International Journal of Research in Marketing 38 (2021) 770–791

Contents lists available at ScienceDirect

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Full length article

Machines learn neuromarketing: Improving preference prediction from self-reports using multiple EEG measures and machine learning

Adam Hakim [a,*], Shira Klorfeld [a], Tal Sela [e], Doron Friedman [c], Maytal Shabat-Simon [d], Dino J. Levy [a,b]

[a] Sagol School of Neuroscience, Tel Aviv University, Israel
[b] Coller School of Management, Tel Aviv University, Israel
[c] Sammy Ofer School of Communications, Interdisciplinary Center Herzelia, Israel
[d] Ohalo College, Israel
[e] Department of Behavioral Sciences, School of Social Sciences and Humanities, Kinneret Academic College, Israel

ARTICLE INFO

Article history:
First received on June 16, 2019 and was under review for 9½ months
Available online 7 December 2020

Area Editor: Bernd Skiera

Keywords:
Neuromarketing
Electroencephalography (EEG)
Consumer neuroscience
Preference prediction
Machine learning
Forecasting

ABSTRACT

A basic aim of marketing research is to predict consumers' preferences and the success of marketing campaigns at the population-level. However, traditional marketing tools have various limitations, calling for novel measures to improve predictive power. In this study, we use multiple types of measures extracted from electroencephalography (EEG) recordings and machine learning (ML) algorithms to improve preference prediction based on self-reports alone. Subjects watched video commercials of six food products as we recorded their EEG activity, after which they responded to a questionnaire that served as a self-report benchmark measure. Thereafter, subjects made binary choices over the food products. We attempted to predict within-sample and population level preferences, based on subjects' questionnaire responses and EEG measures extracted during the commercial viewings. We reached 68.5% accuracy in predicting between subjects' most and least preferred products, improving accuracy by 4.07 percentage points compared to prediction based on self-reports alone. Additionally, EEG measures improved within-sample prediction of all six products by 20%, resulting in only a 1.91 root mean squared error (RMSE) compared to 2.39 RMSE with questionnaire-based prediction alone. Moreover, at the population level, assessed using YouTube metrics and an online questionnaire, EEG measures increased prediction by 12.7% and 12.6% respectively, compared to only a questionnaire-based prediction. We found that the most predictive EEG measures were frontal powers in the alpha band, hemispheric asymmetry in the beta band, and inter-subject correlation in delta and alpha bands. In summary, our novel approach, employing multiple types of EEG measures and ML models, offers marketing practitioners and researchers a valuable tool for predicting individual preferences and commercials' success in the real world.

* Corresponding author.
E-mail address: adamhak7@gmail.com (A. Hakim).

https://doi.org/10.1016/j.ijresmar.2020.10.005
0167-8116/

1. Introduction

Many researchers and practitioners aim to measure neural and physiological activity to predict future decisions or to assess the success of possible marketing campaigns at the population-level (Genevsky & Knutson, 2018; Hsu & Yoon, 2015; Plassmann, Venkatraman, Huettel, & Yoon, 2015; Smidts et al., 2014). In recent decades, extensive efforts have been invested to identify, using both electroencephalogram (EEG) and functional magnetic resonance imaging (fMRI), which neural factors are most crucial to the formation of subjective values that generate preferences and drive choices (Bartra, McGuire, & Kable, 2013; Levy & Glimcher, 2012). In the current study, we propose a novel strategy for analysis and use of EEG neural signals combined with machine learning (ML) algorithms, to examine factors and modeling approaches that would inform marketing scholars and industry practitioners on how to yield better and more accurate predictions of consumer choices.

There are several traditional market research tools that aim to predict the success of products. These methods are often behavioral and subjective, and include mainly self-reports as elicited by questionnaires, focus groups, and interviews. Questionnaires are easy to distribute, cost efficient, practical, speedy, scalable, and enable access to enormous cohorts of subjects, relative to neuroscientific methods (Birmingham & Wilkinson, 2003). Interviews and focus groups offer valuable insights, yet they often fall short in producing quantitative and reliable measures for prediction. Although self-reports have significant benefits, in some cases they may contain problematic issues.

For example, with regards to questionnaires and interviews, different preference elicitation methods can result in different responses (Buchanan & Henderson, 1992; Day, 1975; Griffin & Hauser, 1993; McDaniel & Kolari, 1987; McDaniel, Verille, & Madden, 1985), using questionnaires can be biased or inaccurate (Fisher, 1993; Johansson, Hall, Sikström, Tärning, & Lind, 2006; MacKenzie & Podsakoff, 2012; Neeley & Cronley, 2004; Nisbett & Wilson, 1977) and choices may not be incentive-compatible due to high cost or unavailability of relevant incentives. With regards to focus groups, dominant individuals within a group permits only one or a few opinions to be heard, skewing the results towards the most outspoken preferences. Moreover, group dynamics may obscure some of the controversial perspectives, due to the tendency of participants to avoid controversial topics and to produce normative discourse when in groups. Additionally, conclusions rely on the researcher's subjective judgement (Smithson, 2000), representative focus groups can be difficult to assemble, and they are not fully confidential or anonymous (Gibbs, 1997).

