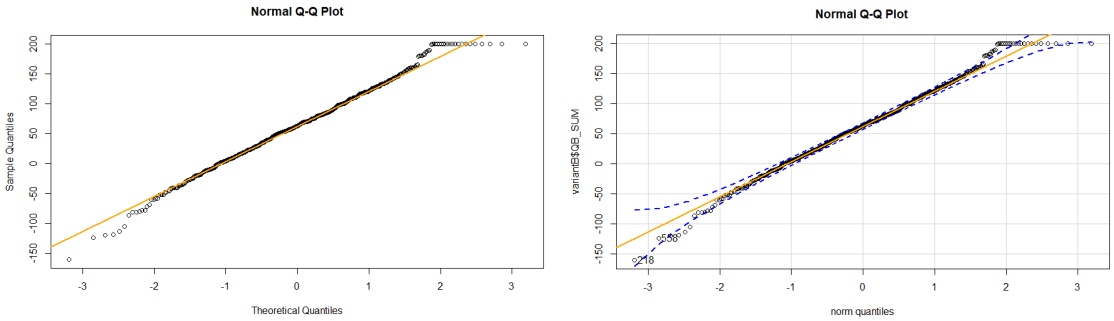

Figure 13: Q-Q plot for the sum of the results in variant B. On the left (13a) we see the data plotted against the diagonal line and on the right (13b) we see the data plotted against both the diagonal line and a 95%-confidence interval

The skewness statistic for the results in variant B is -0.0880, which means that the distribution is very lightly negatively-/left skewed. This is in line with previous findings about the distribution being close to normal. The kurtosis is 3.2447, which is slightly higher than 3 and therefore the distribution is a bit platykurtic, although being close to being perfectly mesokurtic.

Finally, we take a look at the Shapiro-Wilk test for normality. The R-output for the Shapiro-Wilk test is as follows:

```
Shapiro-Wilk normality test

data:  variantB$QB_SUM
W = 0.99319, p-value = 0.002972
```

According to the results of the test, which show a p-value that is smaller than 0.05, we can conclude that the data provide enough evidence to reject the null-hypothesis that the data are normally distributed. This is, even though the data seem to be approximately normally distributed as observed in the visual evaluation.

As stated in the analysis of the distribution of the scores, this is likely mainly due to the serious number of maximum-positive responses. Additionally, this may be the result of the Shapiro-Wilk test being quite strict on normality for larger sample sizes (and 696 is quite a large sample size). Consequently, it will reject the null-hypothesis relatively quickly, even when a distribution is close to being normal (Chan Y. H., 2003). Correspondingly, we should also consider the other measures described before drawing definitive conclusions on the results of this variant.

4.3 Results – variant C – Choice-Matching augmented scale

4.3.1 Initial analysis of the results

Finally, we take a look at the results for variant C. For this variant, we have collected data from 211 participants (24% out of a total of 895 that were selected for this variant) that finished the experiment/survey. This number is already an interesting result by itself and we will return to it later. Now, as for the five features the score distributions are as follows:

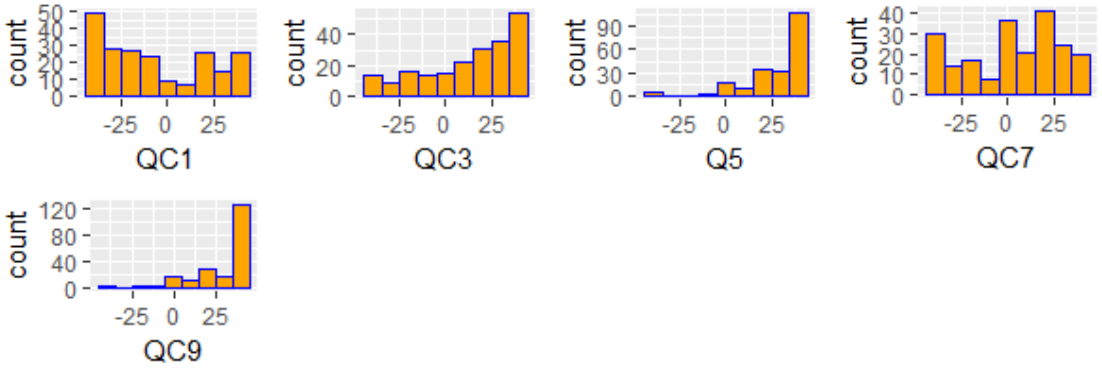

Figure 14: The distribution of the scores for each of the five features (QC1 = Pokémon Catching instead of wild battles, QC3 = Pokémon encounters in the overworld, Q5 = Play Together, QC7 = Pokémon GO Integration, QC9 = Pokémon following)

The graphs in Figure 14 show quite similar distributions of opinion compared to what was observed in variant B. This is not surprising, since the applied scale is the same as the one used for variant B. The main difference with variant B is that C also includes the Choice-Matching process. For the Choice-Matching, respondents were also asked for their prediction of the results for the five features. The results of these predictions (also compared to the actual results) can be found in Appendix E.

4.3.2 Testing the results for normality and symmetry

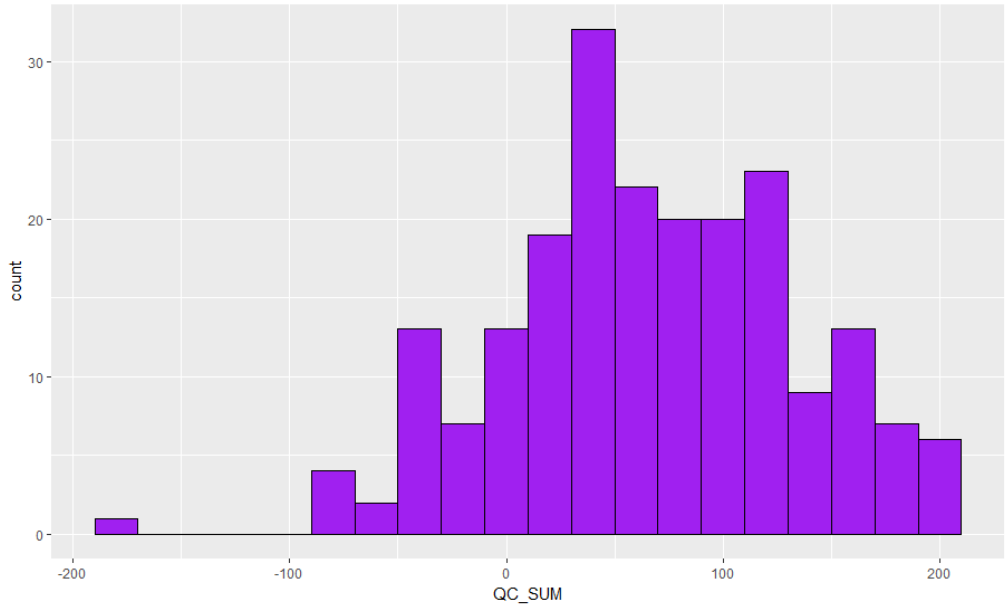

Figure 15: Histogram for the distribution of the participants' sum of the scores for the five features in variant C

According to the results shown in the histogram in figure 15, the aggregate scores are (especially given the smaller sample size compared to variants A and B) relatively symmetrically distributed. The mean value of the aggregate scores is 66.12. The distribution of the scores and the corresponding shape of the histogram are not much different compared to those of variant B. However, the number of respondents that gave a maximal positive score to all five features is not that significant for variant C.

When we take a look at the Q-Q plots in figure 16, we see that almost all datapoints lie on or very much near the diagonal line (see figure 16a). Moreover, all 211 datapoints fall within the 95%-confidence interval among the diagonal line (see figure 16b). Therefore, the Q-Q plots provide us with another piece of evidence for the data being normally distributed in variant C.

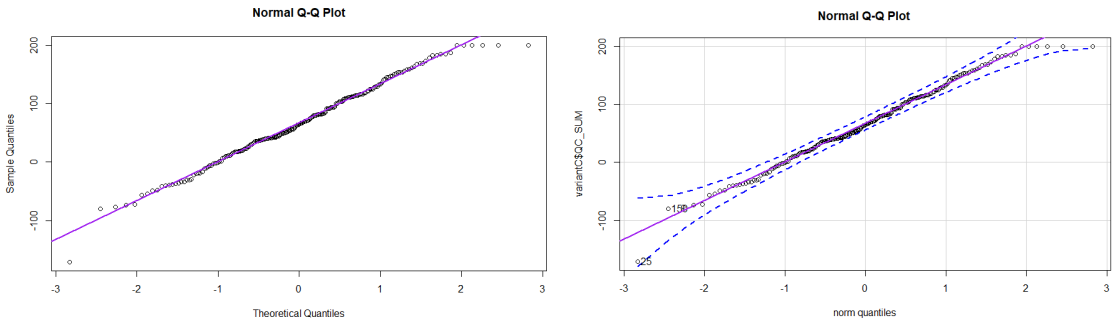

Figure 16: Q-Q plot for the sum of the results in variant C. On the left (16a) we see the data plotted against the diagonal line and on the right (16b) we see the data plotted against both the diagonal line and a 95%-confidence interval

The skewness statistic for the results in variant B is -0.1485, which means that the distribution is lightly negatively-/left skewed. This is in line with previous findings about the distribution being close to normal. The kurtosis is 2.966, which is just a bit lower than 3. This means that the distribution is almost perfectly mesokurtic. It can just barely be called a bit platykurtic.

Finally, we take a look at the Shapiro-Wilk test for normality. The R-output for the Shapiro-Wilk test is as follows:

```
Shapiro-Wilk normality test

data:  variantC$QC_SUM
W = 0.99002, p-value = 0.1524
```

The Shapiro-Wilk test for variant C shows a p-value that is larger than 0.05. As a result, the null-hypothesis can not be rejected due to a lack of statistical evidence for the data not being normally distributed. Since this might be (partly) due to the smaller number of respondents for this variant, the differences in sample sizes will be investigated more thoroughly in the next section. Furthermore, this section will draw comparisons between the survey results that were produced by the three experimental variants.

4.4 Comparing the results among the three variants

As variant A clearly shows similar biases in the data as in Pokémon Study, B and C prove to be improvements. According to most of the analyses conducted in the previous paragraphs, the aggregate score distributions that are produced by variant B and C seem to be approximately normally distributed. According to the Shapiro-Wilk statistic, however, variant B fails the normality test, while for variant C the null-hypothesis of normality can not be rejected. In this section we will mainly compare the results of variant B and C and take a look at the potential sources of differences regarding the results that these two produced.

Although this is not something we can easily check, variant C might have filtered out less interested and/or serious respondents, for example. This is a possibility, since variant C clearly asks more of the respondent compared to variant B, both in time and effort. As a result, less interested and/or serious respondents might have been less likely to finish the survey and more interested and/or serious respondents might have been more likely to finish the survey if they were allocated to variant C, rather than variant B. This specific, more dedicated group that has finished variant C, might have been more accurate and serious in their answers. This might explain the much higher number of extremely positive (maximum positive score for all five features) responses in variant B, compared to variant C. Finally, participants might also be more accurate and honest in their responses because of the truth-inducing Choice-Matching technique that was part of the experimental design of variant C.

On the other hand, there is another possible concern in something I described before: the Shapiro-Wilk normality test is known for more quickly rejecting H_0 if N is very large. Since there is a big discrepancy between the sample sizes of variant B and C, we take a few random samples of the responses from variant B to see how the Shapiro-Wilk test statistic would be with a similar N to variant C. To be sure, we also do the same for variant A, which also had a considerably larger sample size compared to variant C. The full results for the resamples of variant A and variant B can be found in Appendix F.

For the ten resamples of N = 211 of Variant A, the resulting p-values for the Shapiro-Wilk test are all lower than 0.001 and therefore clearly lower than our alpha of 0.05 (see also Figure 17). As a result, this confirms once more that the results produced by variant A are clearly not (close to being) normally distributed. This is, even though we used a significantly smaller sample size this time, similar to the size of Variant C.

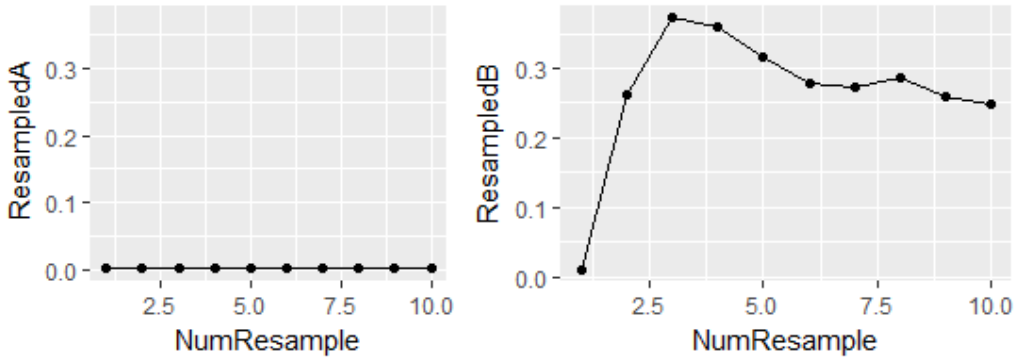

Figure 17: Two graphs that show the means of the (up to that moment) completed resamples. The left graph shows (on the Y-axis) the p-values for the resamples of A and the graph on the right shows (also on the Y-axis) the p-values for the resamples of B.

For the ten resamples of N = 211 of Variant B, these are the resulting p-values for the Shapiro-Wilk test:

P = 0.011 P = 0.513 P = 0.592 P = 0.318 P = 0.152 P = 0.086 P = 0.242 P = 0.374 P = 0.043 P = 0.160

The resulting mean of those p-values for the ten resamples is 0.249, which is a lot higher than the value of 0.002972. This is a direct effect of the smaller sample size of the resamples, which allows for a better comparison to Variant C.

Even though more data should be collected, especially of Variant C, to be able to make statistically robust comparisons and conclusions about the normality of Variant B and C, we can already make some remarks based on this smaller set of resamples of Variant B. If we compare Variant B and Variant C with the same sample size, we see that the mean p-value of the resamples for Variant B is actually higher than the p-value for the results from Variant C (0.1524). While this does not firmly indicate that the distribution of the data from Variant C is closer to a normal distribution, it rather does not give us any evidence to think that the opposite would

be more likely. Based on the current data, we would expect Variant B to produce results that are on average at least as likely as data from Variant C to be normally distributed.

The difference in sample size does reveal one additional concern for variant C, though, and that is sample selection bias. Sample selection bias occurs when non-random samples are used to compare behaviour (for example an expression of opinion) among different groups (Heckman, 1979). Because participants were randomly entered in one of the variants, it is evident that the number of participants that finished variant C is an anomaly compared to the numbers for variant A and B (See also Figure 5). One could be inclined to think that the participants who finished variant C are a non-random selection of the complete pool of participants. Of course, we can also not rule out the possibility of there being significant bias in the type of respondents that finished variants A and B.

Reaching back to the remarks about the exploratory research on variant C (section 3.1.3), there were some serious concerns over the textual length of the explanation for this variant. Adding to this, while the number of participants of the exploratory research was not that high, there was still an interesting trend to be observed. That was, that those with a higher education level stressed significantly less concerns over the textual length and clarity of the procedure, compared to those with lower education levels. Again, the sample size of the exploratory study is not sufficient for any serious inference to be drawn, but this might be a trend that is upheld. Possibly, the textual complexity and length of this variant might cause a selection bias, where participants with a lower education level are less likely to finish the survey than those with a higher education level (also compared to the other two variants).
Unfortunately, given the topic of research, there was no clear reason beforehand to include demographics in the survey, so we can not robustly test this hypothesis.

Nevertheless, some might argue that, in the case that this hypothesis was to be confirmed, this is no big concern for Choice-Matching. Additionally, they might argue that it is perhaps not that bad of an idea to listen more to opinions expressed by those that are sufficiently smart and willing to complete this more complicated variant, compared to those that are either unwilling or unable to do this. Personally, I firmly reject this approach, as it dismisses and depreciates the opinions of those that might already be significantly underrepresented in market research. As found by Berlin et al. (Berlin, Mohadjer, Waksberg, Kolstad, Kirsch, Rock & Yamamoto, 1992), survey respondents with a lower level of education and/or literacy are less likely to respond to surveys, especially when the tasks within the survey require the participant to read long and complex parts of text. Applying methods like Choice-Matching might therefore lead to an even bigger respondent selection bias and hurt the representativeness of the results.

Additionally, as discussed before, multiple respondents approached me individually and commented on the complex nature of variant C (when they were assigned to that variant). I personally know some of these people as being very vocal and clear about their preferences regarding Pokémon games, but they have difficulties with reading, especially with complex, longer texts. In my opinion, it would be a travesty to not take the opinions of these avid Pokémon players into consideration when conducting a Study about preferences of Pokémon fans for a new game. Obviously, the same holds for any other topic of research.

Concluding, those that are able to finish a variant that applies Choice-Matching might certainly be more intelligent or more acquainted with the topic of research. As a result, they might be able to express their opinions more accurately (although one could also ask themselves why they would not also be able to do that without CM then). However, a smaller selective group of 'accurate opinions', might well give a worse representation of the opinions of the full population than a larger, less selective group that is less accurate, especially when their 'errors' are relatively randomly distributed (and not exceptionally skewed to one side).

Now, we return to the main point made about selection bias. Luckily, given that the survey was about features for a new Pokémon game, respondents were asked about their previous experience with Pokémon. This question was included, because this is a variable that might play an important role in their behaviour as a (potential) consumer (Chi, Yeh, & Yang, 2009) for Pokémon games. In particular, the respondents were asked which Pokémon games they played in the past. We will use these data to compare those who finished the three different variants. This will be done both between the different variants and with the full sample of people that at least answered the question about their previous playing experiences (N = 2688).

We will first look at the data of Table 3, which contains the relative and mean relative frequencies of game playment for the six categories of Pokémon games. These relative frequencies were derived from the absolute frequencies that can be found in Appendix G. The relative frequencies of Table 3 show that there are major differences in game playment among those that finished different variants and also compared to and among those that did not finish the survey. To assess which variant produced the results with a population that is most similar to the full respondent pool (and has the least selection bias), additional statistical measures will be used.

To test which of the variants succeeded best regarding representativeness and whether a potential selection bias had a significant effect on the behavioural variables that are compared in Table 3, we use Pearson's Chi-Squared test for Goodness of Fit

Complete dataset that answered INTRO4 (N = 2688)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other	Mean
Did not play games	0.2049851	0.262277	0.40104167	0.04	0.501860119	0.586	0.333147
Did play games	0.7950149	0.737723	0.59895833	0.96	0.498139881	0.414	0.666853

FinishedA (N = 572)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other	Mean
Did not play games	0.1748252	0.218531	0.32692308	0.03	0.451048951	0.538	0.290501
Did play games	0.8251748	0.781469	0.67307692	0.97	0.548951049	0.462	0.709499

FinishedB (N = 696)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other	Mean
Did not play games	0.1997126	0.229885	0.34913793	0.05	0.468390805	0.542	0.306513
Did play games	0.8002874	0.770115	0.65086207	0.95	0.531609195	0.458	0.693487

FinishedC (N = 211)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other	Mean
Did not play games	0.1563981	0.189573	0.33175355	0.04	0.3507109	0.389	0.242496
Did play games	0.8436019	0.810427	0.66824645	0.96	0.6492891	0.611	0.757504

NotFinishedAB (N = 525)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other	Mean
Did not play games	0.232381	0.335238	0.51428571	0.05	0.63047619	0.691	0.409524
Did play games	0.767619	0.664762	0.48571429	0.95	0.36952381	0.309	0.590476

NotFinishedC (N = 684)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other	Mean
Did not play games	0.2280702	0.298246	0.4502924	0.04	0.526315789	0.652	0.36501
Did play games	0.7719298	0.701754	0.5497076	0.96	0.473684211	0.348	0.63499

TotalAB (N = 1793)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other	Mean
Did not play games	0.2013385	0.257111	0.39040714	0.05	0.510317903	0.584	0.331567
Did play games	0.7986615	0.742889	0.60959286	0.95	0.489682097	0.416	0.668433

TotalC (N = 895)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other	Mean
Did not play games	0.2111732	0.272626	0.42234637	0.04	0.484916201	0.59	0.336127
Did play games	0.7888268	0.727374	0.57765363	0.96	0.515083799	0.41	0.663873

Table 3: The relative and mean relative frequencies (on a scale from 0 to 1) of game playment for the six categories of Pokémon games and for each of the completion states for the survey. *FinishedA = finished variant A, FinishedB = finished variant B, FinishedC = finished variant C, NotFinishedAB = started variant A or B (but did not finish), NotFinishedC = started variant C (but did not finish), TotalAB = finished or started variant A or B, TotalC = finished or started variant C. RBYGSCRS = played at least one of Pokémon Red, Blue, Yellow, Gold, Silver, Crystal, Ruby and Sapphire. EDPPtBW = played at least one of Pokémon Emerald, Diamond, Pearl, Platinum, Black and White. B2W2XYSM = played at least one of Pokémon Black2, White2, X, Y, Sun and Moon. GO = played Pokémon GO. MysteryRanger = played at least one of the games in the Pokémon Mystery Dungeon or Pokémon Ranger series. Other = played at least one of the games not specified here. For each individual, a 1 indicates that they have played one of the games in the category, a 0 indicates that they did not. No data is available to distinguish whether those that did not finish either A or B, finished one of these specifically.*

Pearson's Chi-Squared test for Goodness of Fit tests whether the distribution of a set of data is similar to the distribution of / representative of another set of data. It is therefore similar to the Shapiro-Wilk test. The Shapiro-Wilk test is in fact a specific variant of a Goodness of Fit-test that compares a dataset to a hypothetical dataset with the same mean and standard deviation.

Back to the Chi-Squared test for Goodness of Fit that will be applied for the analysis of the results in the different variants. The test produces a Chi-Square value that (as a result of the degrees of freedom, which in our case are always 2 – 1 = 1) matches with a p-value that can be evaluated using the following framework:

$$H_0 = \textit{the observed value is not significantly different from the expected value}$$
$$\text{and } H_a = \textit{the observed data is not significantly different from the expected value}$$

As with the Shapiro-Wilk test, we will apply an α of 0.05. Therefore, if the p-value of the test is higher than 0.05, the null-hypothesis can not be rejected and we do not have (sufficient) evidence for the data not following the distribution of the data it is compared to. If the p-value of the test is below 0.05, then the null-hypothesis can be rejected and we can conclude that there is sufficient evidence for the data not following the distribution of the data it is compared to.

Table 3 shows the results of the Goodness of Fit-test. The first column compares the results from the different variants (from the rows) for each of the six categories of games with the results from the full dataset of respondents that answered the question on game playership. For those that finished either variant A or C, the results of four of the six categories are significantly different to that of the full dataset. For those that finished variant B, this holds for only two out of the four categories produced. This means that the selection bias was more severe in variants A and C, compared to variant B. As a result, the samples of respondents that finished variants A and C are less representative of the full sample of participants compared to the sample obtained through variant B.

Additionally, the results show that the results for those that did not finish either variant A or B and those that did not finish variant C are also significantly different from those of the full dataset for at least three of the four categories. As these are the complements for the categories that did finish those variants, these results are of no surprise. Finally, the results for the total (finished and unfinished) groups that were allocated to A and B and group C, individually, do not significantly differ from those of the full dataset. This is also to be expected given that the respondents were randomly distributed among the variants.

Now, taking a look at the second and third column, the results from the finished and unfinished respondents in variants A and B are compared to the full sample of respondents allocated to one of these variants. Again, the tests show that there is a significant difference between the results of those that finished and those that did not finish, individually, compared to the total pool of respondents that were allocated to these two variants. The same holds for variant C, for which the

comparison is shown in the third column. This confirms our expectation of there being selection bias between those that finished and the full sample of participants that were allocated to the variant. Those that are allocated to either variant A or B *or* to variant C are not found to be significantly different in their game playing behaviour, according to the results of the Chi-Squared test. Consequently, the more severe selection bias for variant A and C, compared to variant B, can only be the result of a more selective, specific group of respondents finishing these variants, compared to variant B.

Based on the results in Table 3, it might seem that those who finished the survey and especially those that finished variant C were more avid Pokémon fans. The main thing that sets them apart from those that did not finished the survey is that, on average, they played games out of more of the six categories. Especially among the (seemingly) more niche game categories of MysteryRanger and Other, the playment numbers are a lot higher for those that finished the survey compared to those that did not. On the other hand, a game like Pokémon GO or the Pokémon games from before 2004 [5] that have been played very ubiquitously, has not been played significantly more by one group, compared to the others.

According to Table 4, there are indeed some statistically significant differences between the different groups. Comparing the results from Table 4 with those of Table 3, it is likely that the differences between those that did finish and did not finish the survey are to a significant extent attributable to how acquainted they are with the Pokémon franchise. Therefore, we will have a final check to see whether this might be the main source of the discrepancies between the results of the different groups.

To test this, we will use the independent samples Z-test for proportions to check whether the proportions of average relative Pokémon game playing behaviour for the six categories are significantly different between the groups. Specifically, we will test whether the average relative proportions of Pokémon game playing behaviour for the six categories are significantly higher for those that finished the survey, compared to the complete group of respondents (also including those that did not finish the survey). Additionally, we will also test whether the average relative propotions of Pokémon game playing behaviour for the six categories are significantly lower for those that did not finish the survey, compared to the complete group of respondents (also including those that did finish the survey). As a final check, we also compare the finished/not finished groups to the total groups *for that variant* and we also test whether there are differences between the complete dataset and the total groups for the variants.

[5] Pokémon Red, Blue, Yellow, Gold, Silver, Crystal, Ruby and Sapphire.

<- Observed	Expected ->	Complete dataset	TotalAB	TotalC
FinishedA	RBYGSCRS	3.1957 (0,07383)	2.4936 (0.1143)	
	EDPPtBW	**5.6629 (0.01733)**	**4.4548 (0.0348)**	
	B2W2XYSM	**13.067 (0.0003)**	**9.6844 (0.0019)**	
	GO	0.6854 (0.4077)	3.392 (0.0655)	
	MysteryRanger	**5.9165 (0.015)**	**8.0359 (0.0046)**	
	Other	**5.3283 (0.0210)**	**4.8825 (0.0271)**	
FinishedB	RBYGSCRS	0.11939 (0.7297)	0.0109 (0.9168)	
	EDPPtBW	3.7794 (0.05189)	2.6989 (0.1004)	
	B2W2XYSM	**7.7936 (0.0052)**	**4.9792 (0.0257)**	
	GO	1.9182 (0.1661)	0.0012 (0.9723)	
	MysteryRanger	3.1261 (0.0771)	**4.8918 (0.0270)**	
	Other	**5.6386 (0.0176)**	**5.1341 (0.0235)**	
FinishedC	RBYGSCRS	3.0582 (0.0803)		3.7912 (0.05152)
	EDPPtBW	**5.7675 (0.0163)**		**7.3353 (0.0068)**
	B2W2XYSM	**4.2122 (0.0401)**		**7.0909 (0.0077)**
	GO	0.0239 (0.8772)		0.0239 (0.8772)
	MysteryRanger	**19.293 (<0.0001)**		**15.212 (<0.0001)**
	Other	**33.882 (<0.0001)**		**35.372 (<0.0001)**
NotFinishedAB	RBYGSCRS	2.4151 (0.1202)	3.1544 (0.0757)	
	EDPPtBW	**14.434 (0.0001)**	**16.782 (<0.0001)**	
	B2W2XYSM	**28.05 (<0.0001)**	**33.857 (<0.0001)**	
	GO	2.4306 (0.119)	0.12281 (0.726)	
	MysteryRanger	**34.717 (<0.0001)**	**30.342 (<0.0001)**	
	Other	**24.053 (<0.0001)**	**24.94 (<0.0001)**	
NotFinishedC	RBYGSCRS	2.2338 (0.135)		1.1828 (0.2768)
	EDPPtBW	**4.5674 (0.0326)**		2.2687 (0.132)
	B2W2XYSM	**6.919 (0.0085)**		2.1969 (0.1383)
	GO	0.4298 (0.5121)		0.4298 (0.5121)
	MysteryRanger	1.631 (0.2016)		**4.6973 (0.0321)**
	Other	**12.299 (0.0005)**		**10.886 (0.0010)**
TotalAB	RBYGSCRS	0.1475 (0.7009)		0.5426 (0.4613)
	EDPPtBW	0.2495 (0.6174)		1.1295 (0.2879)
	B2W2XYSM	0.8376 (0.3601)		3.8381 (0.0501)
	GO	1.5349 (0.2154)		3.8239 (0.0505)
	MysteryRanger	0.5082 (0.4759)		2.3077 (0.1287)
	Other	0.0167 (0.8971)		0.1302 (0.7183)
TotalC	RBYGSCRS	0.20928 (0.6473)	0.5426 (0.4613)	
	EDPPtBW	0.4932 (0.4825)	1.1295 (0.2879)	
	B2W2XYSM	1.6979 (0.1926)	3.8381 (0.0501)	
	GO	0.4202 (0.5169)	3.8239 (0.0505)	
	MysteryRanger	1.0327 (0.3095)	2.3077 (0.1287)	
	Other	0.0574 (0.8107)	0.1302 (0.7183)	

Table 4: The values for the Pearson's Chi-Squared test for Goodness of Fit and the corresponding p-values (within brackets). An **enboldened value** indicates that the observed data on this behavioural (gaming behaviour) variable (in the rows) is significantly different from what would be expected under the data distribution of the (sub)set of data (in the colums).

The test statistic for the independent samples test on proportions is as follows:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Where z is the test statistic, $\hat{p}_1$ and $\hat{p}_2$ are the estimated proportions of mean playment for the six categories that resulted from the samples, n_1 and n_2 are the sample sizes of respectively sample 1 and 2 and $\hat{p}$ is the pooled estimator under the null-hypothesis (which is the sample size-weighted average of $\hat{p}_1$ and $\hat{p}_2$[6]) (The University of Sydney, 2013).

The results of the independent samples tests on proportions for the mean relative frequencies of game playment can be found in Table 5.

	Compared to ->	Complete dataset	TotalAB	TotalC
FinishedA	Mean relative frequency of game playment for the six categories	1.8770	1.8323	
Finished	Mean relative frequency of game playment for the six categories	1.3325	1.2003	
FinishedC	Mean relative frequency of game playment for the six categories	**2.7024**		**2.6258**
NotFinishedAB	Mean relative frequency of game playment for the six categories	**-3.3667**	**-3.2929**	
NotFinished	Mean relative frequency of game playment for the six categories	-1.5729		-1.1941
TotalAB	Mean relative frequency of game playment for the six categories	0.1042		-0.2333
TotalC	Mean relative frequency of game playment for the six categories	0.1648	-0.2333	

Table 5: The values for the independent samples tests of proportions and the corresponding p-values (within brackets). An enboldened value indicates that the mean relative frequencies of game playment (for the six categories) are significantly different between the two groups. The groups on the left/the row groups, form group 1 with $\hat{p}_1$ *and* n_1, and the groups that are above the table/the column groups, form group 2 with $\hat{p}_2$ *and* n_2.

In this table, a negative value of lower than -1.96 indicates that the mean relative frequency for group 1 is significantly lower than that of group 2. A positive value of

[6] $\hat{p}$ *, the sample size-weighted average of* $\hat{p}_1$ *and* $\hat{p}_2$ *is estimated by* $\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$.

higher than 1.96 means that that the mean relative frequency for group 1 is significantly higher than that of group 2. Consequently, the only groups that produced significantly different results among them where the following two groups. The implications of these results will also be discussed shortly.

Firstly, we see a significant positive result for the group of participants that finished variant C. This is both when compared to the full group of participants that answered the question on playment and when compared to all participants that were allocated to variant C, regardless of finishing status. This result indicates that the participants that finished variant C have on average played (at least one of the) Pokémon games out of a significantly higher number of the six specified categories. It confirms that there is indeed significant selection bias as a result of the variant that participants go through. Specifically, being allocated to variant C has this effect, because there is a high and selective drop-off rate.

It seems that participants that are more avid Pokémon players are more likely to finish the survey, as all results (for all three variants) for those that finished the survey are positive and all results for those that did not finish the survey are negative. The effects of this are however only translated in a statistically significant magnitude in variant C. Because of the higher drop-off rate compared to variant A and B, the selection bias is so strong that the remaining group of participants that finished variant C can not be deemed representative of the full population of participants. This result is not only statistically significant, but also has important consequences for inference based on these statistics. In this specific case, game playment is potentially a very important behavioural variable that might be one of the factors that induces players to like or dislike certain features more or less.

Secondly, a significant negative result is observed for the group that was allocated to either variant A or variant B and did not finish this variant. This result holds both when compared to the full group of participants that answered the question on playment and also in a comparison with all participants that were allocated to variant A or B, regardless of finishing status. Given the discussion before, this result is likely to indicate that those that were allocated to variant A or B and did not finish the survey, are more likely to be less acquainted with the Pokémon series. This result is the logical antagonist of the positive (although not significant) positive results of the participants that were allocated to variant A or B and that *did* finish the survey. As these results were already discussed before, we will not discuss the results for the antagonist in more detail.

Now, the next and final section will draw conclusions based on the results that were discussed in this section, answering the sub-questions and ultimately the main research question. In addition to that, it will also contain a further discussion of what the implications of the results are for market research. Finally, the challenges and opportunities for follow-up research will be discussed.

5 Conclusion and discussion

In this study, the following main research question was presented:

'Which measurement scale represents people's preferences best and leads to optimal decision-making?'

To answer this question, two sub-questions were formulated, with two corresponding sub-hypotheses that will be tested to answer these sub-questions. To test these sub-hypotheses, a between subjects-designed experiment was conducted in the form of a survey. This survey consisted of three distinct variants, which applied different measures for evaluating the opinion of participants regarding features of a product (in our case a video game).

5.1 Conclusion and discussion regarding sub-hypothesis 1

The first variant, variant A, was a replication of the system used in the Pokémon Study by Brouwer et al., which produced biased results in favour of positively-rated features as per individual. It included two different rating scales, one for preference and one for importance. These were combined in the Importance-Weighted Preference Score to evaluate the preference for the features, weighted by how important participants found them to be relatively.

In the Pokémon Study, the bias was a result of participants giving (on average) more importance points to features they liked, rather than features they disliked. This finding was confirmed to be apparent in variant A in this research. The main question for now was whether this bias might be persistent, even when using a different measurement scale. In that case, there might be a Benefit-Seeking Bias, in stark contrast with general findings of Loss Aversion.

To check whether this bias was the result of the used measurement scale or if it was persistent, efforts were made to construct a new scale that would potentially produce less biased results. The result of these efforts was variant B, which included a single semantic scale as the main measurement standard. This new scale incorporated the concepts of both preference and (relative) importance into one scale, potentially allowing for easier comparison between features and more clearness. Furthermore, the single scale might have mitigated the tendency of respondents to overly express importance to positively evaluated features in the double-scale IWPS approach.

This leads us to the first sub-question of this research, which was formulated as follows:

Sub-question 1: "Is the proposed method of debiasing the IWPS effective in producing a more symmetric distribution of preferences?"

To answer this question, the results of the survey were compared among the two variants, A and B, to see what kind of effect the applied measures had on the way

respondents reported their opinions regarding the features. Statistical tests, measures and visual evaluation methods were applied to assess the normality and symmetry of the distribution of the aggregate ratings for the features. Since we would expect the new scaling to produce a distribution of preferences that is less biased towards the positive compared to the old IWPS-system the following sub-hypothesis can be formulated:

Sub-hypothesis 1: "The proposed method of debiasing the IWPS is effective in producing a more symmetric distribution of preferences"

The results of the visual evaluation methods are clear. The aggregate ratings under the IWPS-system of variant A are not normally distributed since they show a clear left-skew. This visual assessment is confirmed by considering the statistical measures and the statistical test on normality, where the null-hypothesis of normality is rejected by a large margin. For the new rating system that is applied in variant B, the assessment techniques are not in full agreement. The visual evaluation methods and the statistical measures of skewness and kurtosis do not show large deviations from normality, suggesting an approximal normal distribution. However, contrarily, the statistical test on normality rejects the null-hypothesis of normality, albeit with a much smaller margin than for variant A. Nonetheless, all evaluation methods evidently show that the data produced by variant B are considerably closer to being normally distributed than the data produced by variant A. Besides that, as discussed in earlier sections, statistical tests on normality are known for quickly rejecting the null-hypothesis of normality for larger sample sizes (>100), even for very small deviations from normality. As a result, we should also take the other measures into account when drawing a conclusion (Chan Y. H., 2003).

Considering the above, we can conclude the following regarding sub-hypothesis 1:

The results that have been collected in this research give evidence to expect that sub-hypothesis 1 can be accepted. Variant B produces results that are clearly closer to being normally distributed than those that are produced by variant A, more accurately representing the participants' relative opinions regarding products' features. The visual evaluation methods show relatively less outliers for variant B, compared to variant A. Additionally, the Shapiro-Wilk test for normality shows an extremely low p-value for variant A, clearly rejecting normality. While the p-value for SW for variant B is also sufficiently low (<0.05) to reject the null-hypothesis, it is considerably closer to the rejection point. Therefore, the newly designed scale is an improvement over the IWPS-system. Additionally, we can conclude that the benefit-seeking bias does not seem to be persistent when applying the new scale. No evidence could be found of participants (on average) overweighing features they like compared to features they do not like.

On a decision-making level, the IWPS-system might cause companies to include features in their products that are disliked by a significant part of their potential customer base. As can clearly be observed in the results for the individual features for variant B (figure 12), the feature 'Pokémon GO integration' is met with very mixed responses by the respondents. According to the IWPS-system of variant A (see figure 8 for results) and the corresponding decision rule (>0 means a positive advice), however, the results suggest that the feature should be included. This is again the direct result of the positive opinions about a feature being overweighted compared to the negative opinions.

If The Pokémon Company would decide on including a feature in their game and/or emphasizing it in advertisements, it might make the faulty decision of including /emphasizing the Pokémon GO integration feature a lot, when using the IWPS-system. This might not resonate well at all with a significant portion of the potential consumer base. Clearly, the choice of evaluation measures for market research for The Pokémon Company can have huge consequences for the sales of the game and subsequent future installments in the series. Considering the two systems evaluated in this section, it would be better for The Pokémon Company to use the new scaling that is applied in variant B.

5.2 Conclusion and discussion regarding sub-hypothesis 2

In addition to the new measurement scale, we also investigated further improvements of the design for evaluating customer opinions on product features. Specifically, we considered the effects of including a mechanism that would be both incentive compatible and truth-inducing. As a result, the final variant C included the Choice-Matching mechanism as described in the paper by Cvitanić, Prelec, Riley & Tereick. With the Choice-Matching mechanism included, participants are not only asked for their own opinion, but also for their prediction of the opinions of others. As described before, it is in their best interest (given that their potential rewards are bigger) to answer truthfully. If respondents actually answer more truthfully, then one would expect the results from this variant to represent the opinions of the participants more accurately.

This leads us to the first sub-question of this research, which was formulated as follows:

Sub-question 2: "Is the Choice-Matching augmented method of debiasing the rating scale effective in producing a more symmetric distribution of preferences?"

Since we want to extract information about the effect of Choice-Matching, we compare the results for variants B and C to answer this question. The only difference between these two variants is that C includes the Choice-Matching mechanism and B does not. Similar to the analysis for answering sub-question 1, multiple measures

were applied to assess both the normality and symmetry of the aggregate ratings for the features. As a result of the expectation that Choice-Matching will enhance the accuracy of the new scale, the second and final sub-hypothesis was formulated as follows:

Sub-hypothesis 2: "The Choice-Matching augmented method of debiasing the rating scale is effective in producing a more symmetric distribution of preferences"

In the case of variants B and C, the distinction between their results is much less clear, as compared to those between variant A and B. The visual evaluation methods and skewness and kurtosis measures do not show a clear improvement (regarding normality) in the results for variant C, compared to those of variant B. The statistical test on normality, however, shows contrasting results between the two, in favour of variant C. According to the Shapiro-Wilk test on normality, the null-hypothesis for normality is rejected for the results of variant B, but is not rejected for the results in variant C.

When taking a second look at the visual evaluation methods, one of the potential reasons for this becomes evident. The kernel density plot of the results from variant B (Figure 14) shows a minor peak in the results for a small but significant group of participants that gave a maximal positive score to all five features. This peak is less appearant in the results for variant C (Figure 17). This small difference, however, is not the main source of the difference between variant B and C in the normality test.

A major difference between the data collected in variants A and B on the one hand and C on the other, was that the number of participants that finished variant C, 211, was drastically smaller than that of variants A and B, which was 572 and 696 respectively. This result is substantial, because participants were distributed randomly among the three variants.

As discussed before, the Shapiro-Wilk test for normality is known for more quickly rejecting the null-hypothesis of the data being normally distributed for larger sample sizes. The larger the sample size is, the more data the test has available to it, that might provide evidence for rejecting normality for those data. Given that there was a significant difference between the sample sizes of A and B, compared to C, ten resamples of size N = 211 were taken from the data of both A and B. The results for A were in line with what was found before. Even when taking a smaller sample, normality is out of the question and the null-hypothesis is clearly rejected. For variant B, the results for the resamples of variant B did on average not provide sufficient evidence to reject normality for the data. This is in clear contrast with the results for the full sample, where the test found enough evidence to reject the null-hypothesis for normality.

Even more so, the average p-value (0.249) for the ten resamples was higher than that of the results from variant C (0.1524). This means that (on average) the normality

test found less evidence to reject the null-hypothesis for the data in variant B, compared to variant C. Obviously, more research needs to be done to evaluate which of the two variants is more accurate and produces more normally distributed results.

There is one other concern that is highlighted by the large number of drop-offs in variant C and which was discussed before, namely sample selection bias. Sample selection bias means that the sample that is taken from a population is non-randomly selected and therefore not representative of that population. To test for sample selection bias, Chi-Squared tests on Goodness of Fit were applied for behavioural variables on game playment. The different subsets of participants were compared to the full sample of participants (that filled out the question on game playment). Variant A and especially variant C (through its very high drop-off rate) seemed to have significant selection bias, among those that finished these variants, regarding some of the categories of games they had (not) played. Variant B had experienced a lesser drop-off rate and this also resulted in less selection bias among these variables.

Finally, average game playment of the six categories that were specified in the survey, was compared with independent samples tests on proportions. In this test, the participants that finished variant C were found to have a significantly different playing history regarding Pokémon games. In particular, on average, they had played at least one game per category significantly more than those that were (randomly) allocated to variant C and also compared to all that filled out the question on game playment.

Specifically, as playment of more categories of Pokémon games likely indicates a higher-level relationship between the player and the brand Pokémon, participants that are part of the results of variant C are on average more likely to have a higher-level customer relationship with Pokémon. In other words, they are likely to be more dedicated to the franchise and are more avid fanatics about Pokémon (on average). While it is essential to keep in contact with your most loyal and avid customers, it is obviously not desirable to repel potential customers that are currently more casual players (but are potential fanatics) from participating in a survey on their preferences.

Selection bias is something that can be corrected for, for instance with Heckman's two-step estimator (Heckman, The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, 1976). However, the drop-off for variant C is so remarkably high[7] (as well as selective) that it has the potential for (almost) completely excluding specific groups of possible customers. Clearly, if these specific groups do not finish the survey at all (or in a very small number), it will be unworkable to draw robust

[7] 76%, see also Figure 5

statistical inference about their opinions and potential behaviour as a consumer, even when using a reweighting method.

Considering the above, we can conclude the following regarding sub-hypothesis 2:

The results that have been collected in this research give evidence to expect that sub-hypothesis 2 can not be accepted. At first glance, the Choice-Matching augmented variant C produces results that are closer to being normally distributed, compared to variant B. The visual evaluation methods show relatively less outliers for variant C and the Shapiro-Wilk test for normality rejects the null-hypothesis for symmetry for variant B, but not for C. Nevertheless, this is proven to be primarily due to the smaller sample size of variant C, which revealed the other main concern regarding this variant, sample selection bias. Significant sample selection bias was found for (those that finished) variant C regarding game playing behaviour, an essential variable for market research in gaming and potentially one of the main instigators of their preferences regarding game features.

5.3 General recommendations for market researchers/analysts

It is evident that the use of a certain metric can cause the results and subsequently the recommendations that may be inferred from these results to be significantly different. In the specific case discussed in this thesis, this might result in the inclusion of one or more game features that is disliked by significant proportions of the potential consumer base, based on the results from a survey with, for instance, a rating scale that produces an overly left skewed distribution of scores for features. Antagonistically, it might also lead to the exclusion of one or more game features that is liked by a large proportion of the potential consumer base, based on the results from a survey with, for instance, a rating scale that produces an overly right skewed distribution of scores for features.

Therefore, anyone in the field should be aware of the different results that the various rating scales can produce.

5.4 Opportunities for future research

For future research regarding this topic, I would advice to focus primarily on two specific areas.

Firstly, I believe that more research should be done to determine **when** research scales are most accurate. The assumption of normality as an indicator of preciseness for this thesis was mostly the result of the asymmetry in the results caused by the IWPS-system. While it was in hindsind an indicator of the IWPS's bias that made participants overweight positively evaluated features over negatively evaluated

features, it might not be an ubiquitiously applicable indicator for the efficiency of a rating scale.

Secondly, I would strongly advice doing additional research on Choice-Matching. The high drop-off rate is a serious point of concern, not only because it hinders the opportunities of statistically robust analysis (by producing a smaller sample size of results), but also because it is proven to cause sample selection bias. One of the options would be to test and compare different (including some new) variants of Choice-Matching. The new variants would have to allow for less strict and shorter phrasing/explanation, at the expense of having to make some additional assumptions about respondent behaviour. Specifically, and finally, given in by the concerns stressed in the exploratory research, I recommend researching the potential link between education level and finishing status of these Choice-Matching augmented survey variants.

5.5 Conclusion regarding the main research question

Reiterating, the main research question of this thesis was:

'Which measurement scale represents people's preferences best and leads to optimal decision-making?'

Given the disquisition in this thesis, the measurement scale that represents customer's preferences best is, among those considered, **variant B**.

Bibliography

Ahad, N. A., Yin, T. S., Othman, A. R., & Yaacob, C. R. (2011). Sensitivity of normality tests to non-normal data. *Sains Malaysiana*, 637-641.

Berlin, M., Mohadjer, L., Waksberg, J., Kolstad, A., Kirsch, I., Rock, D., & Yamamoto, K. (1992). An experiment in monetary incentives. *In Proceedings of the Survey Research Methods Section of the American Statistical Association*, 393-398.

Brouwer, D., Ibragimova, D., Katerberg, J., & Singh, H. (2018). *The desirability and importance of features for an upcoming Pokémon game.* mimeo.

Bulbapedia. (2018). *Generation*. Retrieved from Bulbapedia: https://bulbapedia.bulbagarden.net/wiki/Generation

Chan, J. C. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement*, 531-540.

Chan, Y. H. (2003). Biostatistics 101: data presentation. *Singapore medical journal*, 280-285.

Chi, H. K., Yeh, H. R., & Yang, Y. T. (2009). The impact of brand awareness on consumer purchase intention: The mediating effect of perceived quality and brand loyalty. *The Journal of International Management Studies*, 135-144.

Cvitanić, J., Prelec, D., Riley, B., & Tereick, B. (2017). Honesty via Choice-Matching. Working paper.

DeCarlo, L. T. (1977). On the meaning and use of kurtosis. *Psychological methods*, 292.

Frank, A. (2017, June 13). *A Pokémon RPG's coming to Switch*. Retrieved from Polygon: https://www.polygon.com/2017/6/13/15788898/pokemon-nintendo-switch-nintendo-e3-2017

Frank, A. (2018, May 30). *Pokémon's next core RPG out in 2019*. Retrieved from Polygon: https://www.polygon.com/2018/5/30/17408632/pokemon-gen-8-nintendo-switch-release-date-2019

Friborg, O., Martinussen, M., & Rosenvinge, J. H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences*, 873-884.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 486.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *In Annals of Economic and Social Measurement*, 475-492.

Heckman, J. J. (1979). Sample selection bias as a specification error (with an application to the estimation of labor supply functions). *Econometrica*, 153-162.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and psychological measurement*, 657-674.

Moorman, C. (2018, February 27). *Marketing Analytics And Marketing Technology Trends To Watch*. Retrieved from Forbes.com: https://www.forbes.com/sites/christinemoorman/2018/02/27/marketing-analytics-and-marketing-technology-trends-to-watch/#6156aed71b8a

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, 462-466.

Radulovic, P. (2018, June 7). *Pokémon: Let's Go!: everything we know*. Retrieved from Polygon: https://www.polygon.com/2018/5/30/17409664/pokemon-lets-go-pikachu-eevee-switch-pokeball-plus

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 21-33.

Reynolds, T. J., & Gutman, J. (1988). Laddering theory, method, analysis, and interpretation. *Journal of advertising research*, 11-31.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 591-611.

The Official Pokémon Youtube Channel. (2018, May 29). *Pokémon: Let's Go, Pikachu! and Pokémon: Let's Go, Eevee! Trailer.* Retrieved from Youtube: https://www.youtube.com/watch?v=9jRtpMKLsts

The University of Sydney. (2013). *MATH1015 Biostatistics Week 10.* Retrieved from University of Sydney - School of Mathematics and Statistics: http://www.maths.usyd.edu.au/u/jchan/MATH1015/bis13_10_nosol.pdf

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity.* Cambridge university press.

Weaver, R., & Prelec, D. (2013). Creating truth-telling incentives with the Bayesian truth serum. *Journal of Marketing Research*, 289-302.

Appendix A - Results of the Pokémon Study

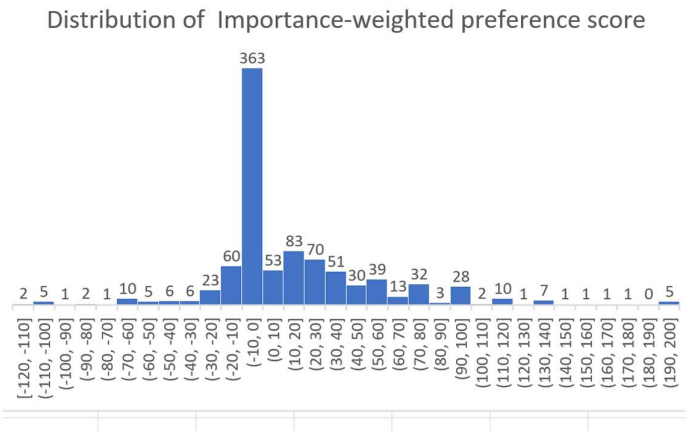

Figure A1: Distribution of Importance-weighted preference score in the Pokémon Study
Note: [-10,0] includes 341 zero-scores (no importance and/or no opinion/neutral on Likert preference scale) and 22 scores between and including -1 and -9. Range = [-120,200], since there were no scores lower than -120.