In order to overcome these problems, researchers attempt to identify neural and other physiological measures that would increase the prediction power of consumers' choices and population-level preferences above traditional self-reports. In fact, business applications of neuromarketing have been around for more than 15 years. There are numerous companies in the industry who use various neural and other physiological measures to improve prediction above the traditional self-reports (see List in NMSBA Website). Unfortunately, often these companies do not provide data with which to evaluate their predictions, nor the algorithms and methodology they apply, due to obvious IP issues. Therefore, it is crucial, in our opinion, to demonstrate this proof of concept and to provide a clear, detailed, and open description of our effort to employ neural information for the prediction of preferences (our data and code is freely available here: https://bit.ly/2Ts48d6). It is also important to note that our goal is not to replace current marketing tools, which contribute to market predictions, but rather to add an additional layer of analysis. Neural data can provide information that is unobtainable using the traditional self-reports, as we aim to show in this study. It adds information about unconscious processes and reduces possible biases that are known to occur when eliciting self-reports.

The interest in neuromarketing stems from various business applications for marketers and managers. For example, assuming there are several versions of a specific commercial, and the marketer is trying to decide on which one to invest the advertising budget. One could present the different versions of the commercial to a group of people and obtain traditional self-reports using a questionnaire as the one in our paper. We show in the current study that combining the questionnaire assessment *together* with the EEG measures we propose, results in a more accurate prediction of which is the superior commercial. Importantly, using the same algorithm and input features, we increase prediction accuracy both at the individual level and at the population level. In principal, the same procedure can be applied to any marketing campaign or research question as an additional layer of information to what would be usually obtained by a standard A/B testing or focus groups. For example, applications may range from understanding which marketing message will succeed the most, what packaging to use, the best location on the shelf, or the color structure or text to use on a webpage design. Finally, by using the trained algorithm we provide in the current study, anyone can record EEG signals of subjects while they observe any set of marketing-related alternatives, extract the measures we suggest here using the code we provide, and generate a prediction – which of the alternatives is most likely to be chosen for individuals within the subject pool and at the population-level.

In brief, in the current study, we focus on using ML algorithms to identify predictive features from various EEG measures, collected while subjects watched commercials. We examined whether these measures contained information helpful in predicting future choices between the products advertised, that would increase prediction rates obtained by traditional self-reports alone. We examine this both by attempting to predict subject's actual choices after commercial viewings, and by trying to predict population-level preferences over the products, as assessed by the commercials' YouTube metrics and self-reports from an online out-of-sample cohort. Importantly, we also address various modeling challenges and examine a variety of ML approaches. By applying various models on multiple types of EEG measures, we hope to elucidate which are most effective and appropriate for value prediction. Potentially, this could inform management and marketing researchers as to the **explanation** behind predictive EEG measures and their modeling, but mostly this provides **prediction** tools for individual and population preferences, two efforts that Yarkoni and Westfall importantly distinguish between in behavioral sciences (Yarkoni & Westfall, 2017).

2. Novelty & comparison with previous literature

In this section we describe relevant previous studies and compare them to the current study and highlight the novelty in our study. We also summarize the most relevant literature in Table 1 and provide a summary of the following overview of methodologies in appendix A ("Methodological Concerns"). Additionally, readers can find an overview of the novelties of our approach to prediction modelling, along with comparisons to relevant previous literature, in appendix B – "Prediction Modeling Approach and Challenges". There we discuss the advantages of using ML models versus standard regression models, and the importance of performing model testing rather than fitting. Those curious in a comparison between using EEG and fMRI for the purpose of preference prediction, can read our examination in appendix C – fMRI or EEG.

2.1. Using EEG to increase prediction accuracy obtained by self-reports

Despite the fact that EEG is commonly used in the neuromarketing industry (see NMSBA Website), and that there is accumulating data linking various EEG signals with value-based choice (Dmochowski, Sajda, Dias, & Parra, 2012; Fuentemilla et al., 2013; Khushaba et al., 2013; San Martin, Appelbaum, Pearson, Huettel, & Woldorff, 2013; Sutton & Davidson, 2000), only several academic studies attempted to predict individual preferences or actual choices (Kong, Zhao, Hu, Vecchiato, & Babiloni, 2013; Ravaja, Somervuori, & Salminen, 2013; Telpaz, Webb, & Levy, 2015; Vecchiato et al., 2011; Yadava, Kumar, Saini, Roy, & Prosad Dogra, 2017; Wei et al., 2018), or population-level preferences (Barnett & Cerf, 2017; Boksem & Smidts, 2015; Christoforou, Papadopoulos, Constantinidou, & Theodorou, 2017; Dmochowski et al., 2014; Guixeres et al., 2017; Venkatraman et al., 2015). However, importantly, all these previous studies (apart from Venkatraman et al., 2015; Christoforou et al., 2017) did not examine if their prediction accuracy based on EEG increased the prediction accuracy of traditional self-reports, which is critical for establishing it as a marketing tool.

Venkatraman et al., (2015) were unsucesfful in employing EEG to the prediction of advertising elasticity beyond traditional measures, while Christoforou et al., (2017) were able to successfully increase prediction accuracy of movies' box-office performance based on the EEG, but they used standard regression models without a test sample, a single type of EEG measure (Cognitive-Congruency) and did not attempt to predict individual preferences. Other studies used EEG mea-

Table 1
Comparison to most influential studies on EEG-based preference prediction. The table summarizes key methodological components of each relevant study. The comparison demonstrates that our study is unique in showing successful contribution of EEG measures to prediction based on self-reports, to employ multiple advanced ML models, to perform reliable and thoroughly reported testing, to employ multiple types of EEG measures, to include an applicable EEG setup, and to predict reliable measures of both individual and population level preferences.