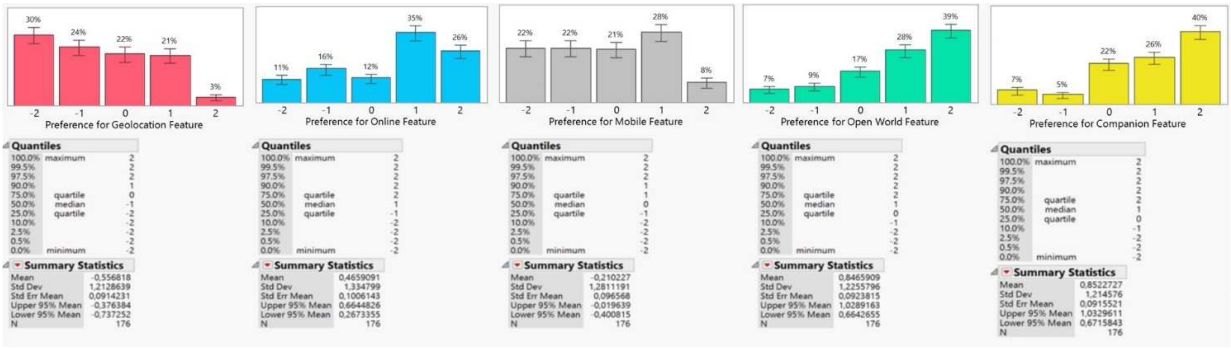

Figure A2: Distribution of Likert-scale preferences scores for potential features of Pokémon

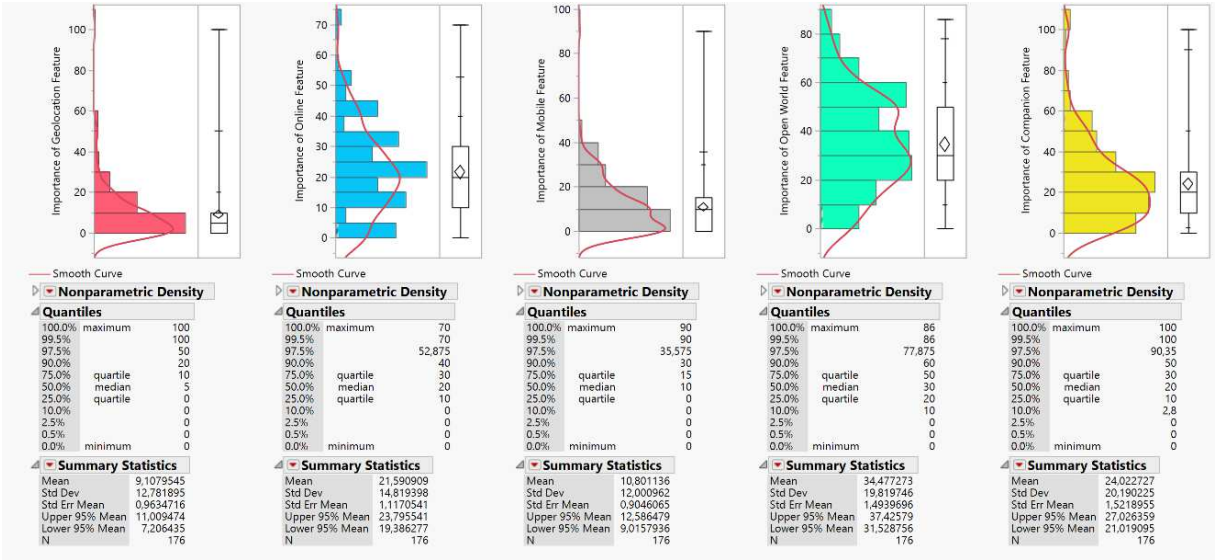

Figure A3: Distribution of Constant-sum importance scores for potential features of Pokémon

Appendix B - Methodology of the Pokémon Study

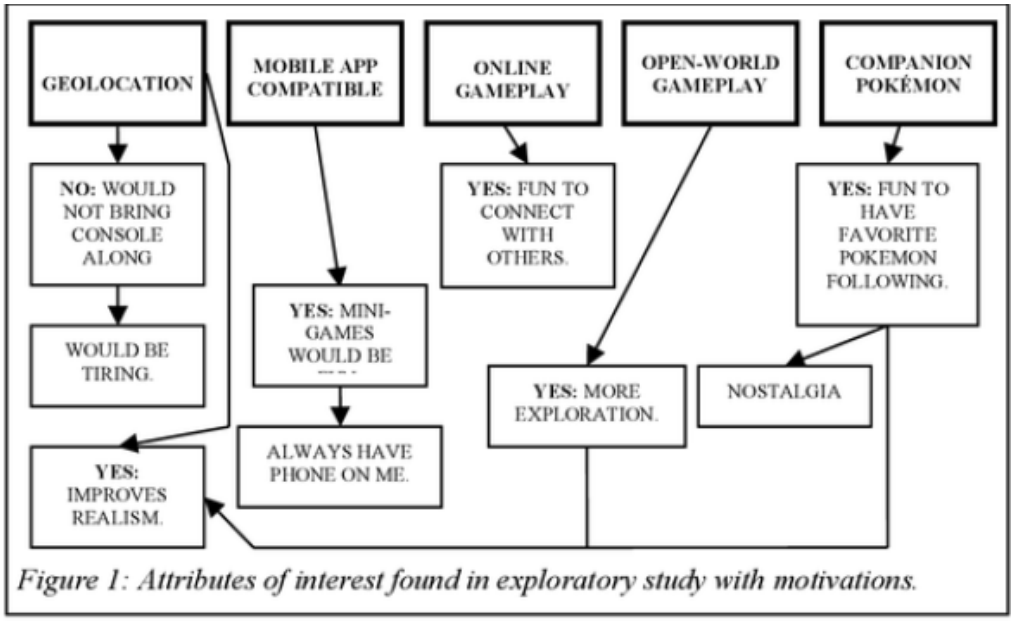

Figure B1: The five features that were selected in the Pokémon Study by Brouwer et al. as a result of the exploratory research

The five features

Geolocation	You predominantly play and interact in the real world
Online gameplay	You predominantly compete and interact with real players
Mobile app compatibility	There are earnable extras for the game. *Example: one could acquire extras by playing minigames on their phone.*
Open-world gameplay	You can freely roam around in the world, in contrast to (more linear and traditional) story-based gameplay
Companion Pokémon	Your favorite Pokémon follows you as you play the game

Appendix C - Survey Design

2019 Pokémon Game Survey

Survey Flow

Block: Start (7 Questions)

BlockRandomizer: 1 - Evenly Present Elements

- Block: Variant A (6 Questions)
- Block: Variant B (6 Questions)
- Standard: Variant C (19 Questions)

Block: End (2 Questions)

Page Break

Introductory Questions

Start of Block: Start

INTROSTART Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I would like to thank you very much for participating in this survey. Your input will certainly help me a lot!

Now, let's first answer some questions about how acquainted you are with Pokémon, before we start answering the other questions. If anything is unclear, please carefully read the instructions again. Pas zo nodig te taal van de vragenlijst aan (hierboven) en bevestig deze hier / choose your language:

- O Nederlands (1)
- O English (2)

INTRO1
Did you play Pokémon in the last five years?

- O Yes, on any of these Nintendo consoles (Game Boy, Gameboy Advance, DS, 3DS, N64, Gamecube, Wii, WiiU or Switch) or on an emulator AND ALSO Pokémon GO (8)
- O Yes, on any of these Nintendo consoles (Game Boy, Gameboy Advance, DS, 3DS, N64, Gamecube, Wii, WiiU or Switch) or on an emulator (but not Pokémon GO) (4)
- O Yes, Pokémon GO on my mobile phone (but not on any Nintendo console) (5)
- O No, but I played it (longer than five years ago) (6)
- O No, and I never have (7)

Display This Question:

If Did you play Pokémon in the last five years? = Yes, Pokémon GO on my mobile phone (but not on any Nintendo console)

Or Did you play Pokémon in the last five years? = No, but I played it (longer than five years ago)

INTRO2 If you have or would have a Nintendo Switch, would you be interested in playing a new main series Pokémon game on the Nintendo Switch in the near future?

- O Yes (10)
- O Maybe (11)
- O No (12)

Display This Question:

If If you have or would have a Nintendo Switch, would you be interested in playing a new main series... = Maybe

INTRO3 Would you want to learn more about the main series Pokémon games and the kind of gameplay that these games offer and would you like to express your opinion on what kind of features you would like and dislike in Pokémon games that are currently in development?

- O Yes (1)
- O No (2)

Display This Question:

If Did you play Pokémon in the last five years? = No, and I never have

Or If you have or would have a Nintendo Switch, would you be interested in playing a new main series... = No

Or Would you want to learn more about the main series Pokémon games and the kind of gameplay that th... = No

ENDDROPOUT According to your answers in the introductory part, you are unfortunately not acquainted enough with the Pokémon games to answer the questions in this survey. However, I would still like to thank you participating in this survey.

Even though you were not able to participate in this survey, you could still be of big help for my research. Therefore, I would like to ask you if you want to share this survey with friends that might be interested in the topic of this survey, Pokémon, and I would like to thank you in advance for doing this.

Skip To: End of Survey If According to your answers in the introductory part, you are unfortunately not acquainted enough w...() Is Displayed

Display This Question:

If Did you play Pokémon in the last five years? = Yes, on any of these Nintendo consoles (Game Boy, Gameboy Advance, DS, 3DS, N64, Gamecube, Wii, WiiU or Switch) or on an emulator (but not Pokémon GO)

Or Did you play Pokémon in the last five years? = Yes, on any of these Nintendo consoles (Game Boy, Gameboy Advance, DS, 3DS, N64, Gamecube, Wii, WiiU or Switch) or on an emulator AND ALSO Pokémon GO

Or If you have or would have a Nintendo Switch, would you be interested in playing a new main series... = Yes

Or Would you want to learn more about the main series Pokémon games and the kind of gameplay that th... = Yes

INTRO4 Which Pokémon Games did ever you play? Choose all categories of which you played at least one of the games.

- ☐ Pokémon Red/Blue/Yellow/Gold/Silver/Crystal/Ruby/Sapphire (1)
- ☐ Pokémon Emerald/Diamond/Pearl/Platinum/Black/White (2)
- ☐ Pokémon Black2/White2/X/Y/Sun/Moon/UltraSun/UltraMoon (3)
- ☐ Pokémon Mystery Dungeon (any) / Pokémon Ranger (any) (4)
- ☐ Pokémon GO (5)
- ☐ One or more Pokémon games that are not listed here (6)
- ☐ None of these (7)

INTROEND Thanks for answering the initial questions in the first part of this survey, please continue to the main part of the survey.

End of Block: Start

Variant A

Start of Block: Variant A

Display This Question:

IfDevice TypeIsMobile

INTROA1MOB In this second and final part of the survey, you will be able to express your opinion regarding five different features that could potentially be included in a new main series Pokémon game that is to be released in 2019. Before this new main series game is released, another Pokémon game will become available later this year, called Pokémon, Lets Go! Pikachu & Eevee. This game will be different compared to the Pokémon games we are used to, and therefore introduce many features that have never been seen before in a Pokémon game or are returning to the games after a long period of absence.

Imagine it's now up to you to decide which new features from the Lets Go games you would want to see in the 2019 main series Pokémon game. I have selected five features which are clearly shown in the first trailer for the Let's Go games. **Please watch this trailer and read the instructions before you start answering the questions.**

Display This Question:

IfDevice TypeIs NotMobile

INTROA1PC In this second and final part of the survey, you will be able to express your opinion regarding five different features that could potentially be included in a new main series Pokémon game that is to be released in 2019. Before this new main series game is released, another Pokémon game will become available later this year, called Pokémon, Lets Go! Pikachu & Eevee. This game will be different compared to the Pokémon games we are used to, and therefore introduce many features that have never been seen before in a Pokémon game or are returning to the games after a long period of absence.

Imagine it's now up to you to decide which new features from the Lets Go games you would want to see in the 2019 main series Pokémon game. I have selected five features which are clearly shown in the first trailer for the Let's Go games. **Please watch this trailer and read the instructions before you start answering the questions:**

Display This Question:

IfDevice TypeIs NotMobile

INTROA2

INTROA3

Please read this instruction carefully. *You can click the links for 'wild Pokémon battles' and 'random encounters', if you are not familiar with these concepts.*

The five features I have selected are (with their timestamps for the Let's Go trailer):

Wild Pokémon catching (instead of wild Pokémon battles): Instead of having wild Pokémon battles, Pokémon can be caught using motion controls (throwing a Poké ball), which is demonstrated from 0:40 to 1:00

Pokémon encounters in overworld (instead of random encounters): Rather than through random encounters (where you randomly encounter Pokémon, for example when walking through grass), Pokémon are encountered in the overworld (where you see a specific Pokémon move on the map and can walk into them to start an encounter), which is shown at 0:43, 1:09 and 1:34

Play together: You can play together with a friend in a local co-op mode, which is demonstrated from 1:01 to 1:28

Pokémon GO integration: You can transfer Pokémon from Pokémon GO to your Nintendo Switch game, which is demonstrated from 1:54 to 2:08

Following Pokémon: Your favorite Pokémon can follow you around and you can ride on them, which is demonstrated from 2:19 to 2:22

Now, I would like to ask you for your opinion regarding these features, would you like or dislike it if they were (also) to be included in the new 2019 Pokémon game?

Please state to what extent you agree/disagree with the following statements and feel free to return to this explanation if anything is unclear about the questions or features.

QA1
The described feature should be included in the new 2019 Pokémon game.

	Strongly disagree (1)	Somewhat disagree (2)	Neither agree nor disagree (3)	Somewhat agree (4)	Strongly agree (5)
Wild Pokémon catching (instead of wild Pokémon battles) (1)	O	O	O	O	O
Pokémon encounters in overworld (instead of random encounters) (2)	O	O	O	O	O
Play together (3)	O	O	O	O	O
Pokémon GO integration (4)	O	O	O	O	O
Following Pokémon (5)	O	O	O	O	O

QA2 Now, additionally, I would like to ask you to express how important these features are to you.

Below are the five features that could potentially be included in the new 2019 Pokémon game. Please distribute 100 points among these features, according to how important they are to you. This is, regardless of whether you would like or dislike them to be included.

Wild Pokémon catching (instead of wild Pokémon battles) : _____ (1)
Pokémon encounters in the overworld (instead of random encounters) : _____ (2)
Play Together : _____ (3)
Pokémon GO integration : _____ (4)
Following Pokémon : _____ (5)

Total : _____

End of Block: Variant A

Variant B

Start of Block: Variant B

Display This Question:

IfDevice TypeIsMobile

INTROB1MOB In this second and final part of the survey, you will be able to express your opinion regarding five different features that could potentially be included in a new main series Pokémon game that is to be released in 2019. Before this new main series game is released, another Pokémon game will become available later this year, called Pokémon, Lets Go! Pikachu & Eevee. This game will be different compared to the Pokémon games we are used to, and therefore introduce many features that have never been seen before in a Pokémon game or are returning to the games after a long period of absence.

Imagine it's now up to you to decide which new features from the Lets Go games you would want to see in the 2019 main series Pokémon game. I have selected five features which are clearly shown in the first trailer for the Let's Go games. **Please watch this trailer and read the instructions before you start answering the questions.**

Display This Question:

IfDevice TypeIs NotMobile

INTROB1PC In this second and final part of the survey, you will be able to express your opinion regarding five different features that could potentially be included in a new main series Pokémon game that is to be released in 2019. Before this new main series game is released, another Pokémon game will become available later this year, called Pokémon, Lets Go! Pikachu & Eevee. This game will be different compared to the Pokémon games we are used to, and therefore introduce many features that are never seen before in a Pokémon game or are returning to the games after a long period of absence.

Imagine it's now up to you, to decide which new features from the Lets Go games, you would want to see in the 2019 main series Pokémon game. I have selected five features, which are clearly shown in the first trailer for the Let's Go games. **Please watch this trailer and read the instructions before you start answering the questions:**

Display This Question:

IfDevice TypeIs NotMobile

INTROB2PC

INTROB3

Please read this instruction carefully. *You can click the links for 'wild Pokémon battles' and 'random encounters', if you are not familiar with these concepts.*

The five features I have selected are (with their timestamps for the trailer):

Wild Pokémon catching (instead of wild Pokémon battles): Instead of having wild Pokémon battles, Pokémon can be caught using motion controls (throwing a Poké ball), which is demonstrated from 0:40 to 1:00

Pokémon encounters in overworld (instead of random encounters): Rather than through random encounters (where you randomly encounter Pokémon, for example when walking through grass), Pokémon are encountered in the overworld (where you see a specific Pokémon move on the map and can walk into them to start an encounter), which is shown at 0:43, 1:09 and 1:34

Play together: You can play together with a friend in the new local co-op mode, which is demonstrated from 1:01 to 1:28

Pokémon GO integration: You can transfer Pokémon from Pokémon GO to your Nintendo Switch game, which is demonstrated from 1:54 to 2:08

Following Pokémon: Your favorite Pokémon can follow you around and you can ride on them, which is demonstrated from 2:19 to 2:22

Now, I would like to ask you for your opinion regarding these features, would you like or dislike it if they were (also) to be included in the new 2019 Pokémon game?

Please state to what extent you agree/disagree with the following statements and feel free to return to this explanation if anything is unclear about the questions or features.

Display This Question:

If Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I wou... = Nederlands

QB1NL

Please state for each of the five features, to what extent you would (not) like them to (also) be included in the new 2019 Pokémon game. The three labels indicate the two extreme responses, -40 and 40 for 'The feature should definitely not be included' and 'The feature should definitely be included' respectively, and the 'middle' response 0 for 'I don't care whether the feature is included in the game or not'.

Please respond according to your opinion and feel free to choose any spot on the slider that captures your opinion regarding the feature as accurately as possible.

- - - - - - - -5 0 5 10 15 20 25 30 35 40
40 35 30 25 20 15 10

Wild Pokémon catching (instead of wild Pokémon battles) ()	
Pokémon encounters in overworld (instead of random encounters) ()	
Play together ()	
Pokémon GO integration ()	
Following Pokémon ()	

Display This Question:

If Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I wou... = English

QB1ENG

Please state for each of the five features, to what extent you would (not) like them to (also) be included in the new 2019 Pokémon game. The three labels indicate the two extreme responses, -40 and 40 for 'The feature should definitely not be included' and 'The feature should definitely be included' respectively, and the 'middle' response 0 for 'I don't care whether the feature is included in the game or not'.

Please respond according to your opinion and feel free to choose any spot on the slider that captures your opinion regarding the feature as accurately as possible.

	- - - - - - - -5 0 5 10 15 20 25 30 35 40 40 35 30 25 20 15 10
Wild Pokémon catching (instead of wild Pokémon battles) ()	
Pokémon encounters in overworld (instead of random encounters) ()	
Play together ()	
Pokémon GO integration ()	
Following Pokémon ()	

End of Block: Variant B

Variant C

Start of Block: Variant C

Display This Question:

IfDevice TypeIsMobile

INTROC1MOB In this second and final part of the survey, you will be able to express your opinion regarding five different features that could potentially be included in a new main series Pokémon game that is to be released in 2019. Before this new main series game is released, another Pokémon game will become available later this year, called Pokémon, Lets Go! Pikachu & Eevee. This game will be different compared to the Pokémon games we are used to, and therefore introduce many features that have never been seen before in a Pokémon game or are returning to the games after a long period of absence.

Imagine it's now up to you to decide which new features from the Lets Go games you would want to see in the 2019 main series Pokémon game. I have selected five features which are clearly shown in the first trailer for the Let's Go games. **Please watch this trailer and read the instructions before you start answering the questions.**

Display This Question:

IfDevice TypeIs NotMobile

INTROC1PC In this second and final part of the survey, you will be able to express your opinion regarding five different features that could potentially be included in a new main series Pokémon game that is to be released in 2019. Before this new main series game is released, another Pokémon game will become available later this year, called Pokémon, Lets Go! Pikachu & Eevee. This game will be different compared to the Pokémon games we are used to and therefore introduce many features that have never been seen before in a Pokémon game or are returning to the games after a long period of absence.

Imagine it's now up to you to decide which new features from the Lets Go games you would want to see in the 2019 main series Pokémon game. I have selected five features which are clearly shown in the first trailer for the Let's Go games. **Please watch this trailer and read the instructions before you start answering the questions:**

Display This Question:

If Device Type Is Not Mobile

INTROC2PC

INTROC3

Please read this instruction carefully. *You can click the links for 'wild Pokémon battles' and 'random encounters', if you are not familiar with these concepts.*

The five features I have selected are (with their timestamps for the trailer):

Wild Pokémon catching (instead of wild Pokémon battles): Instead of having wild Pokémon battles, Pokémon can be caught using motion controls (throwing a Poké ball), which is demonstrated from 0:40 to 1:00

Pokémon encounters in overworld (instead of random encounters): Rather than through random encounters (where you randomly encounter Pokémon, for example when walking through grass), Pokémon are encountered in the overworld (where you see a specific Pokémon move on the map and can walk into them to start an encounter), which is shown at 0:43, 1:09 and 1:34

Play together: You can play together with a friend in the new local co-op mode, which is demonstrated from 1:01 to 1:28

Pokémon GO integration: You can transfer Pokémon from Pokémon GO to your Nintendo Switch game, which is demonstrated from 1:54 to 2:08

Following Pokémon: Your favorite Pokémon can follow you around and you can ride on them, which is demonstrated from 2:19 to 2:22

Now, I would like to ask you for two things regarding each of these five features:

1. Your opinion regarding these features, would you like or dislike it if they were (also) to be included in the new 2019 Pokémon game?

2. How you predict that other participants of this survey will evaluate the five features.

At the end of the survey, some participants will be selected randomly and will be *rewarded financially*. The amount that is rewarded to those selected, will depend on a *score* that consists of two factors:

A. How well you predicted the opinions of other participants

B. How well others that expressed opinions similar to yours, predicted the opinions of other participants

Please read this next section very carefully:

Note that this mechanism has two effects if you are selected to be paid. First, it rewards you for the accuracy of your own predictions about the choices of other participants. Second, it rewards you for the accuracy of the predictions of others who expressed similar opinions as you. Note that this means that it is in your best interest to state your opinions truthfully. This is because you possess valuable information that nobody else does: **you know your own honest opinion.** The details are a bit complicated, but can be shown mathematically that you maximize your chance to receive a high reward if you make use of this information (that you have). According to how well your predictions and those of others like you are, you might be able to win a maximum of up till EUR20/USD23 (if you are selected, I will reach out to you to ask how you would like to get paid out).

Now, let us continue to the questions. *Please state for each of the five features, to what extent you would (not) like them to (also) be included in the new 2019 Pokémon game*. The three labels indicate the two extreme responses, -40 and 40 for 'The feature should definitely not be included' and 'The feature should definitely be included' respectively, and the 'middle' response 0 for 'I don't care whether the feature is included in the game or not'. Furthermore, please state how you predict that others will evaluate these features.

Please respond according to your opinion and feel free to choose any spot on the slider that captures your opinion regarding the feature as accurately as possible.

Display This Question:

If Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I wou... = Nederlands

QC1NL Please state for this feature, to what extent you would (not) like them to be included in the new 2019 Pokémon game.

	- - - - - - - -40 35 30 25 20 15 10 -5 0 5 10 15 20 25 30 35 40
Wild Pokémon catching (instead of wild Pokémon battles) ()	

Display This Question:

If Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I wou... = English

QC1ENG Please state for this feature, to what extent you would (not) like them to be included in the new 2019 Pokémon game.

	- - - - - - - - -5 0 5 10 15 20 25 30 35 40 40 35 30 25 20 15 10
Wild Pokémon catching (instead of wild Pokémon battles) ()	

QC2 Out of 100 other participants in this survey, how many would you expect to rate the feature 'Wild Pokémon catching (instead of wild Pokémon battles)' within these ranges?

Feature should definitely not be included (between -40 and -24) : _____ (1)
Rather not have the feature be included (between -24 and -8) : _____ (2)
Don't care much whether the feature is included or not (between -8 and 8) : _____ (3)
Would be nice if feature is included (between 8 and 24) : _____ (4)
Feature should definitely be included (between 24 and 40) : _____ (5)

Total : _____

Display This Question:

If Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I wou... = Nederlands

QC3NL Please state for this feature, to what extent you would (not) like them to be included in the new 2019 Pokémon game.

	- - - - - - - - -5 0 5 10 15 20 25 30 35 40 40 35 30 25 20 15 10
Pokémon encounters in overworld (instead of random encounters) ()	

Display This Question:

If Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I wou... = English

QC3ENG Please state for this feature, to what extent you would (not) like them to be included in the new 2019 Pokémon game.

- - - - - - - -5 0 5 10 15 20 25 30 35 40
40 35 30 25 20 15 10

Pokémon encounters in overworld (instead of random encounters) ()	

QC4 Out of 100 other participants in this survey, how many would you expect to rate the feature 'Pokémon encounters in overworld (instead of random encounters)' within these ranges?

Feature should definitely not be included (between -40 and -24) : _____ (1)
Rather not have the feature be included (between -24 and -8) : _____ (2)
Don't care much whether the feature is included or not (between -8 and 8) : _____ (3)
Would be nice if feature is included (between 8 and 24) : _____ (4)
Feature should definitely be included (between 24 and 40) : _____ (5)

Total : _____

Display This Question:

If Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I wou... = Nederlands

Q5NL Please state for this feature, to what extent you would (not) like them to be included in the new 2019 Pokémon game.

- - - - - - - -5 0 5 10 15 20 25 30 35 40
40 35 30 25 20 15 10

Play together ()	

Display This Question:

If Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I wou... = English

Q5ENG Please state for this feature, to what extent you would (not) like them to be included in the new 2019 Pokémon game.

- - - - - - - -5 0 5 10 15 20 25 30 35 40
40 35 30 25 20 15 10

Play together ()	

QC6 Out of 100 other participants in this survey, how many would you expect to rate the feature 'Play together' within these ranges?

Feature should definitely not be included (between -40 and -24) : _____ (1)
Rather not have the feature be included (between -24 and -8) : _____ (2)
Don't care much whether the feature is included or not (between -8 and 8) : _____ (3)
Would be nice if feature is included (between 8 and 24) : _____ (4)
Feature should definitely be included (between 24 and 40) : _____ (5)

Total : _____

Display This Question:

If Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I wou... = Nederlands

QC7NL Please state for this feature, to what extent you would (not) like them to be included in the new 2019 Pokémon game.

- - - - - - - -5 0 5 10 15 20 25 30 35 40
40 35 30 25 20 15 10

Pokémon GO integration ()	

Display This Question:

If Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I wou... = English

QC7ENG Please state for this feature, to what extent you would (not) like them to be included in the new 2019 Pokémon game.

- - - - - - - -5 0 5 10 15 20 25 30 35 40
40 35 30 25 20 15 10

Pokémon GO integration ()	

QC8 Out of 100 other participants in this survey, how many would you expect to rate the feature 'Pokémon GO integration' within these ranges?

Feature should definitely not be included (between -40 and -24) : _____ (1)
Rather not have the feature be included (between -24 and -8) : _____ (2)
Don't care much whether the feature is included or not (between -8 and 8) : _____ (3)
Would be nice if feature is included (between 8 and 24) : _____ (4)
Feature should definitely be included (between 24 and 40) : _____ (5)

Total : _____

Display This Question:

If Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I wou... = Nederlands

QC9NL Please state for this feature, to what extent you would (not) like them to be included in the new 2019 Pokémon game.

- - - - - - - -5 0 5 10 15 20 25 30 35 40
40 35 30 25 20 15 10

Following Pokémon ()	

Display This Question:

If Welcome to this survey about Pokémon, which is part of my bachelor thesis research project. I wou... = English

QC9ENG Please state for this feature, to what extent you would (not) like them to be included in the new 2019 Pokémon game.

- - - - - - - -5 0 5 10 15 20 25 30 35 40
40 35 30 25 20 15 10

Following Pokémon ()	

QC10 Out of 100 other participants in this survey, how many would you expect to rate the feature 'Following Pokémon':

Feature should definitely not be included (between -40 and -24) : _____ (1)
Rather not have the feature be included (between -24 and -8) : _____ (2)
Don't care much whether the feature is included or not (between -8 and 8) : _____ (3)
Would be nice if feature is included (between 8 and 24) : _____ (4)
Feature should definitely be included (between 24 and 40) : _____ (5)

Total : _____

End of Block: Variant C

End of Survey

Start of Block: End

END1 Thank you very much for finishing the survey. I would be very grateful and it would help me so much, if you could ask other Pokémon fans you know, to also fill out this survey.

To thank you all for your participation, I will (after I am done with the research) randomly select some participants and will reward those with a monetary payment. For that, I will need your email-adresses.

Of course, this information will only be used for reaching out to the selected participants and for nothing else. Furthermore, you are free not to enter your e-mailadress, if you don't want to.

If you want to participate in the draw, please indicate your emailadress here:

_____

END2 **Thanks again for your time and do not forget to share this survey among your friends.**

To finish this survey and record your responses and(/or) emailadress, please click the button below.