Study	Demonstrated Contribution Beyond Traditional Marketing Tools?	Prediction Technique	Testing or Fitting?	Types of EEG measures	#Electrodes	Predict Within Sample Preferences	Predict Population-level Success
Current Paper	Standard ARF Questionnaire	Various Machine Learning Models	Testing	5 Spectral Power Bands, 5 ISC, 5 Bands Asymmetry	8 electrodes	Binary Choice Rankings	YouTube Metrics, Online Questionnaire Rankings
Boksem and Smidts (2015)	WTP, Liking	Regression	Fitting	100 Spectral Power Bands	64 electrodes	Ordered Preferences	Box Office Income
Barnett and Cerf (2017)	WTP, Free Recall, Ratings	Correlation	Fitting	1 CBC (similar to ISC)	32 electrodes	Ad Recall	Movie Sales
Venkatraman et al. (2015)	Standard ARF Questionnaire – Unsuccessful Contribution	Regression	Fitting	1 Band Asymmetry, 1 Spectral Power Band	129 electrodes	None	Commercial Elasticity
Guixeres et al. (2017)	None (collected ad recall and liking, but were not used to demonstrate contribution beyond them)	1 Layer MLP	Testing	5 GFP, 2 Asymmetry, 1 count of peaks	32 electrodes	None	YouTube Views
Wei et al. (2018)	None	SVM	10% Test set	Band Powers, Indexes,	1 electrode (headband)	Questionnaire Responses	None
Ramsøy et al. (2018)	None	Regression	Fitting	Gamma & Beta Frontal Asymmetry	14 electrodes	WTP Values	None
Shestyuk et al., 2019	None	Regression	Fitting	Frontal Asymmetry and Band Powers	32 electrodes	None	TV Viewership / Twitter Volume

sures to predict or explain preferences (Barnett & Cerf, 2017; Boksem & Smidts, 2015; Ramsøy, Skov, Christensen, & Stahlhut, 2018). However, they did not compare their predictions to a questionnaire (multiple items), but rather to a single item measure, such as subjects' ratings and likings (which are still commonly used in the advertising industry) or willingness to pay (not a standard practice in the advertising industry).

2.2. Multiple EEG measures

Another novelty in the current study, is that we developed a method that uses several EEG measures in order to increase the prediction power of the EEG signal. Most previous studies (Barnett & Cerf, 2017; Boksem & Smidts, 2015; Ramsøy et al., 2018) focused their analyses only on a single type of EEG measure, or two at most (Shestyuk, Kasinathan, Karapoondinott, Knight, & Gurumoorthy, 2019; Venkatraman et al., 2015). Shestyuk et al. (2019), for example, used EEG measures based on frequency band powers and hemispheric asymmetry to predict twitter volumes and TV viewership of various episodes from shows, but did not utilize a measure for inter-subject correlations. Their results demonstrated that the combination of their different types of measures did contribute to prediction of engagement with the episodes (however, only via fitting regression models).

We suggest that a combination of several types of EEG measures would increase predictive power, because each measure captures a different cognitive aspect of the valuation process that in combination generates a better proxy for the overall value signal that a subject construct towards a marketing message. Specifically, we combine information on frequency band powers for estimating valuation (Boksem & Smidts, 2015; Braeutigam, Rose, Swithenby, & Ambler, 2004; Khushaba et al., 2012, 2013; Ravaja et al., 2013; Yadava et al., 2017), hemispheric asymmetries for estimating approach/avoidance tendencies (Davidson, 1998; Laurence & Gerhold, 2016; Ramsøy et al., 2018; Ravaja et al., 2013; Vecchiato et al., 2011), and inter-subject correlations (Barnett & Cerf, 2017; Chan, Smidts, Schoots, Dietvorst, & Boksem, 2019; Dmochowski et al., 2014; Hasson, Nir, Levy, Fuhrmann, & Malach, 2004) for estimating engagement. These measure types, which capture different aspects of the valuation process, have never been used in combination for the prediction of preferences. The only other paper that used several measure types was Guixeres and colleagues (Guixeres et al., 2017). Although this study contributed greatly to our research field, in contrary to the current study, they used only a few EEG measures, did not predict using the same algorithms both individual and population level preferences, and did not compare their predictions to self-reports.

2.3. Predicting preferences of individuals and at the population-level

The combination of both individual and population-level predictions in the same study has not been attempted in nearly all previous papers (except for Boksem & Smidts 2015, albeit using a regression analysis and WTP). This offers an important contribution not just for the industry but for the basic understanding of consumer's behavior, as this shows that the value representation of a small group of subjects could be generalized to the population.

2.3.1. Industry applicability – electrode array

Our study is also industry applicable relative to many previous studies. While some studies suggest that eight electrodes are insufficient for various measures (Lau, Gwin, & Ferris, 2012), we found that EEG could contribute to prediction even with a smaller number of electrodes. In the industry, many neuromarketing companies use very simple and mobile headsets with only a few (dry) electrodes. In most cases, these devices have poor signal-to-noise ratio and limited validity. We do not recommend using these devices, and we believe that quality devices with dense electrode arrays are the most reliable and valid. However, we also understand the allure in using smaller number of electrodes. Hence, in the current study, we attempted to balance the reliability of a quality EEG system (Neuroelectrics, Spain) with practicality (using a mobile device with only a few wet electrodes), without resigning to cheap and undependable devices. Mounting an EEG device on multiple subjects could be strenuous for marketers, particularly in comparison to administering a questionnaire. Each additional electrode comes with various costs: time to apply the gel and place each electrode, and verifying the quality of each electrode; resources such as replacing eroded electrodes, and computational load from analyzing additional time-series signals; dense layouts could negatively affect the overall experience of subjects due to discomfort, which could make it more difficult to identify positive responses to the stimuli.