End of Block: End

Appendix D - Full results for the three variants

Variant A

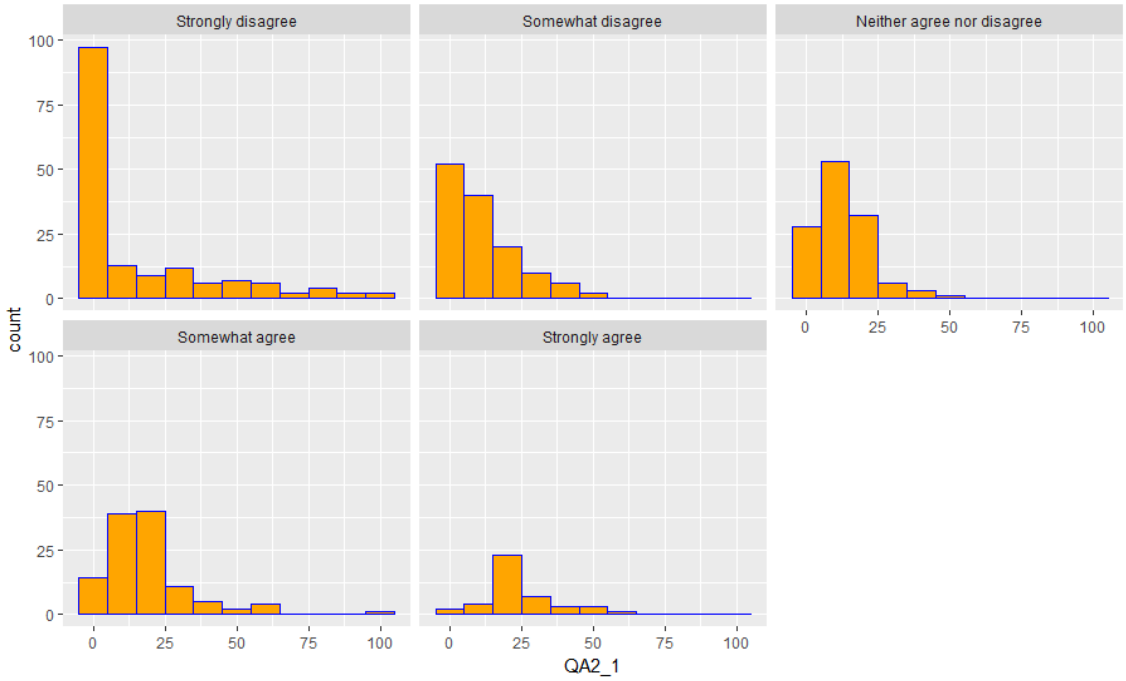

Figure D1: Distribution of the importance scores for each of the five Likert-scale preference categories for the feature 'Pokémon Catching (instead of Wild Pokémon Battles)'

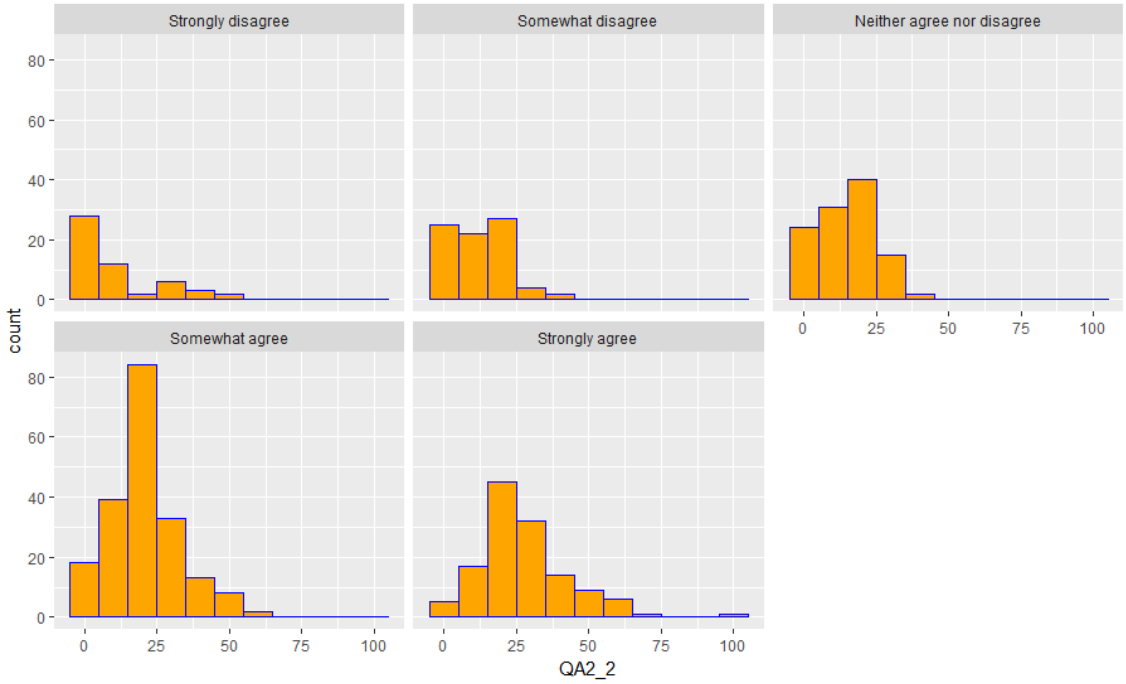

Figure D2: Distribution of the importance scores for each of the five Likert-scale preference categories for the feature 'Pokémon in overworld (instead of random encounters)'

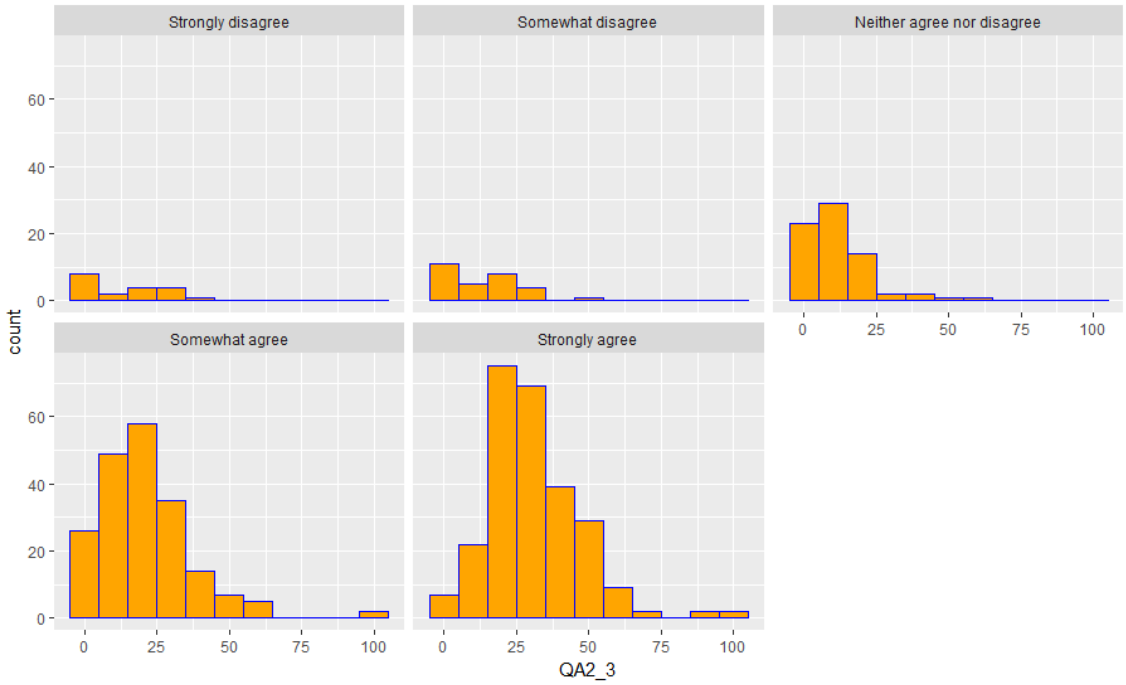

Figure D3: Distribution of the importance scores for each of the five Likert-scale preference categories for the feature 'Play Together'

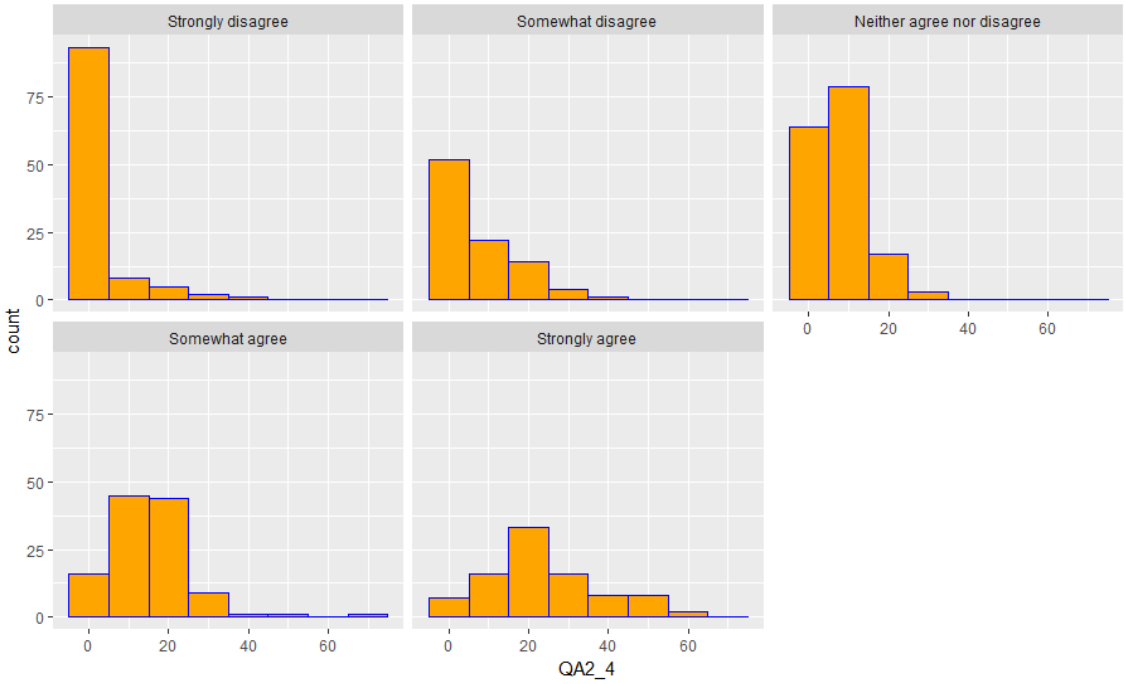

Figure D4: Distribution of the importance scores for each of the five Likert-scale preference categories for the feature 'Pokémon GO Integration'

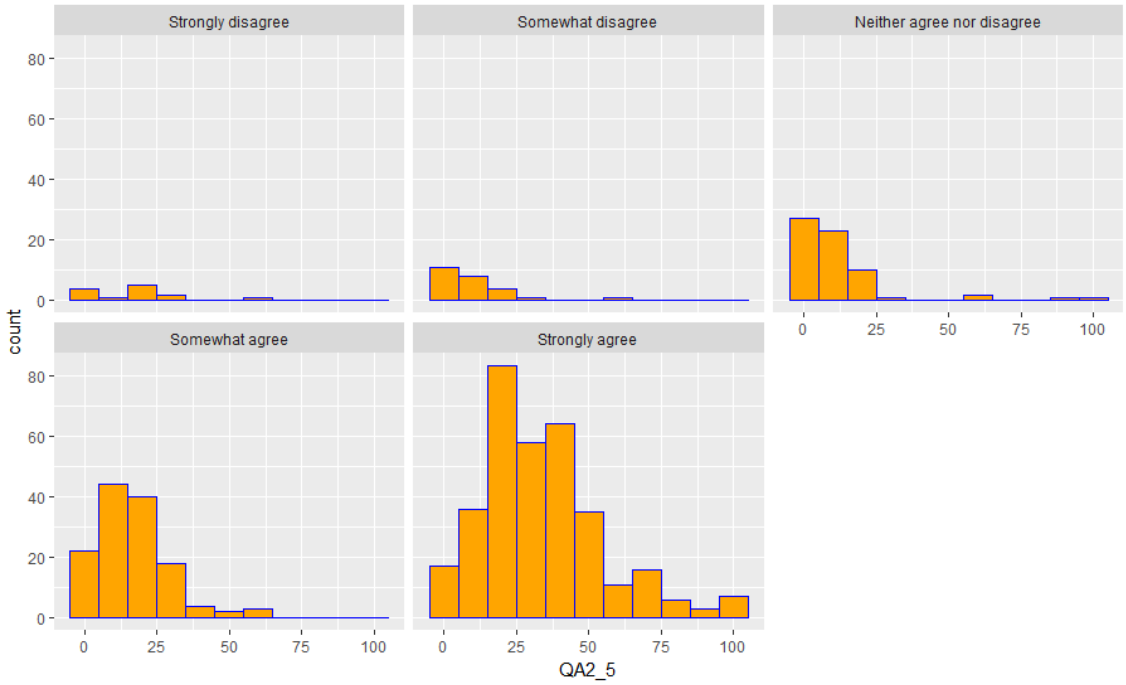

Figure D5: Distribution of the importance scores for each of the five Likert-scale preference categories for the feature 'Following Pokémon'

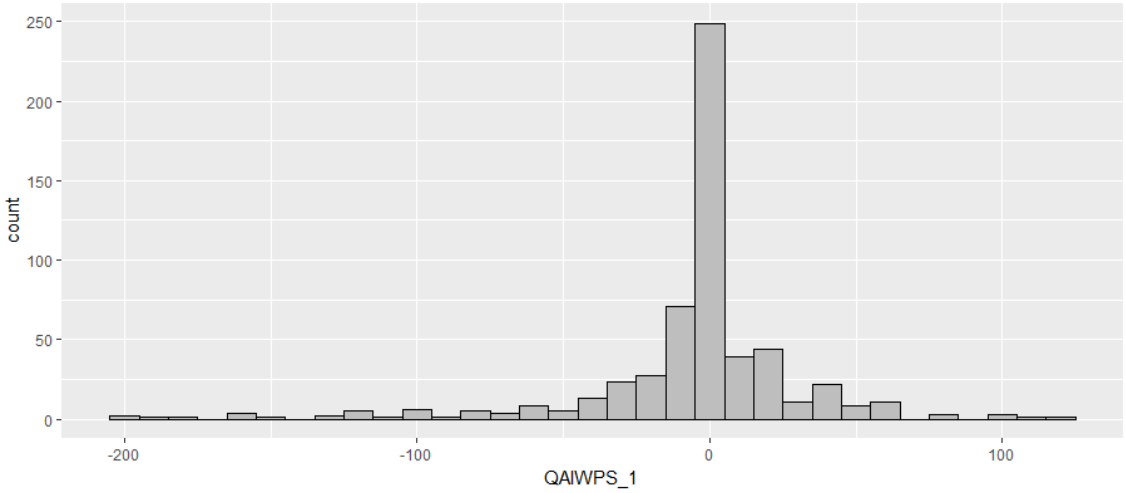

Figure D6: Histogram for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the feature 'Pokémon Catching (instead of Wild Pokémon Battles) ' in variant A

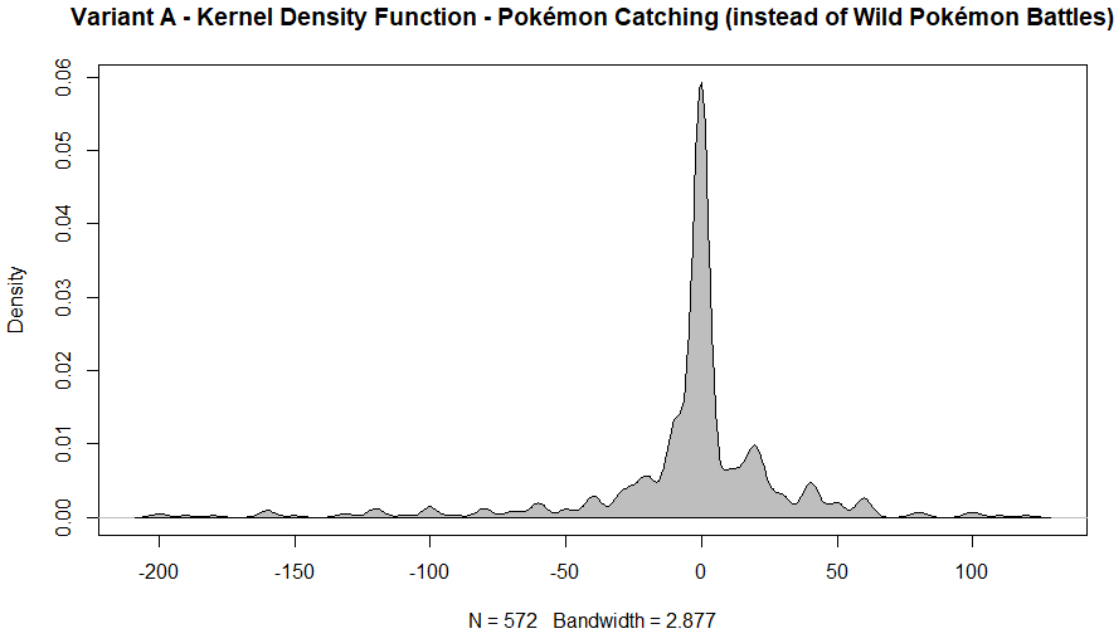

Figure D7: Kernel density plot for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the feature 'Pokémon Catching (instead of Wild Pokémon Battles)' in variant A

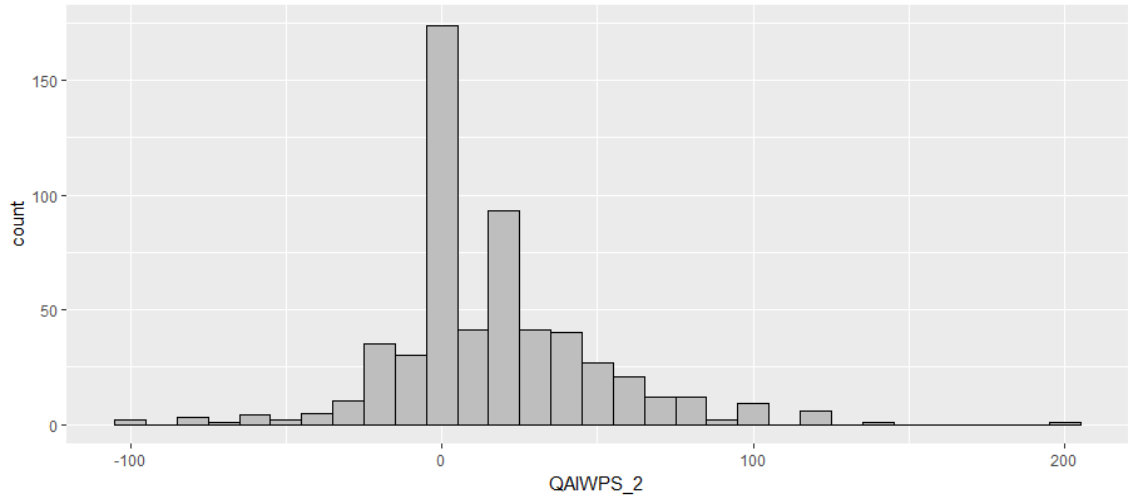

Figure D8: Histogram for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the feature 'Pokémon in overworld (instead of random encounters)' in variant A

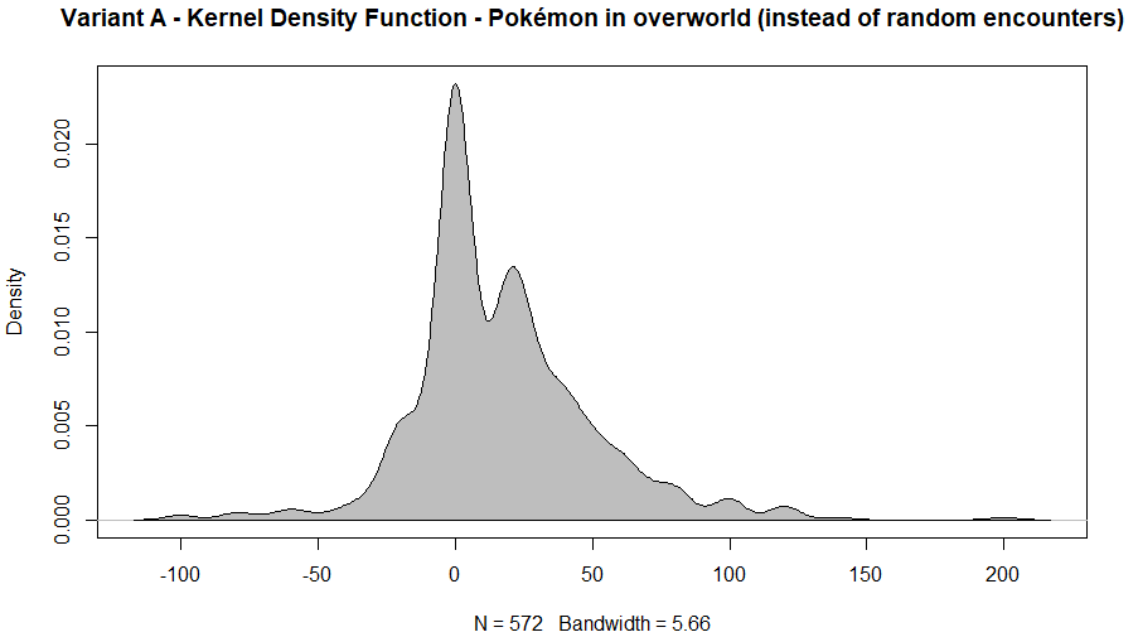

Figure D9: Kernel density plot for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the feature 'Pokémon in overworld (instead of random encounters)' in variant A

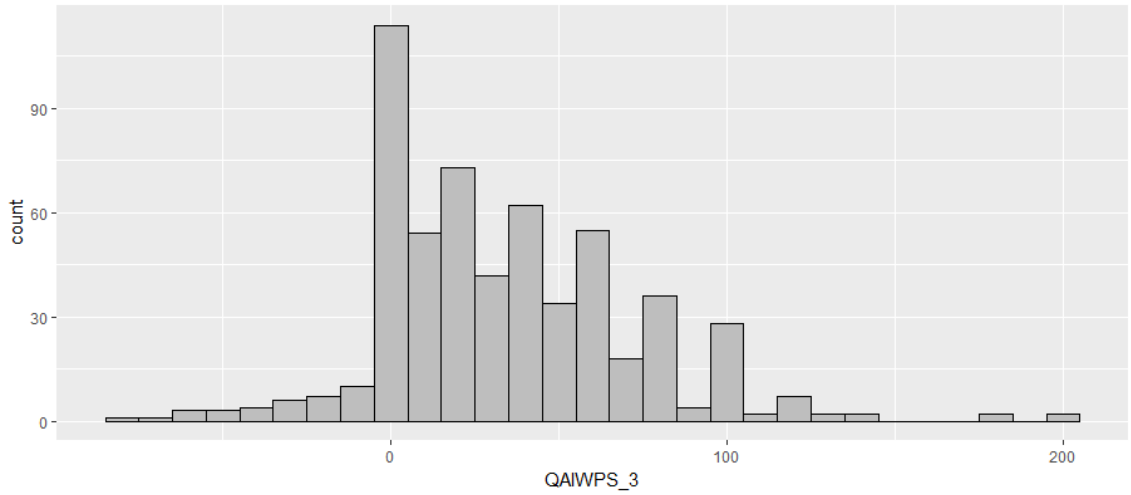

Figure D10: Histogram for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the feature 'Play Together' in variant A

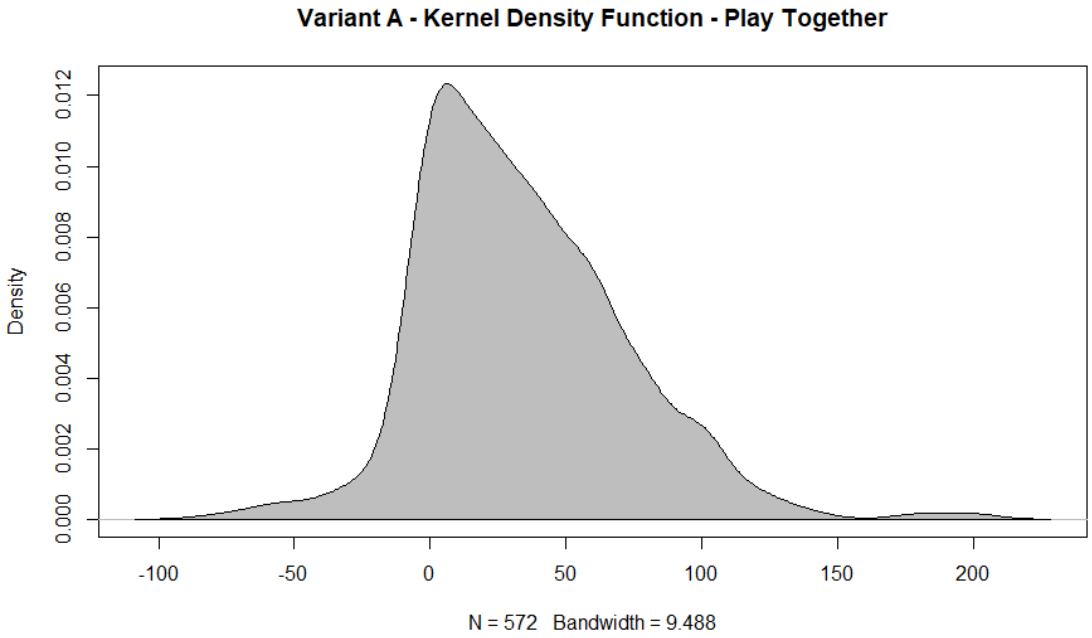

Figure D11: Kernel density plot for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the feature 'Play Together' in variant A

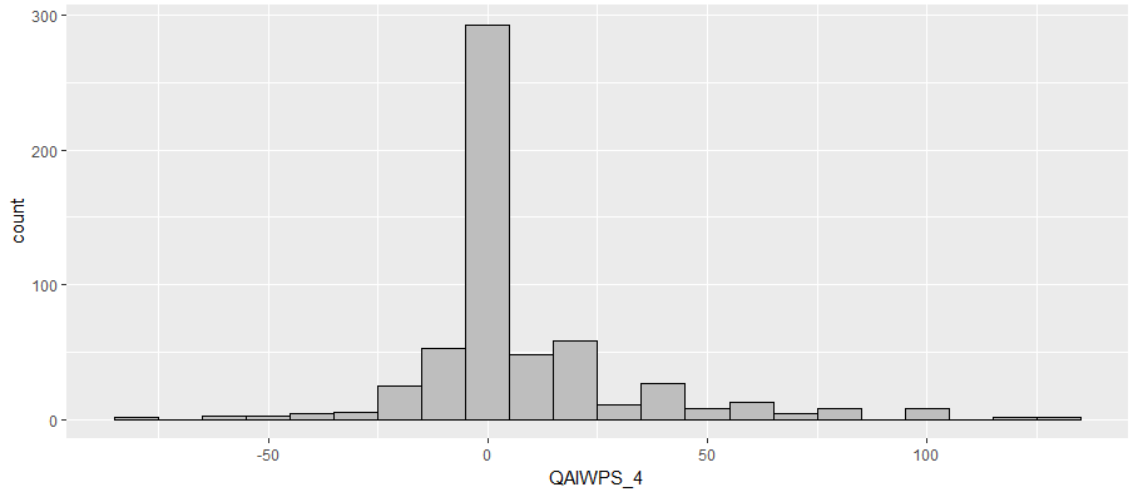

Figure D12: Histogram for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the feature 'Pokémon GO Integration' in variant A