In academia, the use of a dense electrode array is done mainly to be able to use source localization algorithms in order to identify the brain areas which the signal originates from (Song et al., 2015), and to improve signal quality and cleaning through interpolation procedures (Luck, 2014). However, for prediction purposes, we do not need to know the exact brain location of the signal. Additionally, electrodes that are physically close to each other are highly correlated, on the order of $r = 0.8–0.98$ (Bhavsar et al., 2018). Therefore, a common practice in a dense array setup is to average the signal across several electrodes before conducting any analysis, hence eventually looking only at a limited number of input channels. Nevertheless, we agree that using more electrodes while covering additional scalp locations will likely increase any neural-based prediction attempts. We advise practitioners to use dense arrays when possible and affordable and consider recording both central and occipital activity to obtain a richer representation of neural responses. In the current study, we demonstrate that the contribution of EEG measures to prediction with a lean and cost-effective 8 (wet)-electrode array is possible, but it is probable that the predictions would improve had we used more electrodes.

2.3.2. Industry applicability – sample size

We used a subject pool of 33 participants, which is a common sample size in consumer neuroscience based prediction studies (Guixeres et al., 2017; Koelstra et al., 2012; Ravaja et al., 2013; Venkatraman et al., 2015; Yadava et al., 2017). This relatively small dataset can be considered a limitation of this study and could restrict the robustness of our results. However, we aim to demonstrate that significant prediction can be achieved without the cost of acquiring data from hundreds of subjects, as is common in some behavioral and online studies (Musch & Reips, 2000). Reaching useful results on a small data set can help persuade practitioners that EEG-based predictions can be helpful, applicable, and practical.

In summary, there are excellent previous studies that used EEG signals in order to predict some aspects of population-level preferences (Barnett & Cerf, 2017; Boksem & Smidts, 2015; Guixeres et al., 2017; Ramsøy et al., 2018; Venkatraman et al., 2015). However, these studies did not employ all of the advantages and methods we utilize in the current study. Our study is unique as we increase the prediction accuracy obtained by a questionnaire using multiple types of EEG measures, acquired using a portable 8 wet electrode EEG headset, using ML, for both individual and population level preferences. Lastly, as a managerial implication, we freely provide all our data and code for others to use.

3. Materials & methods

3.1. Experimental procedure

Thirty-three subjects (13 males) participated in the study, aged 19–41. We excluded two subjects out of the original 33 from the analysis due to lost recordings. All subjects gave written informed consent before participating in the study, which was approved by the local ethics committee at our university. We designed the first stage of the experiment to mimic as much as possible the real experience of watching TV. During this stage, subjects watched six cycles that included three consecutive skits from a comedy series followed by three consecutive local commercials regarding six food products. Prices of the food products ranged between \$1.6 and \$3 (6–12 NIS), making them relatively similar in real-world value. Further information on the products, including images and links to the commercials, are provided in appendix D (“Products Information”). The order of the skits and commercials was chosen randomly. Overall, subjects watched six different commercials, three times each (for a total of eighteen commercial views), and a total of twenty-seven different skits (Fig. 1). We chose three exposures per commercial as a compromise between strengthening the statistical power and increasing the signal to noise ratio of the EEG signal, and the loss of the measures’ reliability due to disengagement brought by excessive exposure (however, note, that it is common that the same commercial would repeat during a TV episode). The length of each skit was between 24 – 100 seconds and the length of each commercial was between 25 – 46 seconds. The total duration of the first stage was 30 minutes.

While watching, subjects were connected to an 8-(wet) electrode EEG system at positions F7, Fp1, Fpz, Fp2, F8, Fz, Cz, Pz, sampled at 500 Hz. The EEG device is named StartStim 8, manufactured by the Spanish company Neuroelectrics, and has a bandwidth of 0–125 Hz (DC Coupled), resolution of 24 bits – 0.05 μV, measurement noise is below 1 μV RMS, common mode rejection ratio is 115 dB and input impedance is 1GΩ (more information on the device and layout in appendix E – “Electrode Array”). The EEG recordings were processed and used to perform predictions of subjects’ later choices and the population-level preferences. We focused mainly on the frontal electrodes because they are easy to apply, require less gel, and the simple headsets used in the industry usually have electrodes only in frontal areas. However, we concede that not all EEG measures can be optimally assessed using these electrodes, and therefore recommend opting for denser arrays when possible.

Immediately after the end of the first stage, subjects answered a questionnaire regarding each product for 15 minutes. The questionnaire is based on an ARF (Advertisement Research Foundation in New York) questionnaire, which is considered credible, reliable, and with high validity. It was the exact questionnaire format used in Venkatraman et al., (2015) paper, only translated to Hebrew (the full questionnaire, translated back to English, is provided in appendix F – “Questionnaire Summary”). The questionnaire included questions both on the product itself and on the commercial. Note, that our aim was not to test the effectiveness of the commercials in changing preferences (hence, we did not acquire the same measures before and after observing the commercials), but rather to provide evidence that the EEG recordings acquired during com-

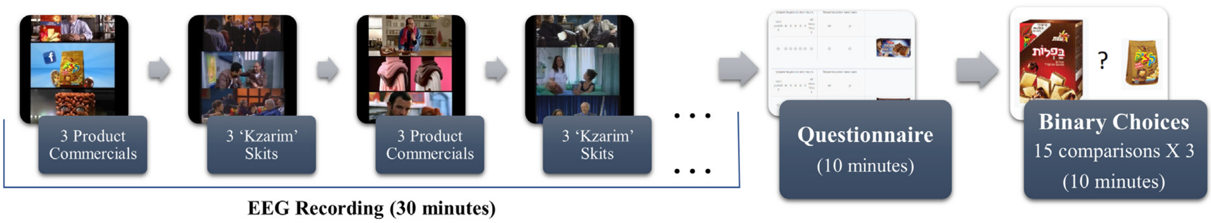

Fig. 1. *Procedure*. Subjects watched three randomly chosen product commercials, followed by three skits, until three viewings of each of the six unique commercials were completed. Thus, subjects watched a total of 18 commercials. After viewing, they filled a questionnaire regarding each product that appeared in the commercials and on the commercial itself. Finally, they performed a binary choice task between the products that appeared in the commercials.

mercial viewing can increase prediction accuracy obtained using the questionnaire only. In addition, we chose a fix order in which subjects first answered the questionnaires and then conducted a binary choice task (see below) to mimic standard marketing practices. In summary, we used the answers from the questionnaire as a baseline for our prediction attempts, which we aimed to improve by adding the EEG measures we extracted while subjects viewed the commercials in the first stage.

Lastly, in the third stage of the experiment, after answering the questionnaire, subjects completed a binary choice task between all product pairs. There were 6 different products, each appearing in one of the six commercials, resulting in a total of 15 unique product pairs. Each pair was presented 6 times, for a total of 90 binary choice trials. Each trial consisted of 0.5 seconds of fixation on a red cross, then a product pair was presented for a maximum of 3 seconds. The subject chose a product by clicking either the left or right mouse buttons. If she did so before the 3 seconds were over, the remaining time was used to present a feedback message with the chosen product. If the subject did not choose within the time limit, the trial was considered a mistrial and was excluded from the analyses. This was followed by a one second black screen as an inter-trial interval. At the end of the experiment, one binary choice trial was chosen randomly, and the subject was given the product she chose on that trial. Since subjects received a product based on their choices, we consider our experimental design to be incentive-compatible, in the sense that subjects were incentivized to choose based on their subjective valuations so that they eventually receive a product they actually desire.

We used subjects' choices from the third stage to obtain the rank order of products' preferences for each subject (from 1, the most preferred product, to 6, the least preferred product). We then attempted to improve prediction obtained using the questionnaire alone, of subject-specific ranked order preferences and the population-level preferences (see below) using the EEG data that was collected while subjects viewed the commercials in the first stage.

3.2. EEG preprocessing

Before applying any ML analyses, we preprocessed the EEG signal. EEG recordings were referenced to the Cz electrode and underwent 0.1 Hz high-pass filtering and a 50 Hz notch filter. Then, signals were transformed through Independent Component Analysis (ICA), and the component most related to eye movements and blinks was subtracted from other components, before transforming components back to the electrode space. Next, we performed Raw Data Inspection to mark apparent artifacts for removal in later processing. Finally, MATLAB's "spectrogram" function was applied to perform Short-Time Fourier Transform (STFT), on each electrode separately, with a window of 2 seconds (1000 samples), and maximal overlap (999 samples). As the function outputs the signals' power in various frequencies, power signals were then aggregated into well-known EEG frequency bandwidths Luck (2014), by choosing the signal of the frequency with maximal power in each bandwidth. The ranges of the bandwidths we chose were: Delta 1–3.5 Hz; Theta 4–7.5 Hz; Alpha 8–12 Hz; Beta 13–25 Hz; Gamma 26–40 Hz. The final product of the preprocessing stages was power signals in the five frequency bands for each electrode and each commercial viewing separately, for every subject.

3.3. Feature extraction

The term "features" is used to describe the various measures we extract from our data through different processing procedures and use these features as predictors.

3.3.1. EEG measures

In contrast to previous studies, we extracted several (as opposed to one) measures from the EEG signal, to gain as many aspects of the individual's evaluation process as possible, in order to maximize our prediction accuracy. It is important to note that we chose to use measures from all five frequency bands for prediction, although only some bands were shown to be related to preference prediction in previous studies. The reason is that we did not want to devoid our models with possibly informative measures and preferred that the models themselves inform us which frequency bands were most predictive. The following EEG measures were extracted from recordings while subjects viewed commercials of products, therefore we hypothesized that they would contain information about both subjects' valuations of the commercials and their a-priori preference toward each product. We do not think these two subjective experiences could be disentangled in our design, but that they would still prove predictive of our chosen labels.

3.3.1.1. Frontal Band Powers (FBP). Powers for each of the five frequency bands (Delta, Theta, Alpha, Beta, and Gamma) were averaged across time per commercial viewing. Some of these measures have been shown to be related to value representations in previous studies (Boksem & Smidts, 2015; Braeutigam et al., 2004; Khushaba et al., 2012, 2013; Koelstra et al., 2012; Luck, 2014; Ravaja et al., 2013; Smith & Gevins, 2004; Telpaz et al., 2015; Yadava et al., 2017). To increase our external validity, we focused on the frontal electrodes to resemble various EEG systems which are commonly used in the industry, that have only a few frontal electrodes. We extracted EEG data from the 3 frontal electrodes – FP1, FP2, FPz, and extracted the maximal activity between the three electrodes for each band, yielding a total of 5 features per commercial viewing.

3.3.1.2. Hemispheric ssymmetry. We calculated, for each frequency power band, the difference between the band powers of the most fronto-lateral electrodes in our setup, F7 and F8. This yielded 5 additional features, out of which alpha-band asymmetry has been related to approach-avoidance behavior (Cartocci et al., 2016; Davidson, 1998; Koelstra et al., 2012; Ohme, Reykowska, Wiener, & Choromanska, 2009, 2010; Ramsøy et al., 2018; Ravaja et al., 2013; Sutton & Davidson, 2000; Vecchiato et al., 2010, 2011; Venkatraman et al., 2015).