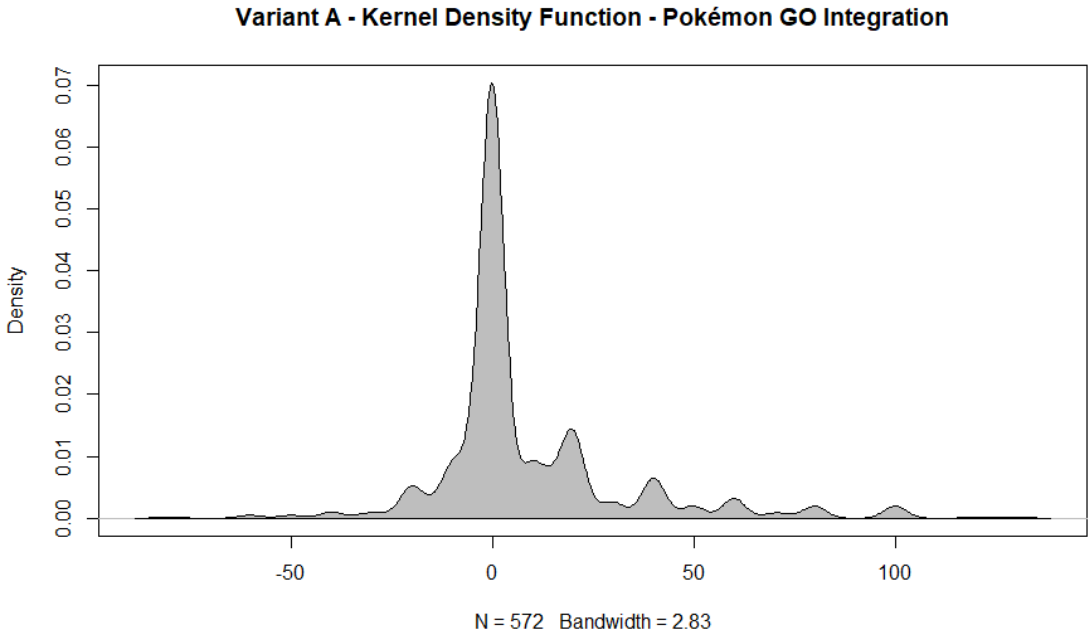

Figure D13: Kernel density plot for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the feature 'Pokémon GO Integration' in variant A

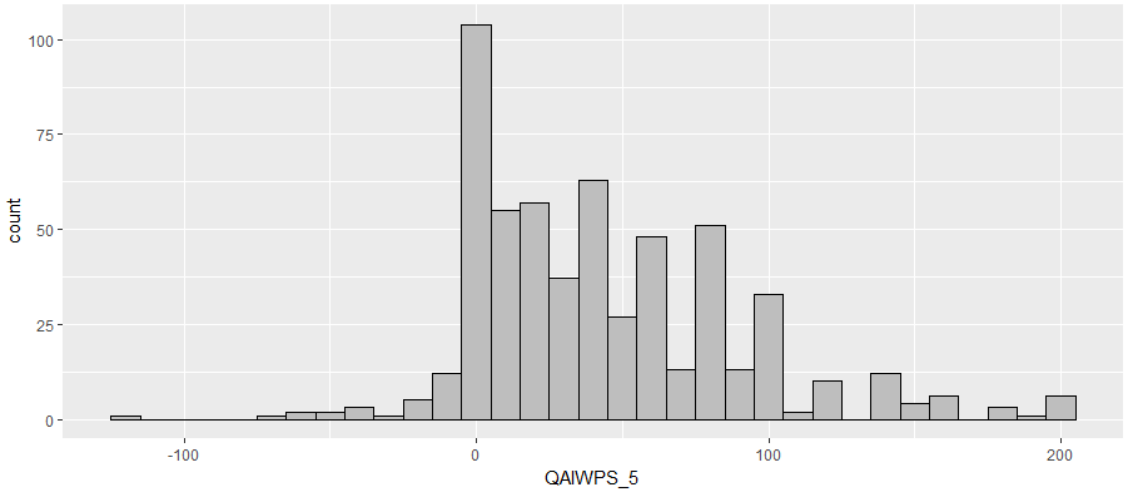

Figure D14: Histogram for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the feature 'Following Pokémon' in variant A

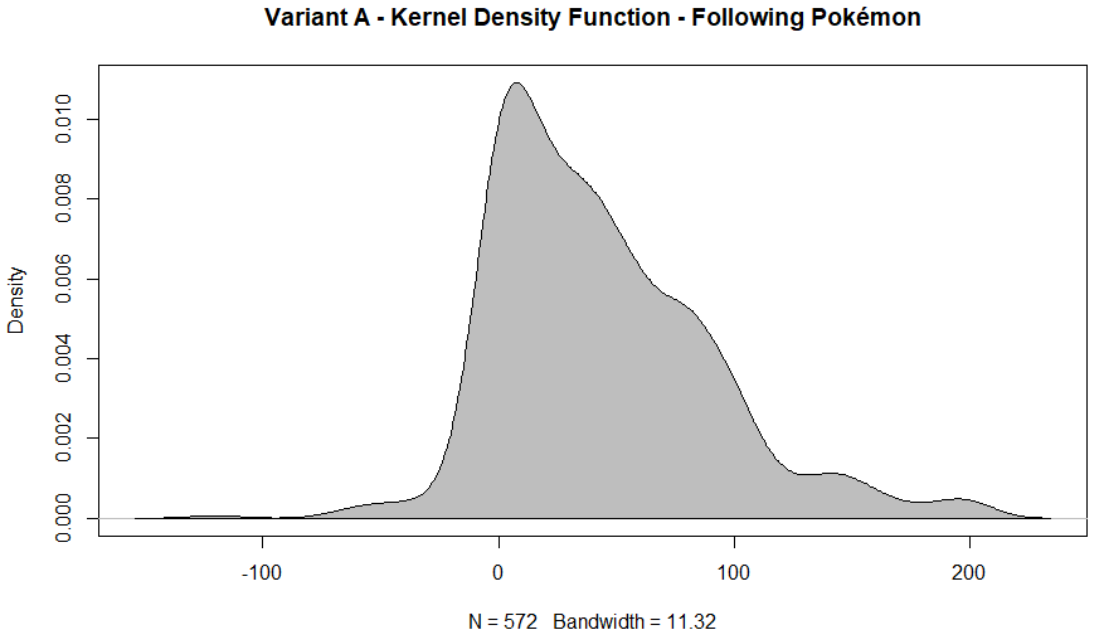

Figure D15: Kernel density plot for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the feature 'Following Pokémon' in variant A

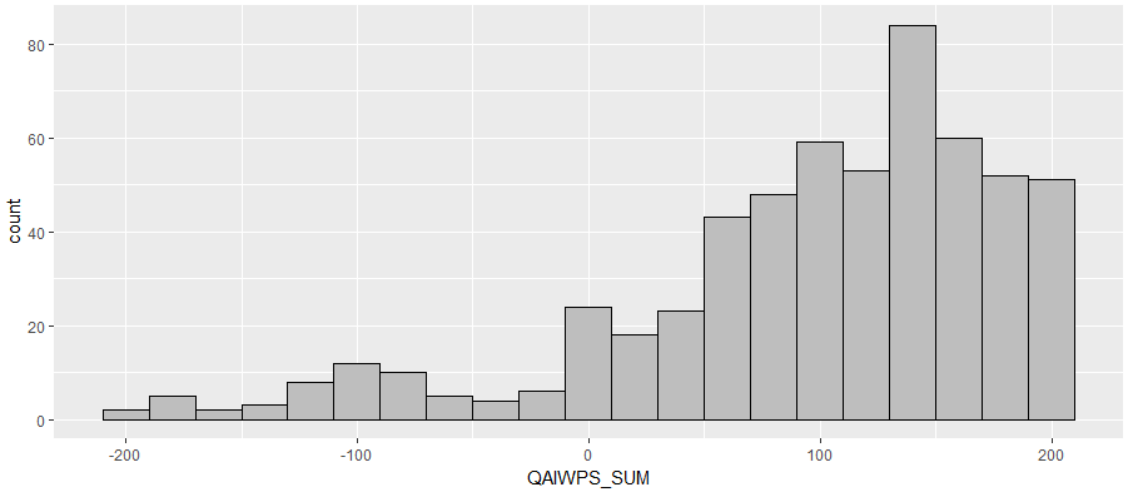

Figure D16: Histogram for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the five features in variant A

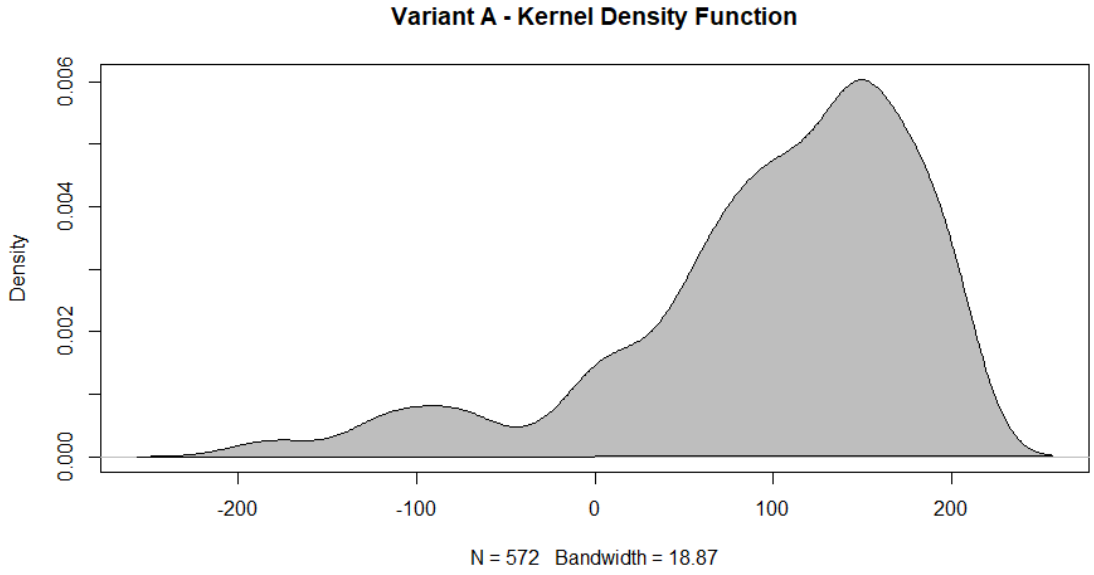

Figure D17: Kernel density plot for the distribution of the participants' sum of the Importance-Weighted Preference Scores for the five features in variant A

Variant B

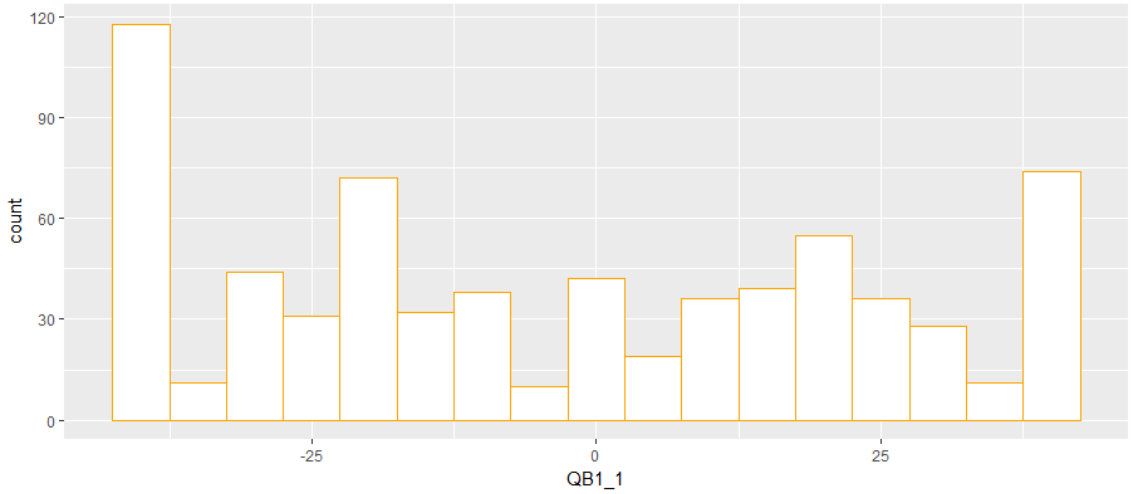

Figure D18: Histogram for the distribution of the participants' sum of the scores for the feature 'Pokémon Catching (instead of Wild Pokémon Battles)' in variant B

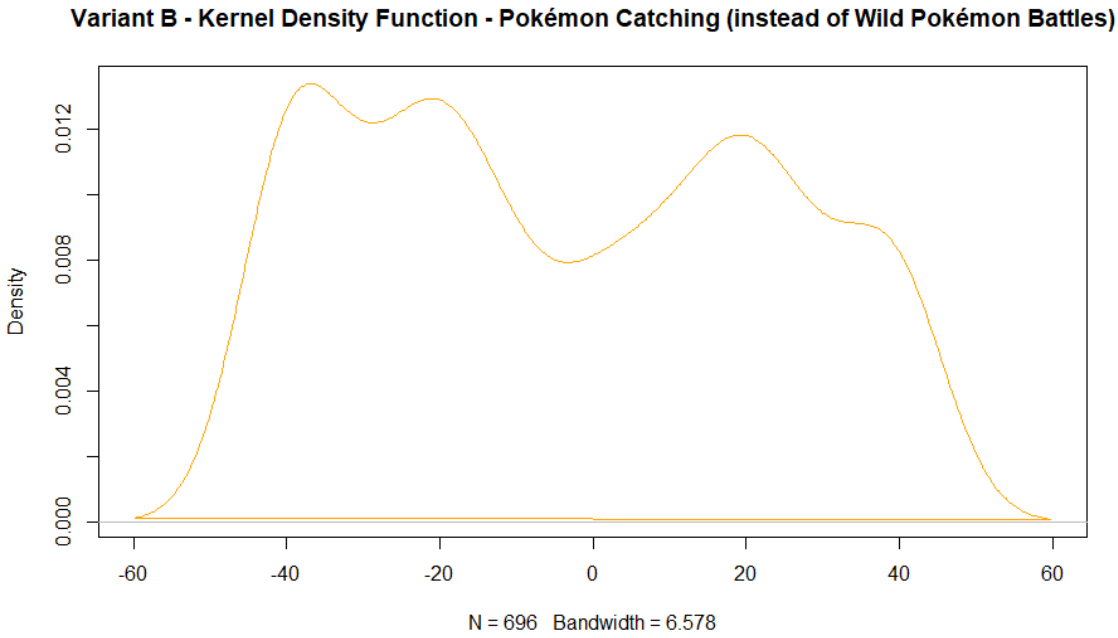

Figure D19: Kernel density plot for the distribution of the participants' sum of the scores for the feature 'Pokémon Catching (instead of Wild Pokémon Battles)' in variant B

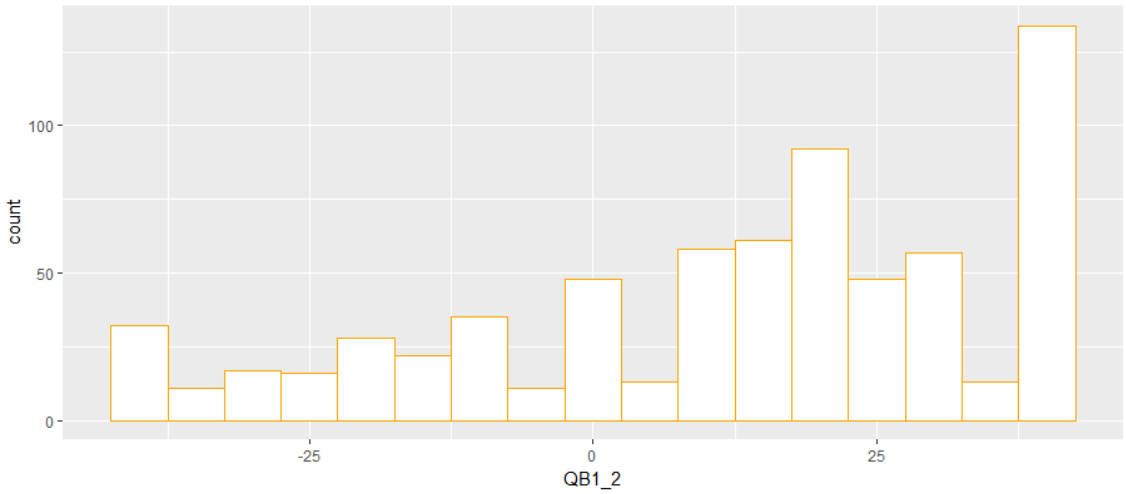

Figure D20: Histogram for the distribution of the participants' sum of the scores for the feature 'Pokémon in overworld (instead of random encounters)' in variant B

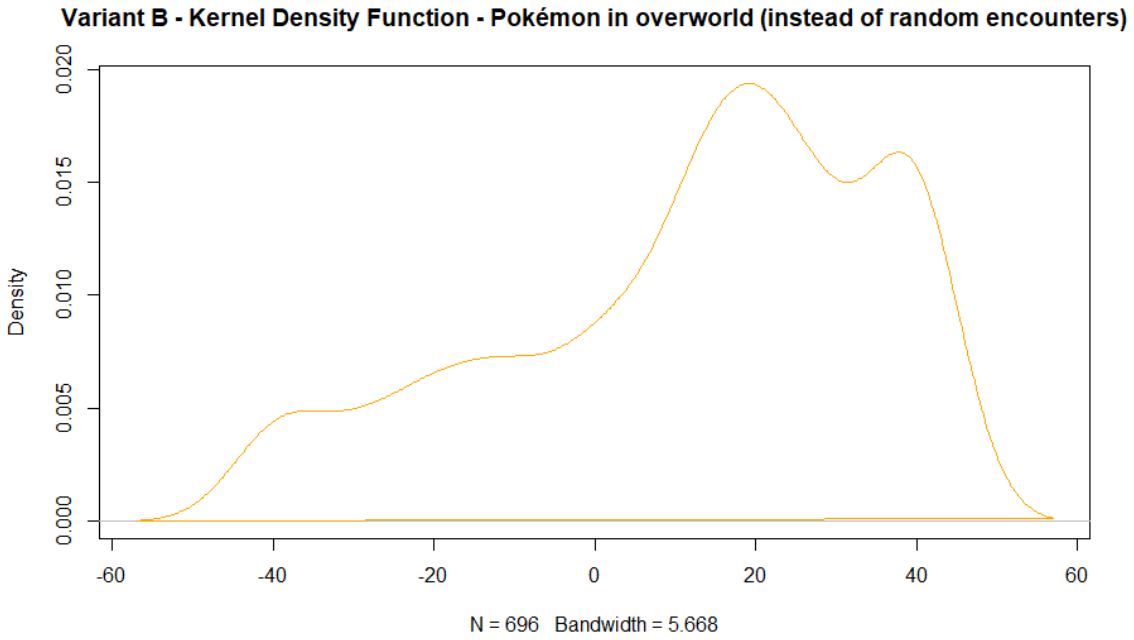

Figure D21: Kernel density plot for the distribution of the participants' sum of the scores for the feature 'Pokémon in overworld (instead of random encounters)' in variant B

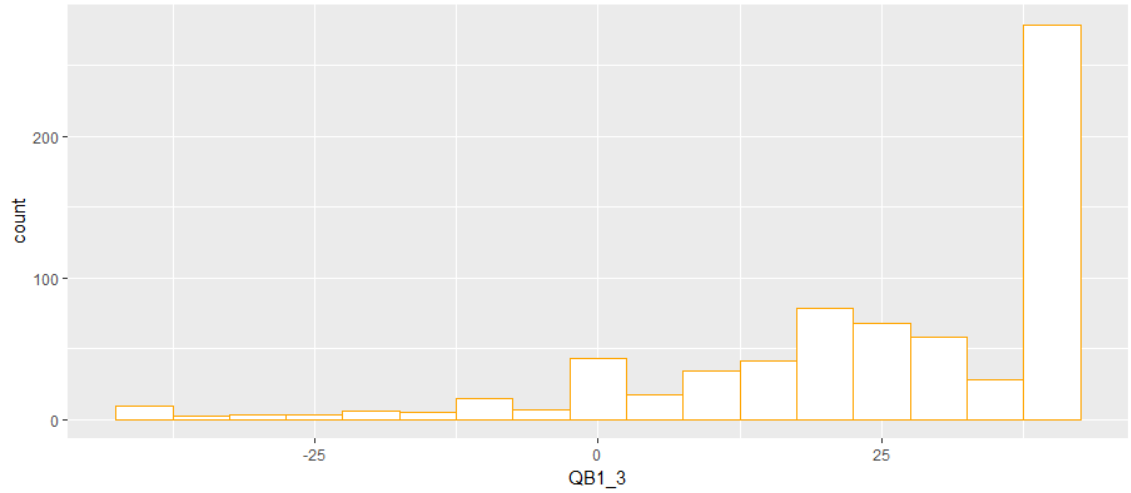

Figure D22: Histogram for the distribution of the participants' sum of the scores for the feature 'Play Together' in variant B

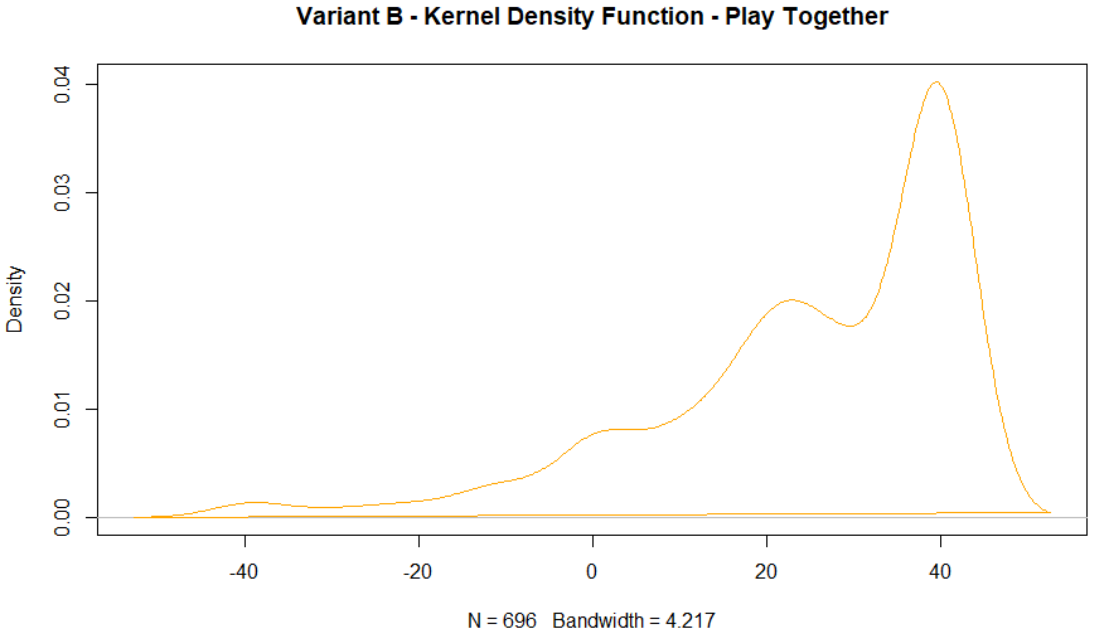

Figure D23: Kernel density plot for the distribution of the participants' sum of the scores for the feature 'Play Together' in variant B

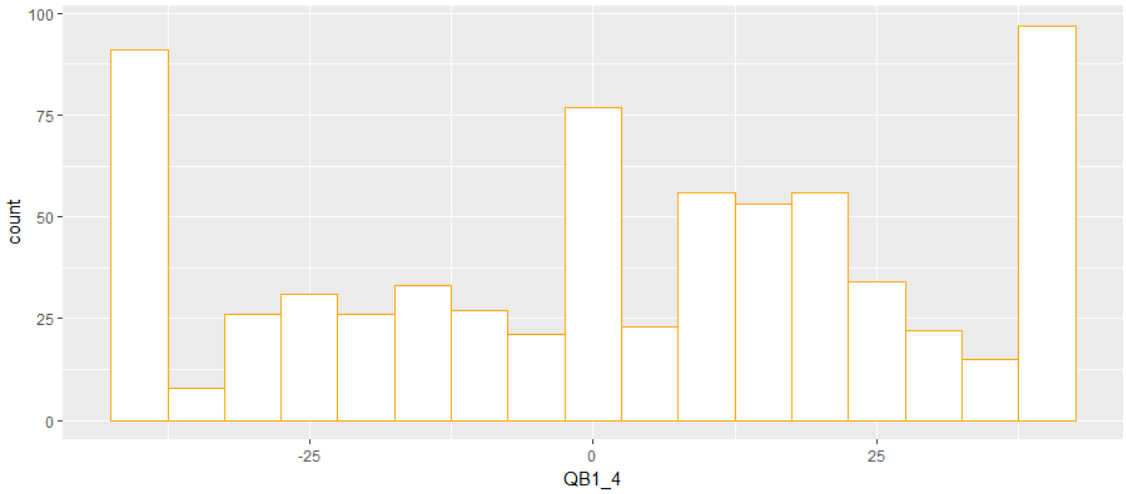

Figure D24: Histogram for the distribution of the participants' sum of the scores for the feature 'Pokémon GO Integration' in variant B

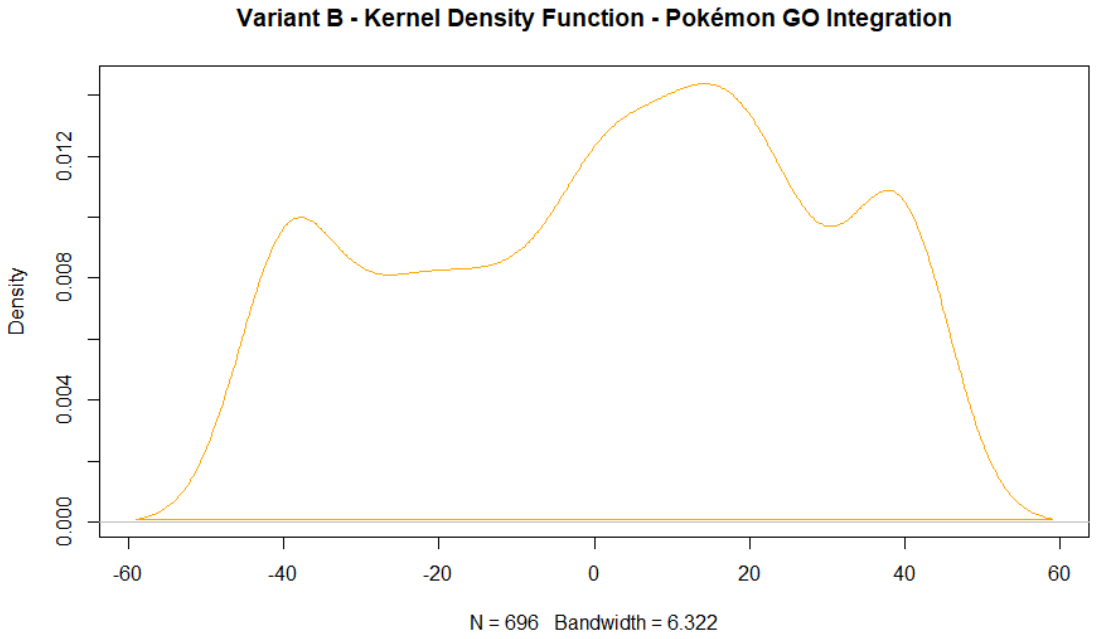

Figure D25: Kernel density plot for the distribution of the participants' sum of the scores for the feature 'Pokémon GO Integration' in variant B

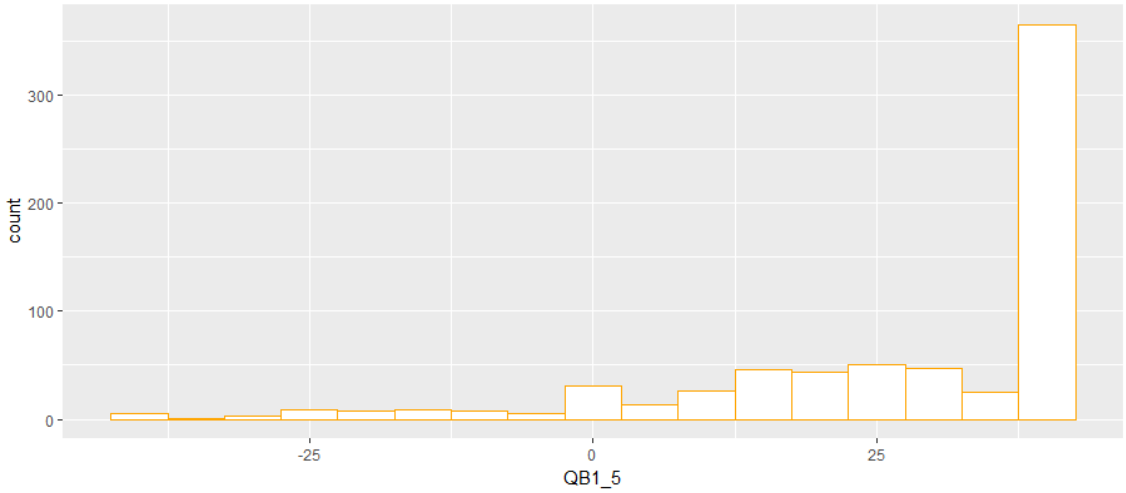

Figure D26: Histogram for the distribution of the participants' sum of the scores for the feature 'Following Pokémon' in variant B

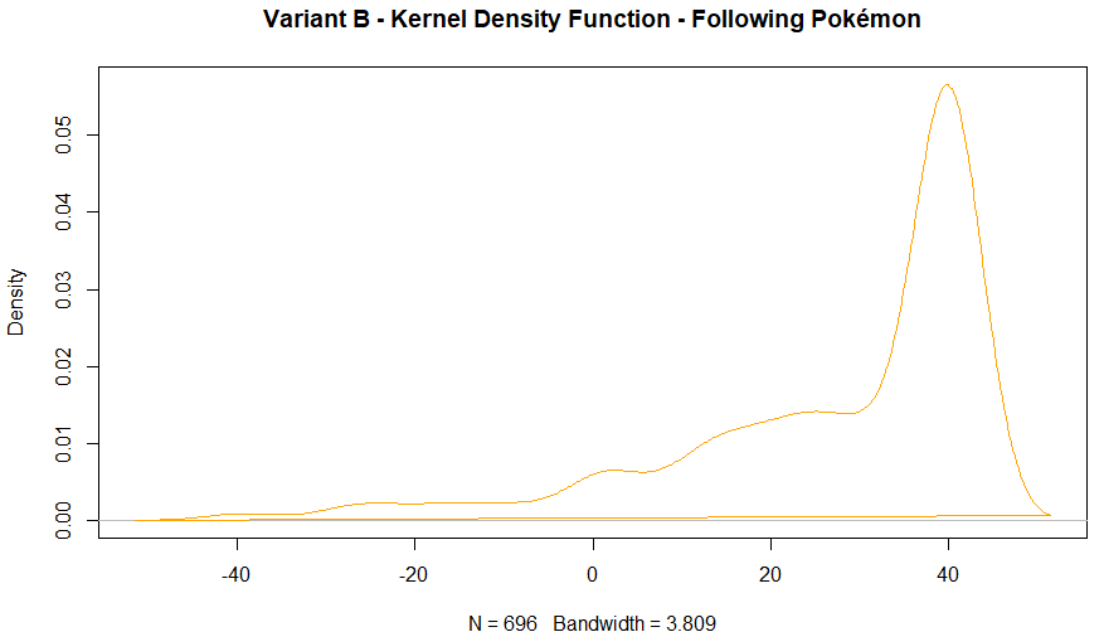

Figure D27: Kernel density plot for the distribution of the participants' sum of the scores for the feature 'Following Pokémon' in variant B

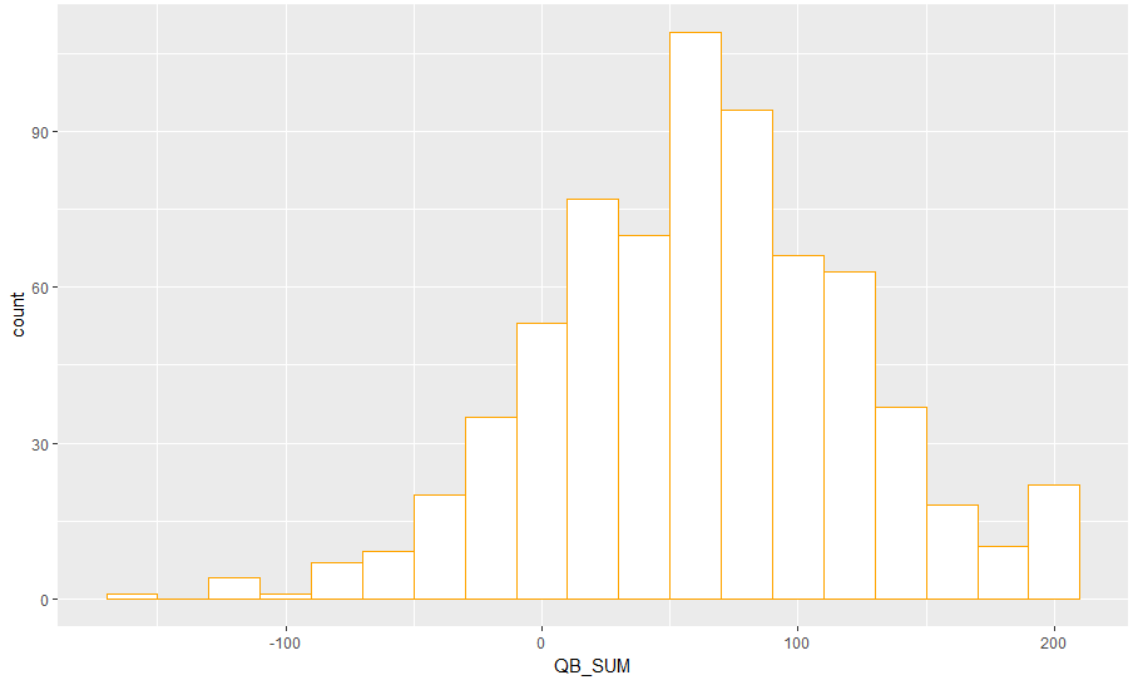

Figure D28: Histogram for the distribution of the participants' sum of the scores for the five features in variant B

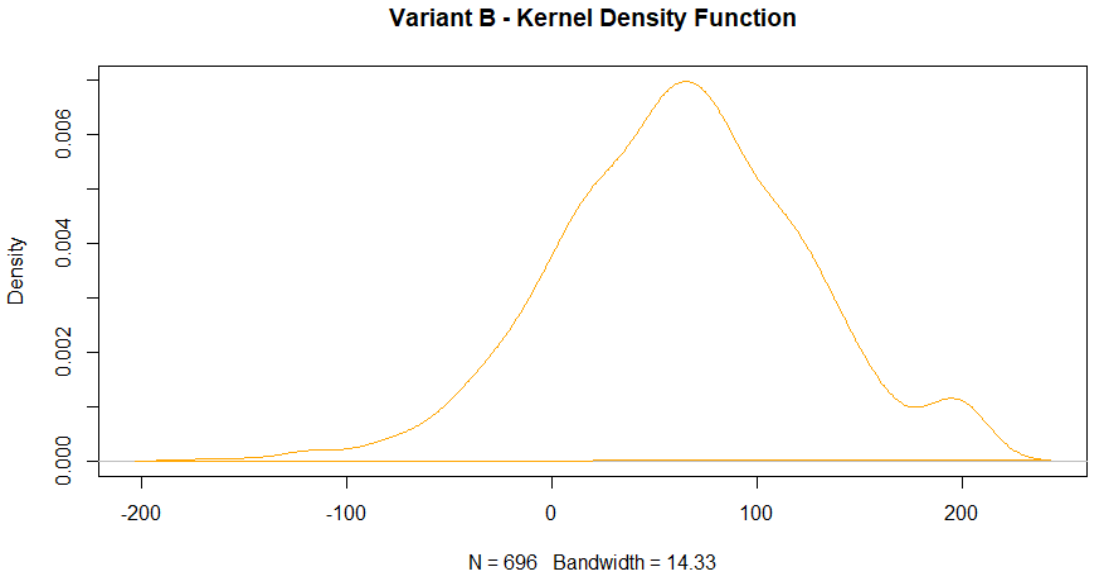

Figure D29: Kernel density plot for the distribution of the participants' sum of the scores for the five features in variant B

Variant C

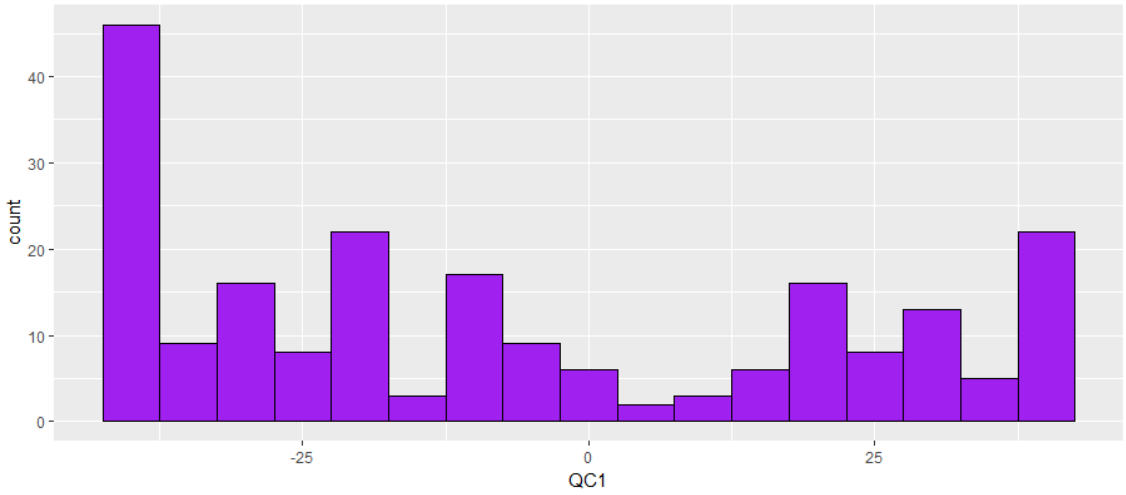

Figure D30: Histogram for the distribution of the participants' sum of the scores for the feature 'Pokémon Catching (instead of Wild Pokémon Battles)' in variant C

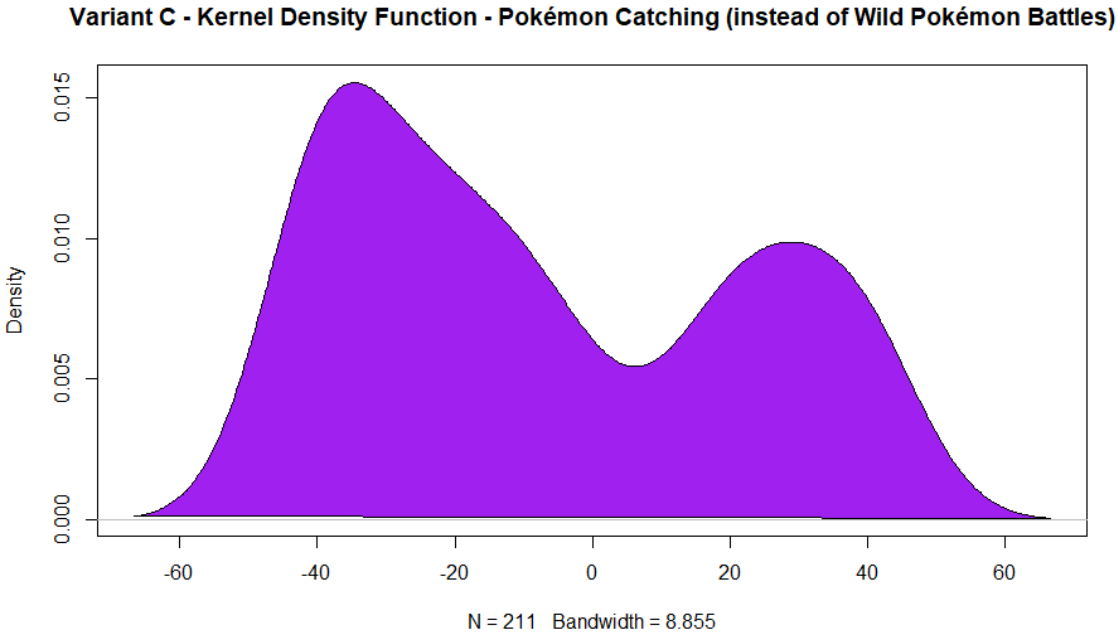

Figure D31: Kernel density plot for the distribution of the participants' sum of the scores for the feature 'Pokémon Catching (instead of Wild Pokémon Battles)' in variant C

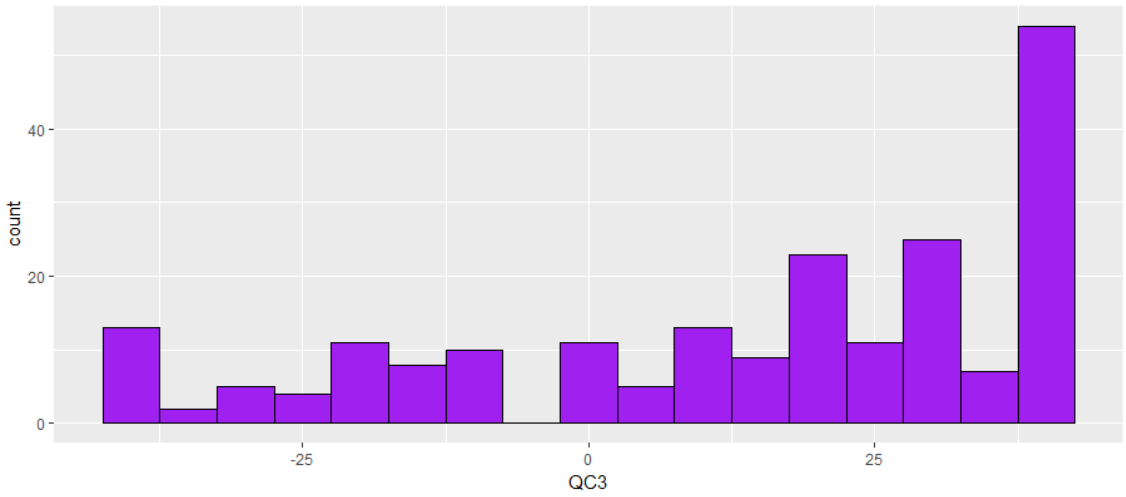

Figure D32: Histogram for the distribution of the participants' sum of the scores for the feature 'Pokémon in overworld (instead of random encounters)' in variant C

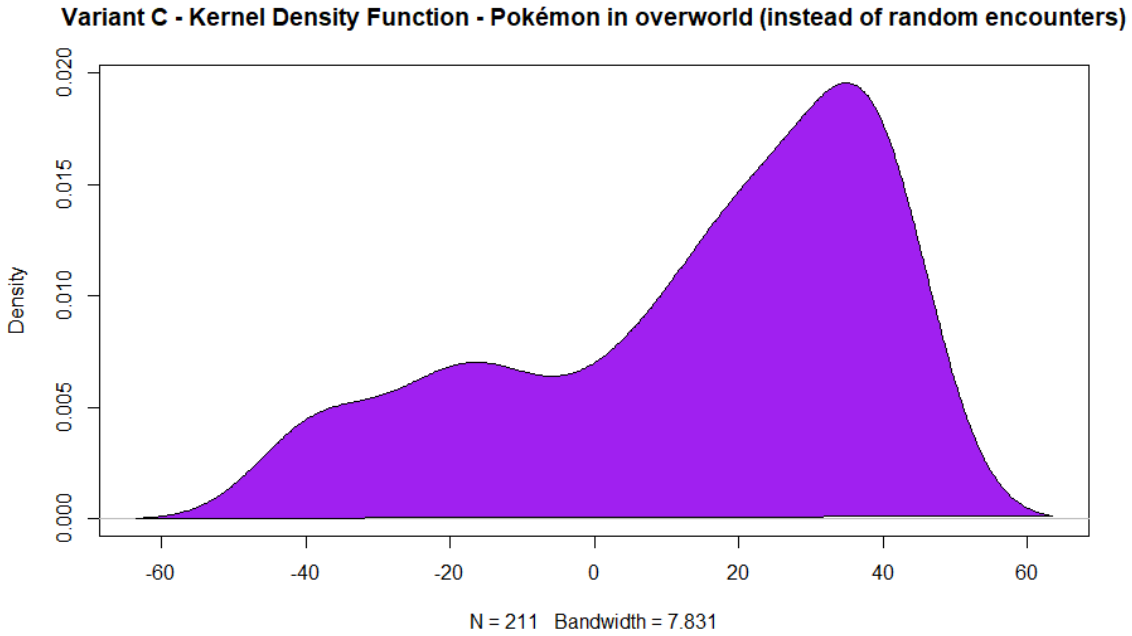

Figure D33: Kernel density plot for the distribution of the participants' sum of the scores for the feature 'Pokémon in overworld (instead of random encounters)' in variant C

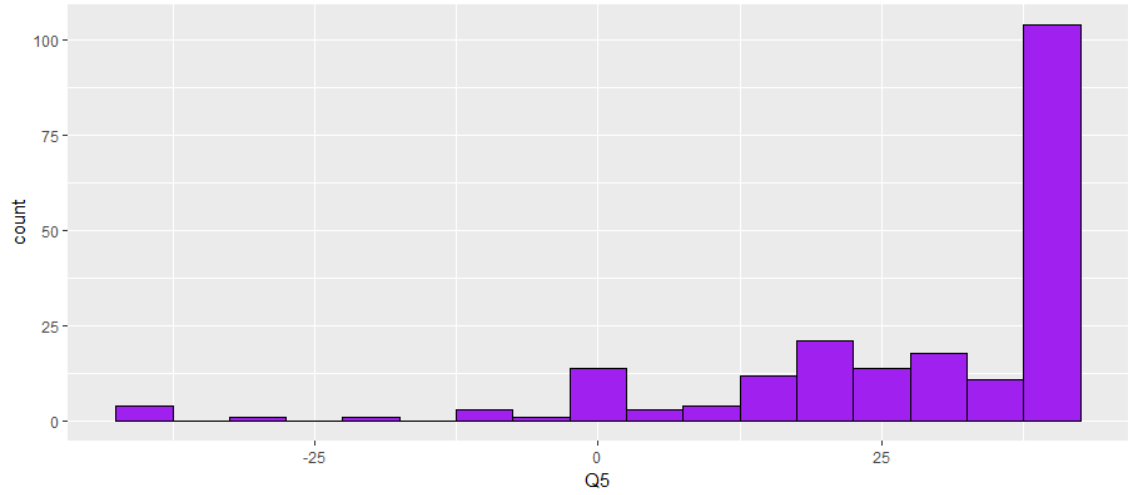

Figure D34: Histogram for the distribution of the participants' sum of the scores for the feature 'Play Together' in variant C

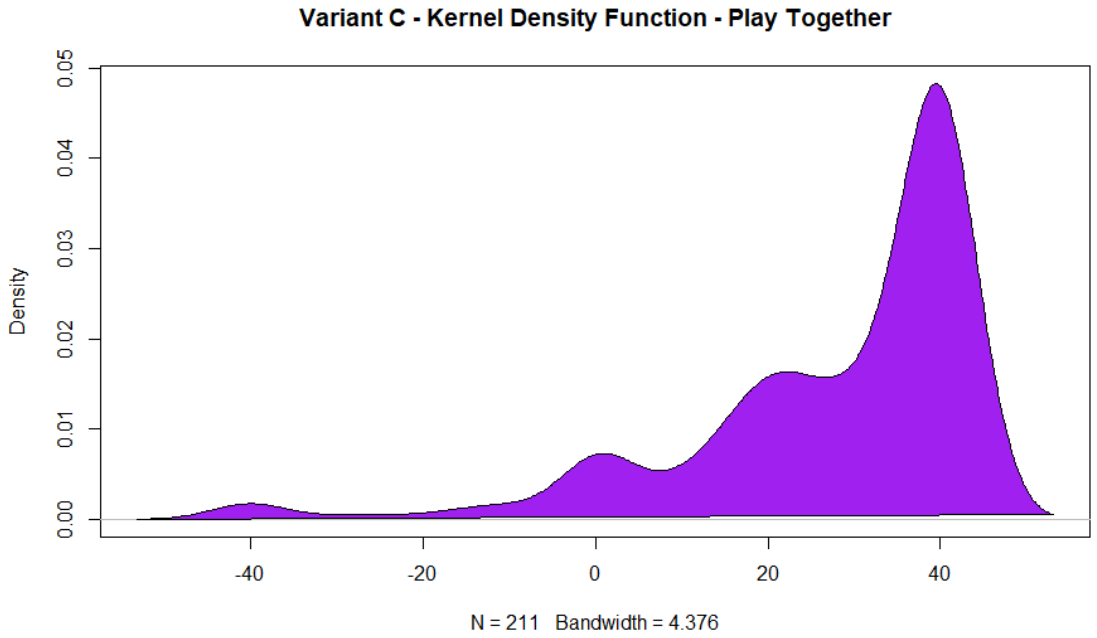

Figure D35: Kernel density plot for the distribution of the participants' sum of the scores for the feature 'Play Together' in variant C

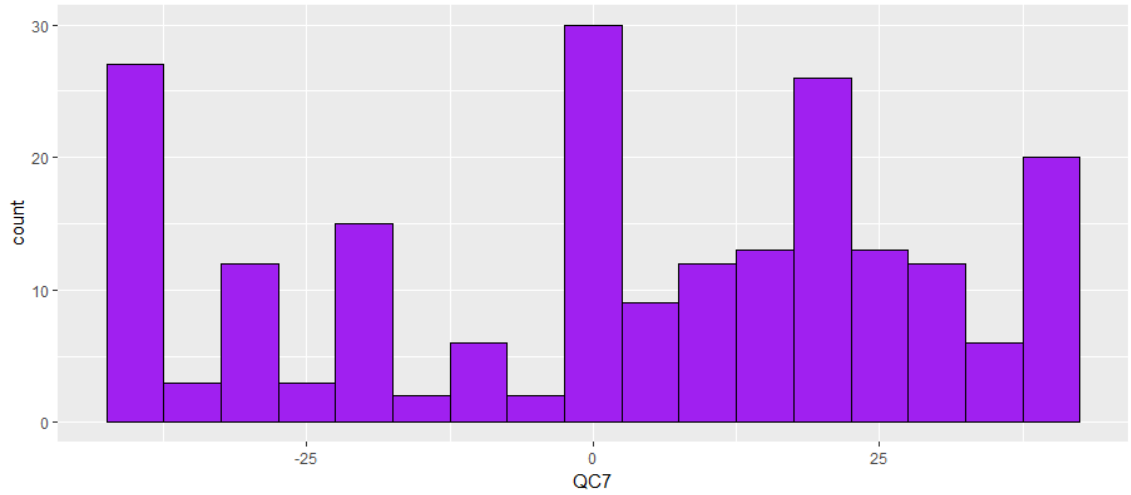

Figure D36: Histogram for the distribution of the participants' sum of the scores for the feature 'Pokémon GO Integration' in variant C

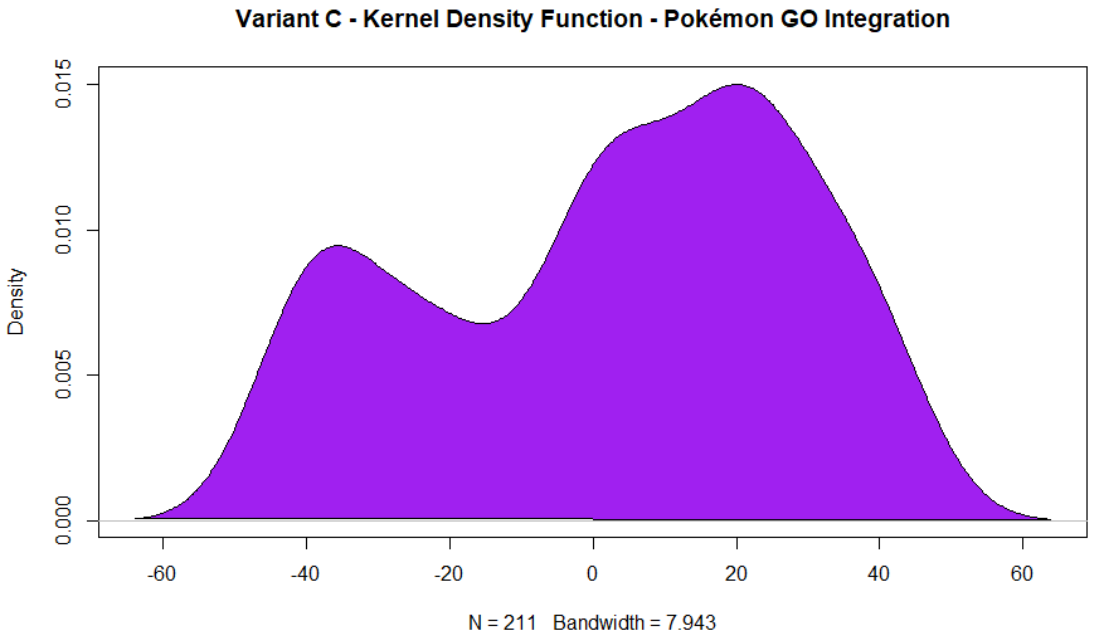

Figure D37: Kernel density plot for the distribution of the participants' sum of the scores for the feature 'Pokémon GO Integration' in variant C

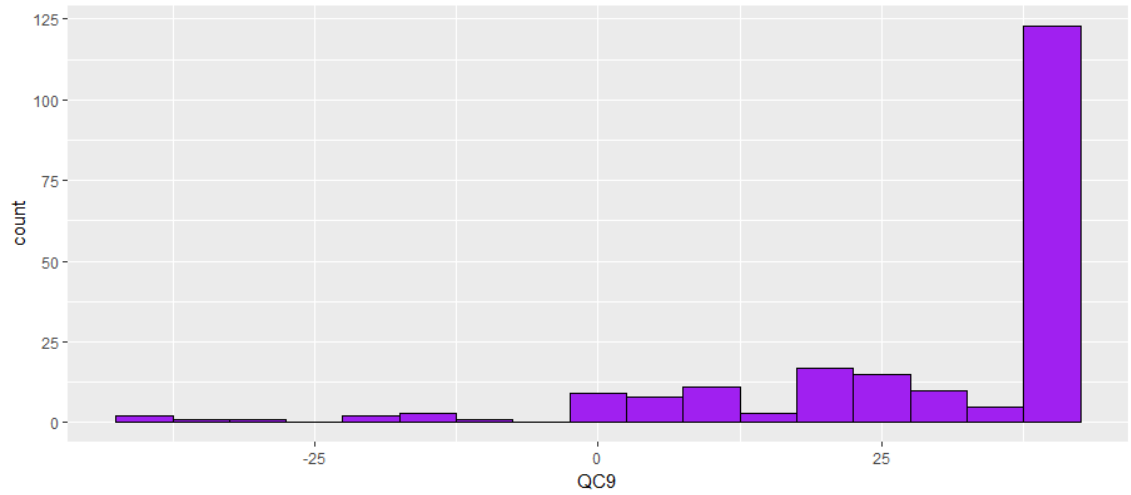

Figure D38: Histogram for the distribution of the participants' sum of the scores for the feature 'Following Pokémon' in variant C