3.3.1.3. Inter-subject Correlations (ISC). Inter-subject correlation was previously used as a measure of engagement but was calculated using different procedures (Barnett & Cerf, 2017; Chan et al., 2019; Dmochowski et al., 2014; Hasson et al., 2004). Our ISC score is created for each specific viewing (repetition) of a commercial. For each subject, frequency band, product, and viewing, we took their power-time-series, and cross-correlated it with the averaged power time-series of the same commercial viewing from all other subjects (excluding the given subject). The cross-correlation resulted in a correlation time-series, from which we took the maximal correlation value across all timepoints, for each frequency band, yielding a total of 5 ISC scores (1 for each frequency band) per commercial viewing for each subject. Thus, for each power band, each subject has a measure of ISC that reflects how much her power time series is similar to the rest of the subjects in a specific viewing of a specific commercial.

3.3.2. Self reports (Questionnaire)

Questionnaire responses were aggregated into eight general measures for each commercial and the product appearing within it: Commercial Negative Elicitation, Commercial Positive Elicitation, Product Liking, Purchase Intent, Product Recall, Familiarity, Commercial Overall impression, and Product Overall Impression. Some measures reflect valuations of the commercials themselves while others reflect preferences towards the viewed product, while some combine both. Our EEG measures also reflect the combined preferences and are therefore on equal footing. Afterwards, we performed a Principal Component Analysis (PCA) on all measures of the questionnaire combined, for each subject. In a separate analysis, we show that combining responses on commercials together with responses on the advertised products when performing PCA, yielded the most predictive components (see appendix G – Product and Commercial Questionnaire Responses). We then extracted the scores for the first three components, which were the only ones that showed some correlation to subjects' rankings (Fig. 3) and together exceeded 75% of the explained variance. We used the three components' scores in order to perform the predictions.

3.4. Label creation

3.4.1. Subject rankings

Based on each subject's choices in the binary choice task in the third stage, we ranked ordered each subject's preferences for the six different products. We ranked the responses such that the product that was chosen the most out of all comparisons was ranked first (Rank 1), and the product that was chosen the least was ranked sixth (Rank 6). We used these six rankings as labels for predictions that reflect subject's preferences.

We chose the rankings as the ground-truth for individual preferences as it was based on incentive-compatible and repeated choices per subject and per unique pair-wise comparison. We think this provides an approximation of subjects' future preferences that we aim to predict. In contrast, we used the questionnaire, which is a common technique in marketing research for estimating subjects' preferences, as the benchmark for the standard practice in the industry. The aim was to show that we can improve upon the prediction of the questionnaire if we add the EEG data to it. Thus, these rankings reflect preferences over products within our subject pool, as formed after viewing the commercials.

3.4.2. Out-of-Sample rankings

To obtain an *Out-of-Sample* preference ranking regarding the commercials and the products within them, 172 subjects watched the commercials online (via Qualtrics, Provo, UT) and then answered the exact same questionnaire as our in-lab subject pool. Subjects were all Israeli residents, Hebrew speakers, average age 29.3 (S.D = 6.64), 114 were female, 71% were active students in various academic institutions. We aggregated the responses from the online cohort of subjects into an overall ranking of the six commercials (between 1–6). We then used the aggregate overall *Out-of-Sample* preference ranking as labels for prediction, under the assumption that they reflect an average general preference towards the commercials and the products advertised within them at the population level.

3.4.3. YouTube metrics

We gathered publicly available metrics from YouTube, regarding each of the commercials. These YouTube metrics served as a proxy for how popular and favorable the commercials and the products within them were at the population-level. The YouTube metrics we collected were the number of likes, dislikes, comments and views. We down-weighted all metrics of a commercial the longer the commercial has been online, by multiplying each metric of each commercial with the ratio between the number of days the oldest commercial has been online divided by the number of days that that commercial was online. Dividing a metric by the number of days a commercial had been online, resulted in the metric per day for the commercial. Thereon, multiplying the result by the number of days of the longest commercial had been online, resulted in an approximation of the metric of a commercial, had it been online as many days as the longest commercial. Accounting for the duration each commercial had been online could be tricky, so we collected all the metrics on the same day, once the newest commercial had already been online for 2 years and 2 months and the oldest commercial had been online for 5 years and 6 months. Since all commercials had been online fairly long, we assumed that their metrics roughly plateaued at the time of data collection, enough to create reliable ordered rankings of the six commercials.

Thereafter, in order to gain relative values of a specific metric across all six commercials, we divided each specific metric in a given commercial by the sum of that same metric in all other commercials. Then, we performed a PCA on the relative metrics, and used the summed score of the two first components, which together explained above 75% of the variance, as the aggregate score for each commercial. These aggregate scores were finally converted into six ordered rankings of the products' (and their commercials') success in YouTube, based on the order of scores. We use these overall rankings as labels for prediction.

3.5. Prediction models

We attempted various ML models, in search for the model that best suited our data and achieved highest predicted accuracy. Most of these models' description and code may be found online thanks to the AYRNA research group (Gutierrez, Perez-Ortiz, Sanchez-Monedero, Fernandez-Navarro, & Hervas-Martinez, 2016) and the rest were obtained with the relevant MATLAB functions. We used the default values for all unspecified parameters. The complete code that runs all the models can be found in the link provided, under folder "MLCodes", named "CreateMLTable.m". Note that some models perform a binary classification, meaning classifying data into two possible classes (labels), some perform a multi-class categorical classification, classifying data into more than two classes, and other perform an ordinal classification, classifying data into multiple classes while considering the rank order between them. These included the following models:

3.5.1. Binary classification models (prediction of two classes):

Support Vector Machine (SVM) with a linear kernel, Logistic Regression (LOG), Boosted Decision Trees with Adaboost M1 (TREE), with 100 trees and a minimum leaf size of 5 for regularization, and K-nearest Neighbor (K-NN) with $K = 5$.