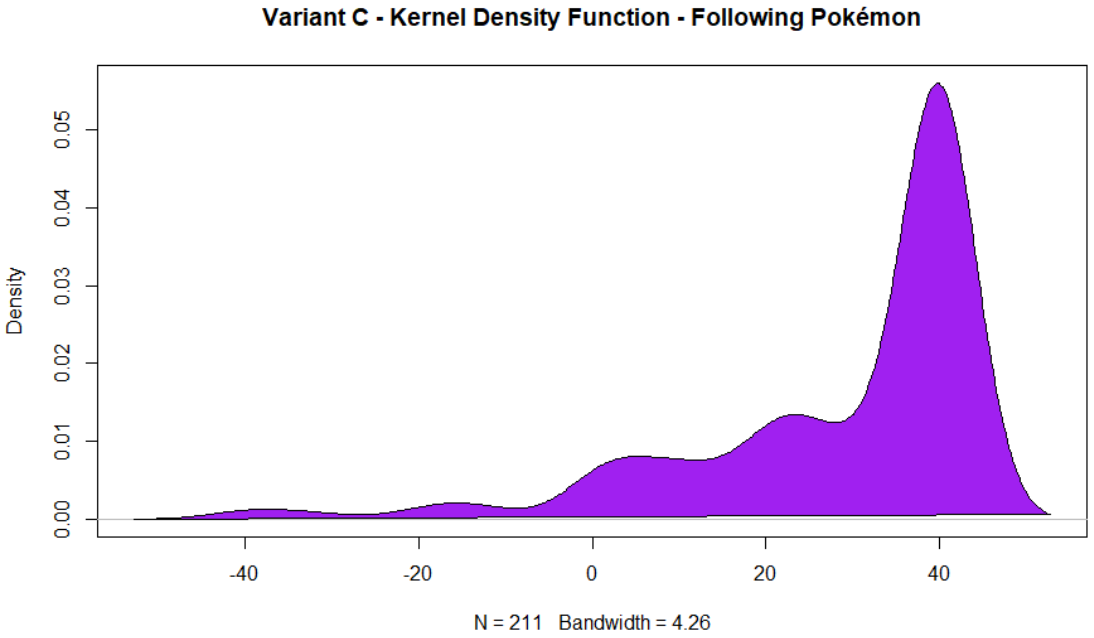

Figure D39: Kernel density plot for the distribution of the participants' sum of the scores for the feature 'Following Pokémon' in variant C

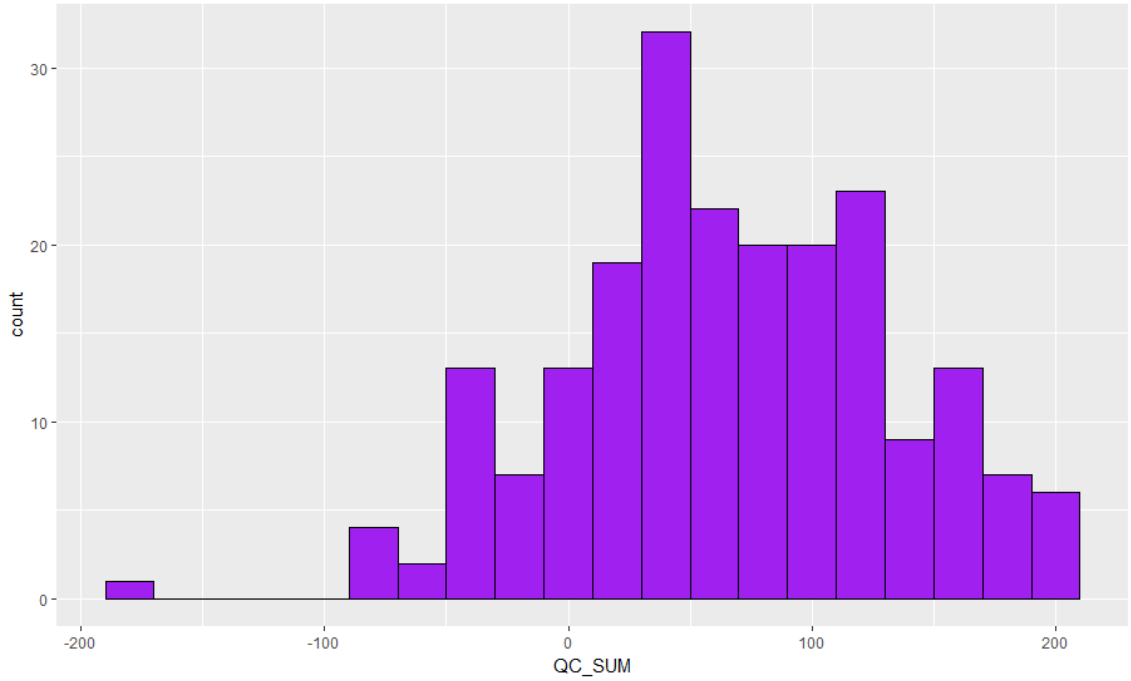

Figure D40: Histogram for the distribution of the participants' sum of the scores for the five features in variant C

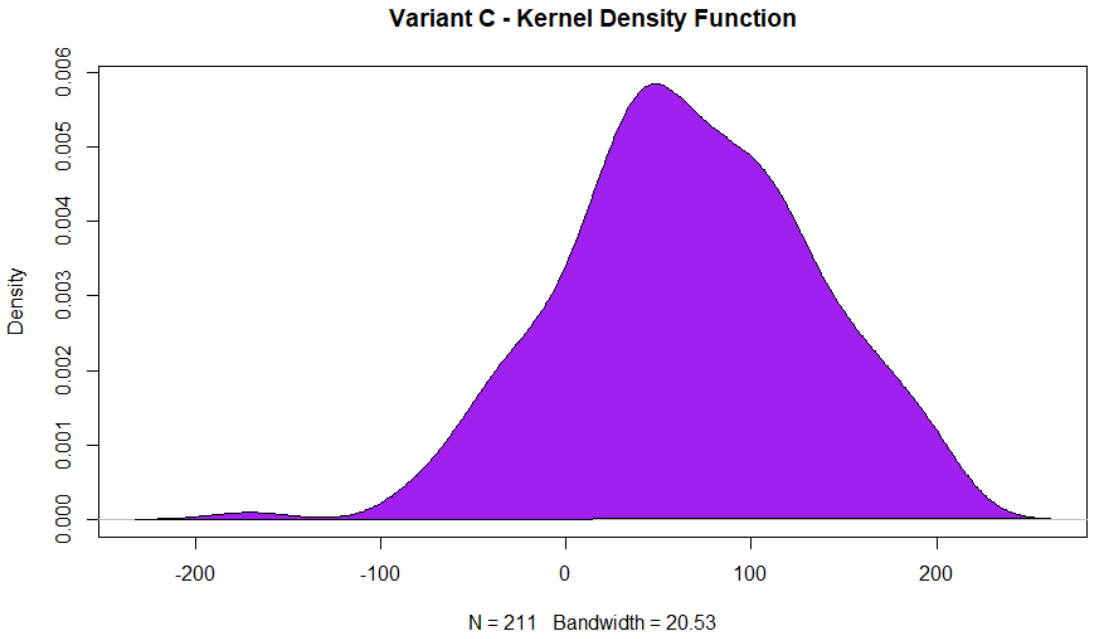

Figure D41: Kernel density plot for the distribution of the participants' sum of the scores for the five features in variant C

Appendix E - Predictions vs. outcomes (variant C)

Feature 1	Predicted	Actual	Relative (over)estimation (%)
-40 to -24	0.3099	0.3697	-16.1753
-24 to -8	0.1655	0.2038	-18.7929
-8 to +8	0.1729	0.0806	114.5161
+8 to +24	0.1557	0.1422	9.4937
+24 to +40	0.1961	0.2038	-3.7782
Feature 2	**Predicted**	**Actual**	**Relative (over)estimation (%)**
-40 to -24	0.1626	0.109	49.1743
-24 to -8	0.1264	0.1422	-11.1111
-8 to +8	0.1996	0.0758	163.3245
+8 to +24	0.2172	0.237	-8.3544
+24 to +40	0.2942	0.436	-32.5229
Feature 3	**Predicted**	**Actual**	**Relative (over)estimation (%)**
-40 to -24	0.0782	0.0237	229.9578
-24 to -8	0.0751	0.019	295.2632
-8 to +8	0.1832	0.0853	114.7714
+8 to +24	0.2246	0.2085	7.7218
+24 to +40	0.4389	0.6635	-33.8508
Feature 4	**Predicted**	**Actual**	**Relative (over)estimation (%)**
-40 to -24	0.2207	0.2133	3.4693
-24 to -8	0.1239	0.109	13.6697
-8 to +8	0.238	0.2133	11.5799
+8 to +24	0.1826	0.2512	-27.3089
+24 to +40	0.2348	0.2133	10.0797
Feature 5	**Predicted**	**Actual**	**Relative (over)estimation (%)**
-40 to -24	0.0612	0.019	222.1053
-24 to -8	0.0569	0.0284	100.3521
-8 to +8	0.1619	0.0853	89.8007
+8 to +24	0.1827	0.1706	7.09261
+24 to +40	0.5373	0.6967	-22.8793

Table E1: This table compares the outcomes (Actual) of the relative frequencies for the five intervals as stated in the Choice-Matching part of variant C to the predictions of the respondents (Predicted) for each of the five features (denoted by Feature 1 to 5). The final column indicates the relative (over)estimation in % of the predicted values versus the actual outcomes for the relative frequencies of the intervals.

Feature 1 = Pokémon Catching (instead of Wild Pokémon Battles)

Feature 2 = Pokémon in overworld (instead of random encounters)

Feature 3 = Play Together

Feature 4 = Pokémon GO Integration

Feature 5 = Following Pokémon

Appendix F - Resamples of N = 211 for variants A and B

B:

```
> sample(variantB$QB_SUM, 211, replace = FALSE, prob = NULL)
  [1]   2  -7  63  70 109  45 160  15  63  89   2   8  -2  80
 [15]  87  89  71  81  54  19 -33 102
 [23]  16  26  56  93  57 108 166 -81  94 141   3 117  20  72
 [37] 109 -23  69  38 -19  86 146 108
 [45]   7  48 -19  57  82 -28 -36   2 119  56  -7 101  91 -13
 [59]  98 126   8 -25  56  63  16  71
 [67] -15 128 129  66  18  -9  47 -80  17  70  75  74 114 133
 [81]  87  57 111 180 124 139  24  80
 [89]   7 120 -42 -124  75  47  68 102 -28  38  25  69  68  76
 [103] 35  23  80  84  97 106  45   6
[111]  62 -30  62 200 200 -21 200 143 135 -20  21  52  25  75
[125]   3 133  70  94  59 137  47 111
[133]  68 124  35  36 118  74  51 141 128  34  97 -10 117  63
[147] -58 200 -160 -10 100  87  42 111
[155] 124  40  31 120 127  89  75 121  98 124 160  65 104  98
[169]   0 -105  67 117  30  83  75  84
[177]  69  54  74  52  99 199 108  17 112  81  86  27  43 120
[191] 163  53 200 -26 132  61  52  72
[199]  85  96  70 161  44  82  72 129 129  26  29  56 -25
> sample(variantB$QB_SUM, 211, replace = FALSE, prob = NULL)
  [1] -41  71  88  29  57 147  85 126  81 200  69  22  49 -52
 [15]  93 -13  72  12  17  -8   2  72
 [23] 128 156 -37 160  17  88  71  34 107  29  56  54 139  61
 [37] 100  -6  22  36   5 179  16  40
 [45]  25 163  70 161  51  13  81 -124  56  46 101 -46 129  69
 [59] 154  94 134 -58  21 -40 139 159
 [67]  66 160  79   8 -52 157 146 -27 -21 132 120 200 113  13
 [81] 115 120   5 -33 131   8   6   7
 [89]  74  82  51 124  53  29  36  47 -53 133  42  63 -78  69
 [103] -9 186 190 103  81  75  63 -119
[111]  40 104  21  48  35  34  45 120  59  83  75  81  74   3
[125] -80  70  35 145  66  11  49  65
[133] 129  -2  32  61 112  77  89  63 -28 106  58  83  45 118
[147]  58  91  10 118 -18 182  48  28
[155]  44 -28  19  -7  19  68  10  65  63  81  47  76  98 122
[169]  69  17 111 199  52 200   2  58
[177]  85  70 -11 111  55 -30 115  50 124  69  24 -12 189 166
[191] 114  71  54 140  34 -31  91  74
[199] 108   4   5 -14  40 -40  98 117 106  14  84  81  46
> sample(variantB$QB_SUM, 211, replace = FALSE, prob = NULL)
  [1]  -9  54   0  98  49  55  80 120  70 -27 100 108 117 -48
 [15]  79 200 189 154 103 132  17 -11
 [23] -52 200 188   5  59  66 200 -81 100  11 129  78 -113   1
 [37]  28 111  75 -69 -15  53  25  75
 [45]  -8  61  61  40  80 -36  64 103  29 -35 128  84  27 121
 [59]  30  66  98  64   4  53  99  23
 [67] 129 -81   6  26 -40 120  53 112  90 109  43  17  52 140
 [81]  76 -58  59  99 -31  58  79 200
 [89]  79 -72  42  21  76 145  24  13  63  70  57  13 -42 -78
 [103] 81 117  72  47  69 -60  56  32
[111]  -3  70  89  99  56 -105 107  70 107   0  51 200  92  28
[125]  38 111 -27 106 147 117  -2  45
```

```
[133]  159  135   68  -14   -1   36   15   38   67  109  102   19   10  120
  163  129   -1  -37   83   11  -40   22
[155]   27   20  -30   51  107  -20   81  149  160  124   85   30   93  180
   38  116   18   -9    7  150   80  -36
[177]  120  111   70   63  111  160   80  151  129  122   29   44   44   14
    8    5   43  149   34  139   31  -11
[199]   63    4   80    2    4  120  126   39  106   66   87   16   23
> sample(variantB$QB_SUM, 211, replace = FALSE, prob = NULL)
  [1]   64   38   54   86  180   70    3  144  -27   74    5  118  107  -81
   75   42  144  158   89  -35   48   54
 [23]  -10   58  163   40   35  132   29   31  160   16   30   22   23  -25
  108    7   53   -1  -48   60  200   98
 [45]    0  -42   98  200   79  -20  101   56   75  119  200   52  -23   52
   -9   80  137   20   22  -21  182   87
 [67]  114  128   52   95    1   92  200   59    2   28  141   54  120   86
   44   41   52   11   71   95  190   21
 [89]  140  -53   74  106   53  107  -26   81   98   66  126   81   72  100
   61  -52  137   62  110   84   35   -8
[111]   81  145   52  -52   62   45   63    2   98   28  -78  143  200   29
   80  133  115  107   39   87  128  -41
[133]   19   21   88   63  141   28   10  -58   47  200   -1  -11   52   -9
   64  -27  188  113   57  126  100   27
[155]  -19  -86   27  119   43    1  111  132   15   97   13   51  -78   69
    4   38   -2   49   75   47   70    7
[177]  200  121   80   94  150   40 -160   59  160   18  182  117  118  112
    2   70   64  -46   75  112   10  -10
[199]  101    3  -37   15   83   56    5   71   51   75   94  122   96
> sample(variantB$QB_SUM, 211, replace = FALSE, prob = NULL)
  [1]   36   69   75   70  117  132  117   70  117   15   72  165   43   19
   98   18   84   28   70   26   89  156
 [23]    0   89  200   91  135  111   72   70 -113    2  -78  140   76   -7
  120   80   12  101   79   54  200  -31
 [45]  127   50   70   69   99   10   75   79  166   66  -20  129  160   38
  186   28   78   56  -40  200   19   -2
 [67]    8   55   85   75   70  111   17  -60  134  100   70   49   49   27
   66   70   57   95   62  -26   30  -28
 [89]   93  -52    6  200  -13   58  126   41  146  145  120  129   10   76
   -8   47   39   48   99   21   57   63
[111]  102   81  144   43   19  -72   63   61    4   56  -28  -18  101   23
   44  200   69  -42  149  -23   86  126
[133]   68   11   10    3   90   60    2  109   59  -60   35   27   80  -11
  200  200   21  -12  -21    7  154  100
[155]   51   40   91   34  -34   80  -42   83  117   24  200   98   72  126
   20   87   45   95  117   70   81   74
[177]  -27   87  -48   54   49   47  111  -14    3   82  160   50   90   80
   58   41  -15   18   69  -40   84  158
[199]   14  115  161   98   98   58  106   53   40   57  200  133   42
> sample(variantB$QB_SUM, 211, replace = FALSE, prob = NULL)
  [1]   70   30   -7  157   10  135   13    5  -10   61    5   11   39   71
   51   57   47   81  137   89   23    3
 [23]   34  -30  136    5    7   96  144  133  139  126   81  120   52  200
   49   22   99  -37   49  -11  -52   55
 [45]   48  110   52   74  104   72   87   84   82  118  -21  -53   54   95
  139  -48  -15   75  -69  126   48   79
 [67]   29   66   99  200  102  200  200   25  200   40  -23   49    2  141
   19   29  200   41   83   45   72  109
 [89]  200   83   -8  200  -42  -10  101  117  160   70  -35   45   89  200
   98   75   88   34  108   53   68   16
[111]  -25   63  111 -105   44  -41 -160  145  111   18    8   80   55  200
  117  -58   62   84   71   47   -6   57
[133]   48  118   15  106  118   24   40   31  -27   63   35   54  -31  132
   64  166    2  122   75    8  107   26
```

```
[155]   61  200  200   56   34   19  112  124   98   75   94   46  -36   16
   22   53   83  129   67   23   23  124
[177]   91   13   71  -48  -72  149  -78   92   89   42   80   -7    7   57
  143    0   28  113   82   40   27   76
[199]   79  128  143  180   91  126   82  111    2   81  -25   99  120
> sample(variantB$QB_SUM, 211, replace = FALSE, prob = NULL)
  [1]   17   93   51   47  -78  154   -8  -23   18   51   16    4   50   31
    8  180    5  118   41   66   44   99
 [23]   53   67  111  100  -29  -20   26   22   51   69   40   83   36  151
  -11  200   15  122   29   -7   75  128
 [45]  100   63   83   38   42   57  160  121  150   61  -28   57  151   25
   63  -13   52  135  200   83   41   10
 [67]   75   50  -11  129  -40  106  -25   48   23   66  129   70   79    7
   81   46  101    4    1  -18   65  132
 [89]   75  150   29  126    2   70   45   98   42   70   70  104  119   98
  -21 -113  126   80  128   49  118    7
[111]  -52  200   58  -46  137   78   35  122   61   72   68  200  120   16
   64  -19   32   20  -15  158  120  -31
[133]  -10   23  -20   99   49   89   82   19   79  129   -3  200  190   54
   40  188  114  106  153  -52   56    2
[155]   51   58   53   44   84  124   49   10   20  -12  -60   13   43  107
   82   75  123  -81  135   72  107   89
[177]   84   57    3  -33   77    2   31  -26   39  180  200  200  200   15
  115   87  122  -35  -37  107   86   52
[199]   34  134  -41  -81   64  157   91   -5   21  -48   75  114  143
> sample(variantB$QB_SUM, 211, replace = FALSE, prob = NULL)
  [1]   40   19   -8   91   34   91  -23   40   32   -5  -21   -9   94   83
   24  106  108  145   25   56   51  117
 [23]  -29   62   92   80   78   72   30   40   23  127  102   99    5  200
   80   78   70   -1   19   66   39   95
 [45]  100   88   30   53  133  101  -25   75   11   76   55  133  -60   65
   21   64   41  -52   43   38  160   82
 [67]   31  144   35   57  100   36  136   55   29  -58   36   80   72   98
   63   71   71  120  200  122   70   90
 [89]  200   83  -58   75  129   38  -53  135   82  126   49   89   63   61
   74  200  -10   13   54  -17   83  129
[111]   31  -42   84  -11  -12   83  -27  117   92  120   13  102   72  160
   44  180  137  118   50  160   67   28
[133]   98   19   -7   89   30   99   34   74  100   93   80  119  153  121
   16   85   -7   -9  124   24   54   20
[155] -120  104   97   21   45   48  121  147   94  118  200  -20   76   84
  140   90    2   46    3   34  -20   20
[177]  111   41  -36   86   42  -60    7   99   51  124  -69   15   74  -35
   13  139  111  139   97   59  118  200
[199]  100  -10  131  -48   63   48   79  -11   29  129   40  -25   42
> sample(variantB$QB_SUM, 211, replace = FALSE, prob = NULL)
  [1]   42   50   75  -34   90   55  135  200  -69   62  -42   20   53   63
  -46  132  114   80  139   14   84    0
 [23]  100   82    8  -18  199   75  151  107  158   31  145   19   38   81
   74   59  141   67 -160  108  101  180
 [45]   39  200   79   80  -58  200  100   82  117   51   10  -19  112  -36
  137   53   -2  120   99   53   54   40
 [67]  157  200   81  143   76   53   81   83  126   71  101  200   83   25
   21   58  109   30  102   85   65  116
 [89]  -11  200   35    4   17   80  182  117  120   82   62  153   39   71
   75  200  -28  -11   44   51  149    2
[111]   81  100   91   78   70   54  137  104  120  -21   98  108   54  200
   79    3   52   30   35   51  118    4
[133]    1    6  -20  124   -1   57  112  120  -10   23   -9   87   26  129
   70   69  -81  124   25   16   75   54
[155]    2   55  -10   91  127   70  -72   13    2  117   83   99   40   -1
   48   96   98   78  129  120  107  -35
```

```
[177]    3   56  -78  140   63   46   69   25  -12   81  -23   25   71   71
   -9   58   36  111   -8   28   57 -105
[199]   71   16   51   95   52   -2   56   72  113   87  -78   68   24
> sample(variantB$QB_SUM, 211, replace = FALSE, prob = NULL)
  [1]  158  200  124   69  -30  106   13  146   67  133   93    7   43   44
   52   75  -20   89  122  200   84   50
 [23]   79  146  -11   49  100   31  200   17   36   99   89  182  -10  -26
    6   58   16   54  126   85    1  106
 [45]   90   -1  -41   86   99   18  179  200  137   66   81  117  -35  129
  200    5   71   68   87   69    4  141
 [67]   39  108  -86   54   24   66  107  -33  135   58   63   26  -21   28
   23  166   65   47   76  -78  186   19
 [89]   66   81    5   54   70   51   51  199   52   -1    5   83   93  111
  144  100   30  -48  118   97   18   93
[111]   52   27   83   57   63  117    7   38  157   72   10   70   42  106
  115   19   70   71  160   72  120  100
[133]  188   31  100   98 -120  182   64   23   22    1   71   86   56   47
   76   24  129  114  109  -37  -12  111
[155]   99   26   46  -27  151  133   57   71  139  -58   63  104   11  180
  -10   79   70   10  200 -124  161   19
[177]  200   72   28  180   13  140  124   -1   53  -28   92  -40   -9   59
  104  118   65  140   92  101   98   84
[199]  143  140  -28  -25   25  126   68   30   71  -25   35   51   57
```

A:

```
> sample(variantA$QAIWPS_SUM, 211, replace = FALSE, prob = NULL)
  [1]   40  165  170 -130  -70  140  180  140  -75  150   35   60  145  200
  200  140   30  146  120  -80  120   70
 [23]  150  150  160  120  120  160  140   80   60   70  160  180  165  180
  150  130  200   60  130  140   55  140
 [45]  100  125  150  125   90   85   80  200  200  150  140  105  110   95
  120    0  165  158  150  180   65   96
 [67]  200 -180  140   20  170  160  145   80  190   60 -200   80  200  140
   15  200  145 -170   10   70   25  100
 [89]   60  160  130  200 -120  135   80  160  -90  100   60    1  130  135
    0  160    0  170   65   40   65  -60
[111]  100   80  155  175   35   12   50    0   80  200  140  170  200   80
   75  110   80  190  120 -100  -90  120
[133]   70  115  120  100  150  155  144  110  130  150  170 -118   90  130
  200  171   80  110  130    5  160  155
[155]  120  -65   55  140  100  200   75  170  200  115  100  140  130  185
  120 -125   70  180  200  120  100  160
[177]   60  120   90   70  200  170  130  143  100  155  153  170  200   29
   45  190  -30  100  145  198  190  110
[199]  145  130   20 -108   80  140  160   90  120  160  180  115  200
> sample(variantA$QAIWPS_SUM, 211, replace = FALSE, prob = NULL)
  [1]  140  100  150 -130 -175  155  150   65  100  170   55  135  150  110
  184   30  115  180   46  140  100    5
 [23]    0  170  195  160  -80  100   50  -10   60  100 -165   30  200  150
   96 -120   65  150  120   95  155  -65
 [45]   20  185  155  140   90  170  125   55  130  140   60  150  200  120
  130   95   61  175    0   55  100   80
```

```
 [67]  200  -20  145  190  190  200  180  -20   75  130  150  140  140  104
   80  200  100  -85  190  200   60  180
 [89]  150  120  140  160  158  170  145   80  174 -118   90  145   20   40
   80   80   20  160  175   75   50   80
[111]  190  120  130  200   65  -60  -30   70   45  110  160   55  175   60
  100 -125  190    1  170  120  143  200
[133]   45   70  130  180    5 -170  120   90  171  115  110   80  100  180
  138  145  200  100  150  130  196  180
[155]  200  145  180   55  144  200   30  185  200  150  165   80  100   75
  140   70   12   80   80  100 -120   80
[177]  175  160  -40  200  -20   20   80  200  165   60 -180  110  170  177
   20   10  180  100  145   60  155   70
[199]  140   65  -75  145  170 -130  -90   80  115   35  180   25   20
> sample(variantA$QAIWPS_SUM, 211, replace = FALSE, prob = NULL)
  [1]  200  110  171  150  175  125  200  170  160  150   10  140  153  200
 160   81  167  200   12   29  160  130
 [23]  -60    0  135  -20  160   65  170  100  -12  170  -75    0  -80   55
 140  200  150  145  160  120  190  200
 [45]   80   50  184  100  150  165   46  200   90  200  180   65   61   60
  80   96  -40  180  100   10 -200  190
 [67]  150   90   90  170   60  140  100  130  160    0   20  200  196 -200
 175    5   60    0   80  140  160  145
 [89]   64   60  115  190 -100   35  140 -185   90  140  -65  145  100  130
 110  125  115  150   95  175   20  100
[111]    7  145  140  110 -118  190  120  170  -80   55  175  120  165   80
 140   45   80  174  -30  155   75  160
[133]  130  150   80  135  200  190  100   80   70 -115  115 -165  171   60
  85  150   85  150  155  180  150  110
[155]  145   35  115   85  200  135  180    0   25   20  200  104   70  200
 200   50  124  110  175  140  -65  120
[177]  100  115  200  -70   75  192  150  -80  100  200 -180  126   80  160
 190  135  185    0  120   45  150  198
[199]    0  200  160  105  160  195   80   60  105  100  105  160  150
> sample(variantA$QAIWPS_SUM, 211, replace = FALSE, prob = NULL)
  [1]  185  150   80  145  190   55  180    0  190   64  115  140  135  -90
 -130 -125  -85  175  130 -108  120  180
 [23]   50    0  170  180   80  -70  160  100  140   70    0  -75  120  147
 200   75  195  100  160  -60  200  155
 [45]  -12   75  -40  150  140  100    5  150   61  120  200  125  140   80
 -185  145  120  130   80 -105  126  120
 [67]  105  200  200  160  130  -10  200   95  190  170   70  110   15  160
  60  104   70   85   20  100  130  -90
 [89]  180  100  130  160  145  120  180   80 -120    5  175   29  100  108
 200  140  100  140   80  140  120  130
[111]  150   60   60  200  140 -125  170  138 -120  200  180  -70  158  167
 110  200  180 -120  185  184   95  140
[133]  172  -40  130  200  160   60  -90  195   55   35  165  120  175 -175
 180   15  160   10   20  155  150   80
[155]  200   60  110  170  125 -130  170  200   80  100    0  130  160  150
 -80   50  140  100  170   90  140  200
[177]  155  100  150   50   20   45  145  115    0  170  120   90  200 -170
  10  -10  185   92  160  -70  190  160
[199]  170  150  200  100   46  140  100   65  100  145  160  180  140
> sample(variantA$QAIWPS_SUM, 211, replace = FALSE, prob = NULL)
  [1]  165   55   85   50  175  200  100   45  200 -118  200  190  200   45
  75  150 -115  120  180   20    5  185
 [23]   65   35   70  -80  200 -120  135   70  -20  200  155  100  115  170
 140  115  160  190   75  200  140  110
 [45]  200   70  100 -105  135  140  160  140  175  150   95  135 -120   20
 192  200  115  120  200   90  200  200
 [67]   70   64  -10  130  110  120  100  120 -175   20  100  137  200  100
 140  180  195  200  100 -200  174  120
```

```
 [89]  -30  160  155  130  160  105   50  150  130 -130  110  150  170    0
  170    5   90   60  180  146  140  155
[111]  160  170 -130   85   55   20   70   85  130  120 -125   30  160   10
   55  -80 -200  140  140 -130  140  170
[133] -105  200  150  100  200  200    0  195  145   30    0  145    0  124
  -20   90  110   80  180   90  185  150
[155]  100   60  180  120  140   46   30  100  160  130  138  120  110   20
   60 -125  160   15  180  -70  105   95
[177]   29  200  200  140  143   50   80  115   80  180  -75    5  108   55
  160 -150  -85  105  171  100  120  190
[199]  147  -90    0  -40  150  200  165  130  150  167  140  200   65
> sample(variantA$QAIWPS_SUM, 211, replace = FALSE, prob = NULL)
  [1]  130 -108  155   60  155  140   60  185   80  170  160  200  110   45
  170   70  -90  190  -90  150   12  160
 [23]  -70  158  195  145  200   60   80  167  110  200  108   60  177  130
   75   80  -70  160  100  200  140   92
 [45]   85  200  145  100  120  110   15   20  200   90  -20  160 -105  180
  155   46  110 -175  190  200  171  120
 [67]   50   20  -75   80  145   75  160  175   60  150   90 -100  -30   80
  145   90  120  100  126  150   55  100
 [89]   85  170  200  150 -165   75  180   96  160  130  175  190  145  200
  120  180   95 -105  140  140 -200   30
[111]  130  140  100   80    7  200  -90 -120  -65  150  200  140  190 -105
 -200  155  150  170  130  180  140   70
[133]   90  200   95   80   80 -120    5  200   70  145  160   25    0  135
  145   80  170    0   10  -90  150  110
[155]   75   45   90  170   50 -100  150  100  150  190  170  180  150  140
  150   70  160  100   85   80  140  120
[177]  192   40    0  -40  140   85 -125  160  175  110  180  200  140   75
   65  170  170   85   60  190  120  160
[199]  160   20   55  198   70  195  -60  115 -130   60   80  100  145
> sample(variantA$QAIWPS_SUM, 211, replace = FALSE, prob = NULL)
  [1]  140  -20  140  180  -10  140  180  115   80  130  115  -80   25  115
  167  100   70  180  140  200  130  170
 [23]  171  200   55   75  170  120   55  200  -70  180   65  130  -90   95
  140  150 -118  180  147  100 -115  171
 [45] -200   50   90  170  200  100  150  100   20  110  180    1   95  195
 -170  160  170  100  180  140  -10   40
 [67]  165  115   30  145  110   55  110  160  200  200  100   60  130   45
  200   80    0  110  200  126  180   80
 [89]  150 -108   80   65   10  180   50  140   35  140  108  170  100  105
  170  140  175   90  110  100   75  190
[111]  100  185  180  120   90  160  -90  160  140  200  120    0 -120   40
   30  100  190  140   50  -65  180 -105
[133]  130  160  140   90  190   95  180   60   80  200  145   75  160  160
  105  140   64  200   30  120   85  180
[155]  115    0  110    7   30  140   81  200  100   55   60  170  140  120
  150  145    0  100  171  145  160  200
[177]   70 -130  115  200  135   70  160  100  175  170  165  130   29   80
  174   70   50  200  150  145  200  140
[199]  200  167  -40  120   95   60  130  120   98  125   20  120  160
> sample(variantA$QAIWPS_SUM, 211, replace = FALSE, prob = NULL)
  [1]   40   25    0  150  192  200    0  135  -30  160  110  120  170  140
    0  140  100  165   90  195   30  165
 [23]   15  115  130  190   80    0  120  138 -105  140  115  140  180   60
   10   50   80  100   50  200  150   35
 [45]  120  150   95   80  115  -80  200  -20   10  170   80   75  175  170
   70  190   75   80 -120  160  -85  150
 [67]  145   20   50  200  160   85  170  140  145  140  120  160  174  150
   40  145  171    5  -90   70  140  172
 [89]  120   70   55  110   65  -12  125  120   35  150  150  200  100   30
   45  135   60  150   95  160   70  146
```

```
[111]  150  100 -118  140  200   55  135   29   60  170  150 -130 -115  100
  160   65  140  140  120  140  200  150
[133]  145  180   95  200   55   85  130  170  147  140  120  100  200  200
 -200  110  140  140   65  145  140   85
[155]  -30   80  145  167  -70  155  185   98  140  200   75  160  150  -80
  150  170  190  110 -120  200  150  170
[177]  108  170   92    7   80 -125  180  200 -175  180  150   80    0  180
   70  170   60  150  130  100  160  -90
[199]  -70  120  150  110  180   30  180    0  100  -90  145  150   80
> sample(variantA$QAIWPS_SUM, 211, replace = FALSE, prob = NULL)
  [1]   64  200  126  120   95  140   60   35   90  155  146  180  175  100
  150  115  120  145  144  -65  -90  200
 [23]   80   30  -12   60  150  200   90   70  153  135  -65   70   60   55
  120  155   40  115  120  105  150  200
 [45]   60  147  195  170   80  160   80   10  200  155  120  100  160  130
  180 -100   55   60   35  160  180   20
 [67]   20  110  190  170  180  150   60  150  155 -130  -80   60  200   80
    0  200  110 -175  100  100  -75  100
 [89]  -80  140  130   65    0  140   95  185   80  200  200 -130  110  -85
  100 -115  165  115  165  110  160  115
[111]  175  100  145  120  140  200   80   60    0  -70    0  120  140  160
    0  140  172  120  -90   50  100   90
[133]  130  170  150  150   75   40  170  100    7  170   80  145  125  140
  120   70    0   90  171    0  -20  180
[155]   75  184  185  105   30  155   80  135  175  -40  170  200   90   75
  110  -20  180   80   50  100   70  198
[177]  200   85  170  190  100   70  200  190  100   15    0  180  -10    0
  150  167   80  100   75  190  -10  160
[199]  130  100  180  140   10  160  200 -200   60  180  145  180   65
> sample(variantA$QAIWPS_SUM, 211, replace = FALSE, prob = NULL)
  [1]  120  200  150    0   55  120  180  174  200   90  150    5  120 -130
    0   75  150  100  -90  160   10   55
 [23]  100  170   15   20   45  140  145   65  115   60  180  145  170  125
  140    0  100  135  135  200   80 -125
 [45]  155   70   60  160  180   90  160  130  115  177   85  140  150  160
  150    0  180  150  100  160   70  190
 [67]   70  200  140   60   85   70  -30  110  110  171   80  100   64  200
  155    7  160  -30  100    0  140  170
 [89]  200  150  -20  145  120  110   80  115  200   90  170  195  125  130
  145  120   75   90  150  200 -120   50
[111]  150  180  200  165  130   55  180  150   80   70   80  167  192 -108
   60  -70   29  -65  115   70   61  110
[133]   25   50  180  160  150   12   10 -130  130  140  135  105  140  -90
  200  198  140  155   65   80   90  140
[155]    1   20  200  100  130  160  120  120  105  200  170  -70  200  175
   50  155    0  190  200  200   90  140
[177]   90   65  200   85  165  180  170  140   55    5  155   45   60  140
   95  196  -90  130 -105  170  171   40
[199]  144   95   81   90  140  100  126 -103  150  110  130  185  190
```

Appendix G – Pokémon Games played by variant and survey finishing status

Complete dataset that answered INTRO4 (N = 2688)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other
Did not play games	551	705	1078	114	1349	1576
Did play games	2137	1983	1610	2574	1339	1112

FinishedA (N = 572)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other
Did not play games	100	125	187	19	258	308
Did play games	472	447	385	553	314	264

FinishedB (N = 696)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other
Did not play games	139	160	243	35	326	377
Did play games	557	536	453	661	370	319

FinishedC (N = 211)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other
Did not play games	33	40	70	8	74	82
Did play games	178	171	141	203	137	129

NotFinishedAB (N = 525)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other
Did not play games	122	176	270	28	331	363
Did play games	403	349	255	497	194	162

NotFinishedC (N = 684)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other
Did not play games	156	204	308	24	360	446
Did play games	528	480	376	660	324	238

TotalAB (N = 1793)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other
Did not play games	361	461	700	82	915	1048
Did play games	1432	1332	1093	1711	878	745

TotalC (N = 895)

	RBYGSCRS	EDPPtBW	B2W2XYSM	GO	MysteryRanger	Other
Did not play games	189	244	378	32	434	528
Did play games	706	651	517	863	461	367

Table G1: The absolute frequencies of game playment for the six categories of Pokémon games and for each of the completion states for the survey. *FinishedA = finished variant A, FinishedB = finished variant B, FinishedC = finished variant C, NotFinishedAB = started variant A or B (but did not finish), NotFinishedC = started variant C (but did not finish), TotalAB = finished or started variant A or B, TotalC = finished or started variant C. RBYGSCRS = played at least one of Pokémon Red, Blue, Yellow, Gold, Silver, Crystal, Ruby and Sapphire. EDPPtBW = played at least one of Pokémon Emerald, Diamond, Pearl, Platinum, Black and White. B2W2XYSM = played at least one of Pokémon Black2, White2, X, Y, Sun and Moon. GO = played Pokémon GO. MysteryRanger = played at least one of the games in the Pokémon Mystery Dungeon or Pokémon Ranger series. Other = played at least one of the games not specified here. No data is available to distinguish whether those that did not finish either A or B, finished one of these specifically.*

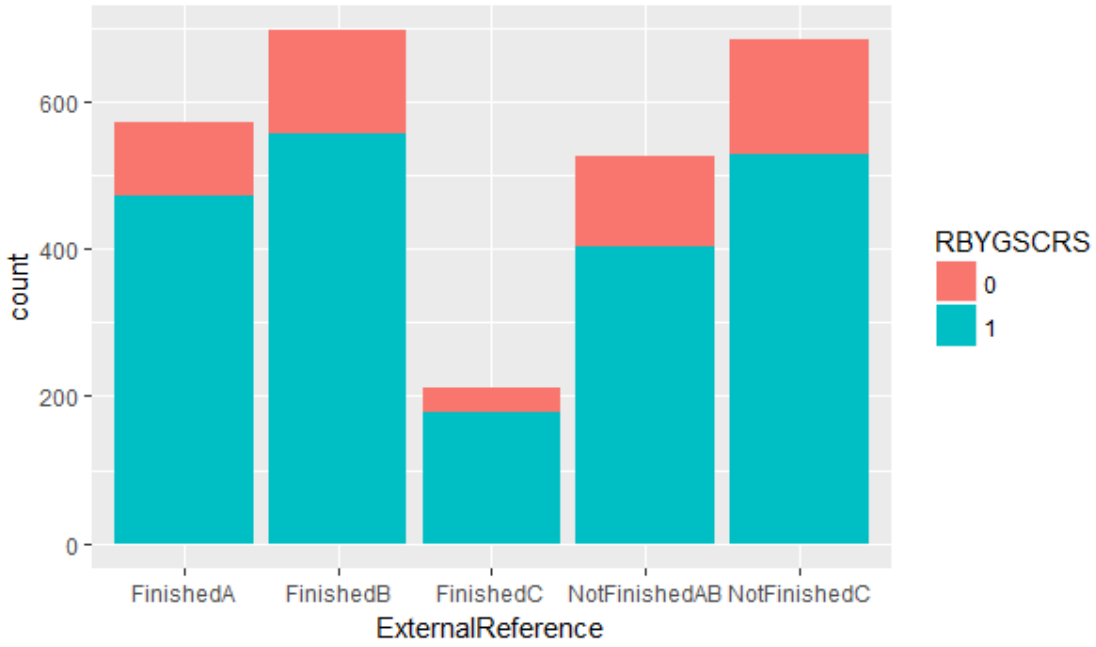

Figure G1: A graph that shows the distribution of game playment for *Pokémon Red, Blue, Yellow, Gold, Silver, Crystal, Ruby and Sapphire*. For each participant, a 0 indicates that this person has never played any game in this category of Pokémon games, a 1 indicates that his person has.

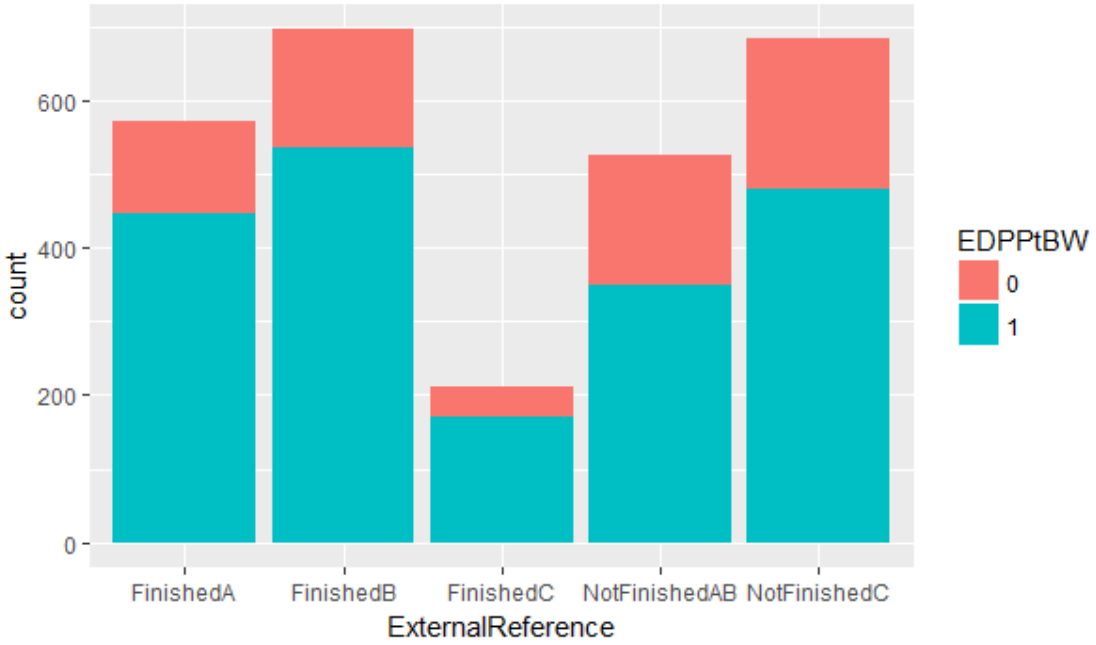

Figure G2: A graph that shows the distribution of game playment for *of Pokémon Emerald, Diamond, Pearl, Platinum, Black and White*. For each participant, a 0 indicates that this person has never played any game in this category of Pokémon games, a 1 indicates that his person has.

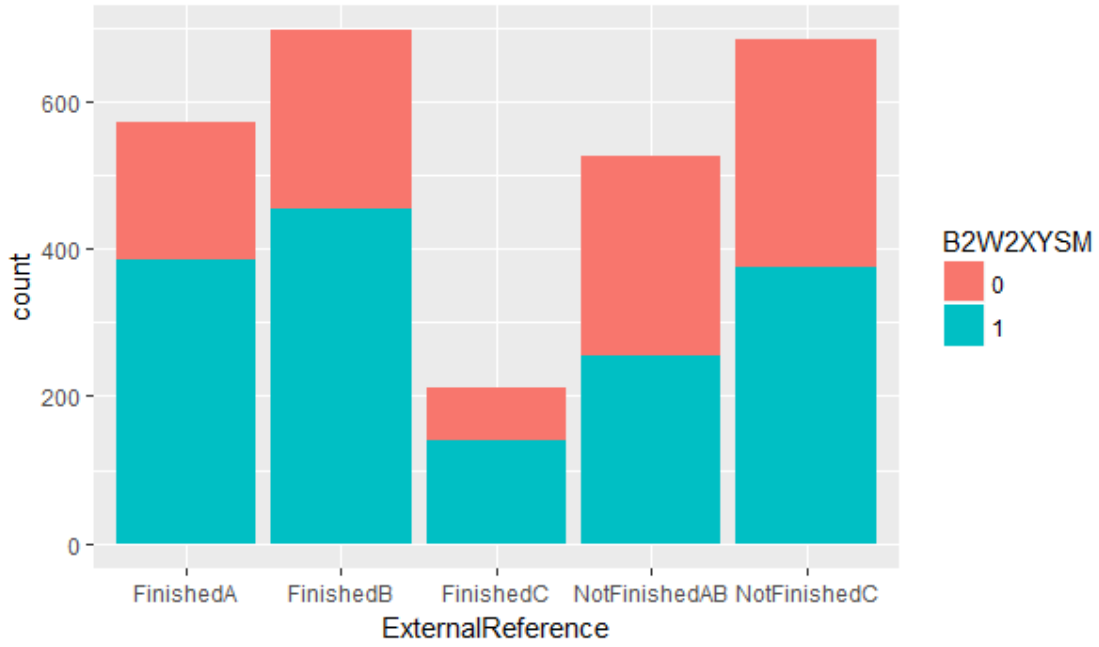

Figure G3: A graph that shows the distribution of game playment for *Pokémon Black2, White2, X, Y, Sun and Moon.* For each participant, a 0 indicates that this person has never played any game in this category of Pokémon games, a 1 indicates that his person has.

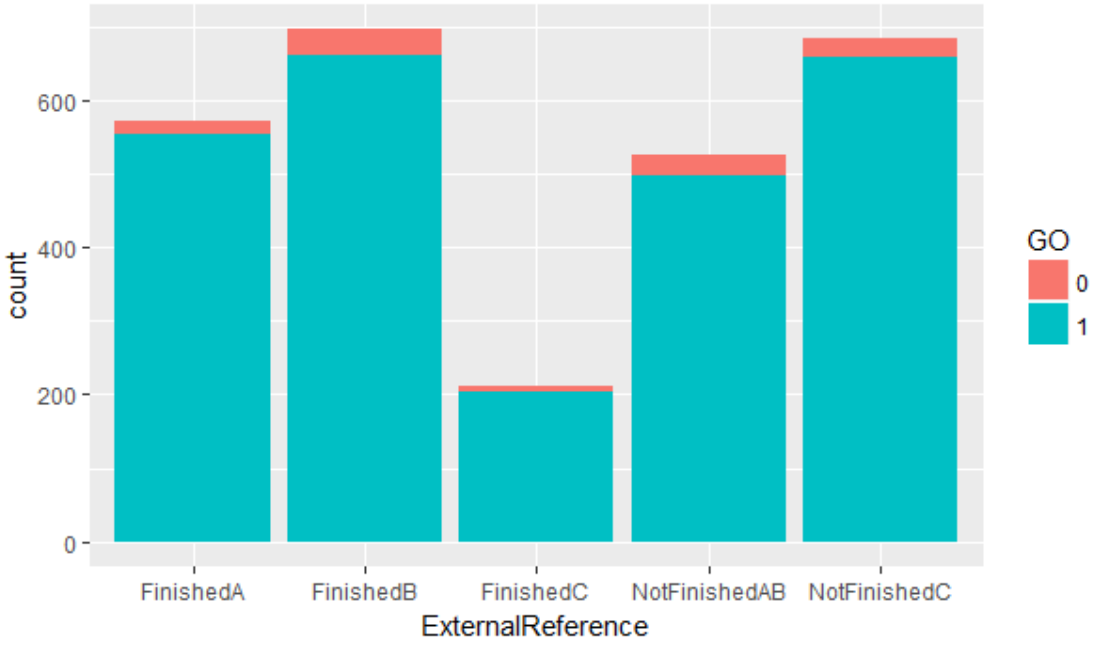

Figure G4: A graph that shows the distribution of game playment for *of Pokémon Go.*. For each participant, a 0 indicates that this person has never played any game in this category of Pokémon games, a 1 indicates that his person has.

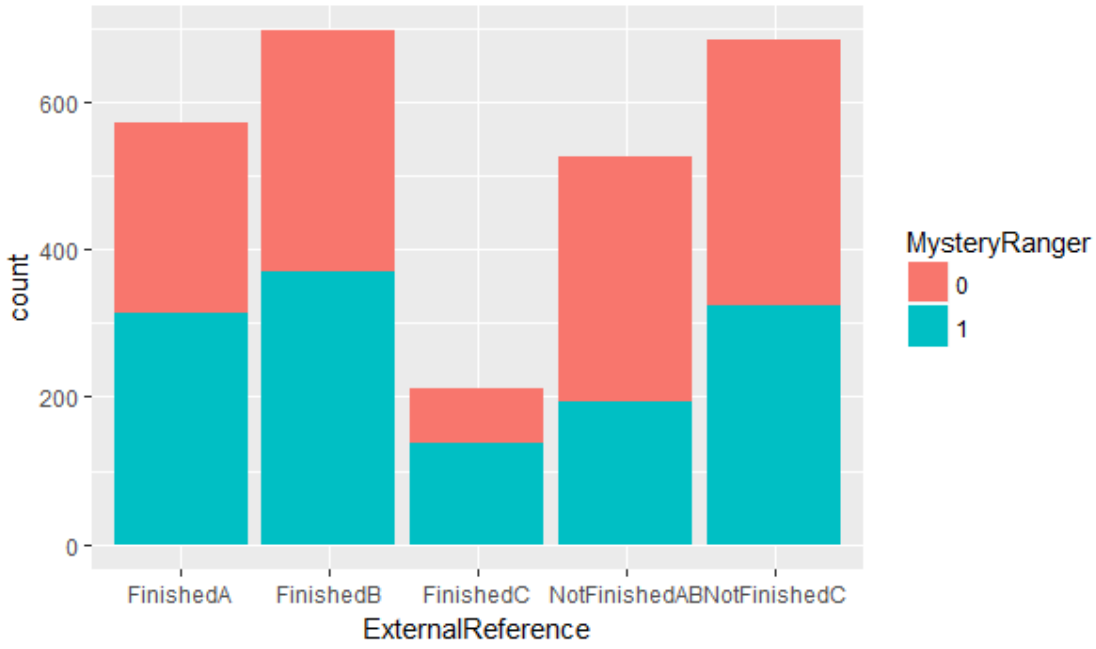

Figure G5: A graph that shows the distribution of game playment for *Pokémon games from the Pokémon Mystery Dungeon and Pokémon Ranger series*. For each participant, a 0 indicates that this person has never played any game in this category of Pokémon games, a 1 indicates that his person has.

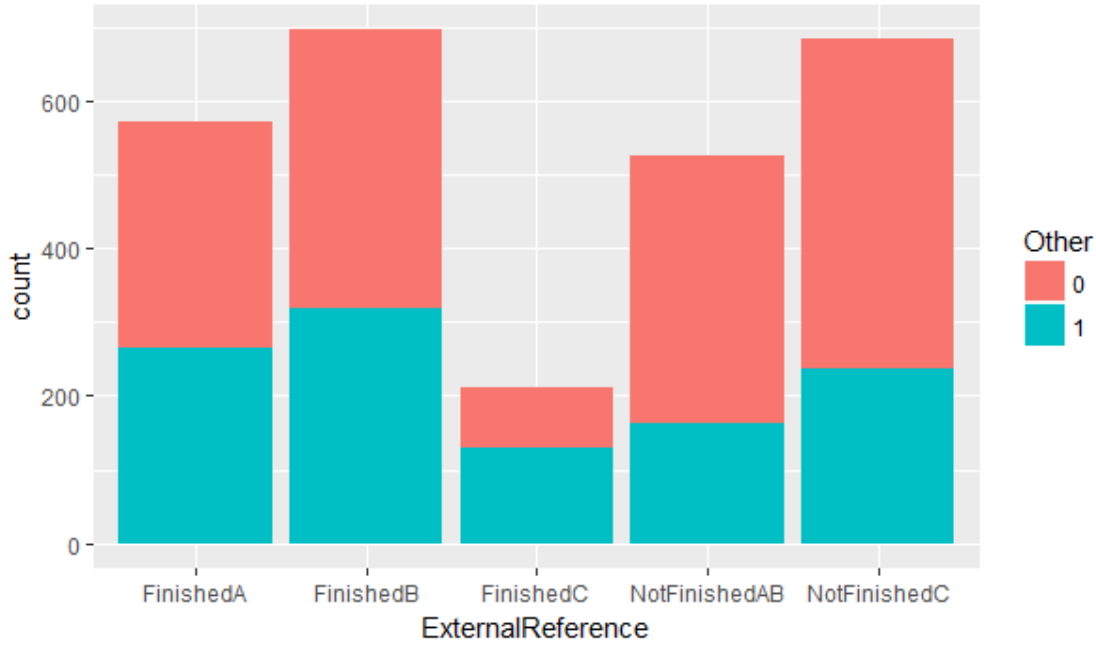

Figure G6: A graph that shows the distribution of game playment for *any other Pokémon game that is not specified in the other 5 categories*. For each participant, a 0 indicates that this person has never played any game in this category of Pokémon games, a 1 indicates that his person has.