3.5.2. Nominal multi-class classification models (prediction of more than two categories):

One versus all support vector machines (SVC1VA), Multinomial Regression (MNR), K-NN, and Boosted Decision Trees with Adaboost M2 (TREE) and 100 trees.

3.5.3. Ordinal classification models:

These models consider the ordinal nature of rankings (1 to 6). We used an ordinal support vector machine (SVOREX), and Kernel Discriminant analysis for ordinal regression (KDLOR), with a radial basis function as a kernel.

3.6. Prediction approach

Both the EEG measures and the questionnaire responses within our subject pool reflect valuations formed for each product and attitude towards the commercials. We hypothesized that the mixture of these preferences as captured by our EEG measures could improve prediction accuracy obtained using only the questionnaire for the following labels: (1) **Subject rankings** (acquired from the binary choice stage) of the products that appeared in the commercials. This served as our proxy for the individual-level future preferences. (2) The average responses of an **out-of-sample** group of subjects to the same questionnaire, which represents a combination of the preference towards both the commercial and the product. This served as our attempt to predict the average response to a questionnaire at the population level. (3) Aggregated **YouTube metrics** of each commercial, which also reflect a combination of both the success of the commercial and the popularity of the product at the population level.

The final data matrix for prediction included several types of features depending on the specific prediction procedure and included a different number of samples according to the type of labels predicted. Generally, each row of the final data matrix constituted a sample containing features from a specific commercial viewing of a particular subject and was assigned a label according to the rank given to that product either by the subject, by the online questionnaire responses for the commercial in that viewing, or by the YouTube metrics for the commercial.

3.6.1. Prediction of highest vs. lowest rankings

For binary prediction between the least and most preferred products, the final data matrix had 183 rows (or samples) – 31 subject's X 2 Product Commercials X 3 viewings per commercial = 186 rows; We excluded 3 rows due to extreme noise in the EEG recordings.

3.6.2. Prediction of all six rankings

For prediction of all six rankings, the final data matrix had 549 rows (or samples) – 31 subject's X 6 Product Commercials X 3 viewings per commercial = 558 rows; 9 rows were excluded during the preprocessing stage due to extreme noise in EEG recordings. Each sample was labeled with a rank from 1 to 6, according to the ranking that was given to the viewed commercial in that sample, by either the subject herself, by the aggregate YouTube score, or by the aggregate Online responses.

3.6.3. Training and testing

The features per sample also changed within each type of label we tried to predict, according to the type of measures used for prediction. For each label we tried to predict, we based prediction on either only the questionnaire responses (3 features – 3 PCA components), only the EEG measures (15 features), or the combination of both (18 features). We performed the predictions by splitting the data randomly into ~15% test set and ~85% train set, while adhering to the principles raised in the "modeling challenges" section in appendix B. That is, for each prediction we randomly excluded to the test set all three viewings of one random product commercial per subject, for 6 randomly chosen subjects in the binary predictions (16.4% of 183 rows of data to test) and for every subject in the multi-class predictions (16.7% of 549 rows of data to test).

We trained the models on the train set, used it to predict the test set, and compared predictions to the test set labels to obtain prediction accuracy – the percentage of correct predictions (for binary predictions), or the root mean square error (RMSE) for the prediction of all six ranks. RMSE is calculated as the square root of the sum of squared differences between each prediction to its corresponding true label. Note, that a lower RMSE score indicates a better prediction. We used RMSE instead of prediction accuracy to account for the *magnitude* of the prediction errors between the six different rankings. That is, when the algorithm assigns a rank of 2 when the true rank is 3, this should be considered a small error compared to when the algorithm assigns a rank of 6 when the true rank is 3. Note, that we lose the sensitivity gained by RMSE scores if we use simple prediction accuracies for predicting more than two classes.

Our model training and predictions were tested for significance via permutation testing (Good, 2013; Nichols & Holmes, 2002). For each feature combination, we repeated the random train-test split and prediction 10,000 times, to form a distribution of accuracies that was unbiased to any specific train-test split. The accuracies presented are the distribution means and standard errors. Also, we performed the exact same procedure after shuffling all labels of the data. This enabled us to obtain a randomized 'baseline' distribution of predictions to improve upon with real unshuffled predictions, that provides an empirical approach to deriving statistical significance (Golland & Fischl, 2003; Ojala & Garriga, 2010), particularly for relatively small datasets (Combrisson & Jerbi, 2015). Thus, p-values were calculated as the proportion of random predictions that were higher than the mean of real predictions. This enabled us to perform significance testing between the distribution of prediction accuracies on shuffled labeled and the distribution of prediction accuracies on the true labels. Additionally, we performed significance testing between subsets of measures within the same procedure, between the distributions of prediction accuracies on true labels of the most accurate model of each subset – Questionnaire, EEG, and all measures together.

3.7. Dataset size considerations

Although there are 31 subjects, there are 18 samples per subject (6 products X 3 repeated viewings), leading to a total of 558 samples. Our maximal number of features is 18, which is considered reasonable to our sample size. It was shown that the number of training samples per class must be greater than three times the number of features per class (Foley, 1972). We have six classes, which results in ~ 90 samples per class, allowing us to use up to 30 features. More modern studies are even more lenient in terms of allowing many more features (Hua, Xiong, Lowey, Suh, & Dougherty, 2005). The models we used are all suitable for small scale datasets, particularly since we did not use any neural network models (as were used in Guixeres et al., 2017) that demand much higher volumes of data. Moreover, as mentioned, we employ a more severe analysis, suggested by Combrisson and Jerbi (2015), designed specifically for small scale datasets. Therefore, we are confident that the number of features is suitable for our prediction attempt, both to avoid the curse of dimensionality while exploiting information from the EEG recording as much as possible. We chose not to use unprocessed EEG signal, since ML models cannot necessarily overcome very noisy signals and a low signal to noise ratio. These methods work best on clean, carefully engineered and interpretable features.

4. Results

4.1. Subject rankings

For meaningful predictions, the correlation of product preferences across subjects must be relatively low, such that each subject ranks the products by her own idiosyncratic subjective evaluation. It is problematic if all subjects have similar preference ordering of the products. If all subjects rank the same product as the most preferred, then predicting that product's individual ranking could be a matter of identifying its commercial's particular identity pattern of neural activation, rather than basing prediction on the value signal it elicited. Fig. 2 shows the mean ranking of each product across subjects. Five out of the six products are well within each-other's confidence intervals, while only one product ("Candy") appears to be significantly better than all other products ($F(30,5) = 20$, $p < 0.001$(. This means that for the most part, there is high variation of product preferences between subjects, such that most products on average are liked the same. The variation in preferences between subjects ensures that we predict their subjective values rather than specific attributes of the commercial that subjects' EEG respond to.

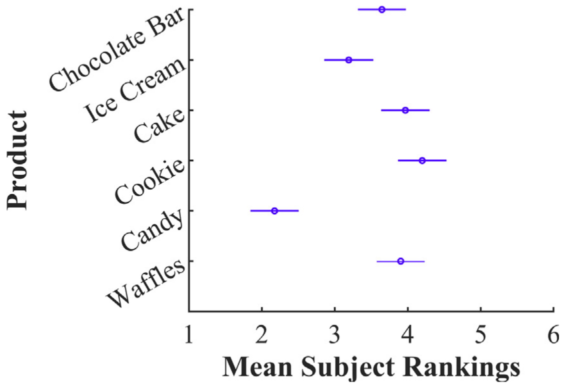

Fig. 2. *Mean Rankings per Product*. Mean rankings and their confidence intervals (alpha = 0.05) for each product in our experiment. Only "Candy" was significantly different than all other products, reaching the best (lowest) ranking of 2.17. (F(30,5) = 20, p < 0.001 on repeated measures ANOVA).

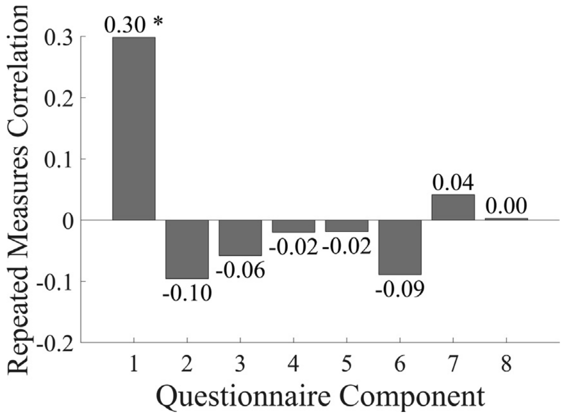

Fig. 3. *Subject Rankings and Questionnaire Principal Components*. The figure shows the repeated measures correlation coefficients (Bakdash & Marusich, 2017) between subject rankings, as obtained by actual choices from the binary choice task, and the Questionnaire's PCA component scores. A moderate connection was found, leaving room for improvement in prediction using EEG measures (* = p < 0. 01). Note, that only the first 3 components were used in the predictions, as explained in the methods section.

4.2. Questionnaire

Before attempting prediction, we explored the relationship between the questionnaire components with subjects' product rankings, as obtained from the binary choice task. It is common practice to explore the relationships between the data and labels, before forming prediction models, to enhance, select, and properly engineer our features. As can be seen in Fig. 3, PCA components of subjects' questionnaire responses were only moderately correlated with subjects' product rankings ($\rho = 0.3$, repeated measures correlation, $p < 0.01$, for the first component, and a $\rho = -0.1$ for the second component n.s., (Bakdash & Marusich, 2017)). This shows that the scores of the questionnaire are related to subjects' actual choices and can be valuable predictors of preferences. Yet, the relatively moderate correlation hints that there is enough room to improve upon using predictive measures of a different type. Additionally, as we show in appendix H ("Questionnaire PCA Inspection"), predictions based on the questionnaire measures after PCA were generally better than without using PCA, likely due to reducing noise by removing components with less explained variance.

4.3. EEG measures

We also examined the relationship between our various EEG measures with subjects' product rankings. Previous studies have shown (Levy, Lazzaro, Rutledge, & Glimcher, 2011; Telpaz et al., 2015) that as the distance between preferences increase, so does the "neural distance", as measured by the neural activity, and therefore prediction accuracy increases. As such, we would expect that for product commercials ranked highest and lowest, our EEG measures would be most disparate, while for those in medium rankings, very little difference would emerge. Indeed, for Delta, Theta, and Alpha frequency bands we found a significant difference ($t(31) = 2.73$; $t(31) = 2.77$; $t(31) = 3.18$ respectively, $p < 0.05$; FDR

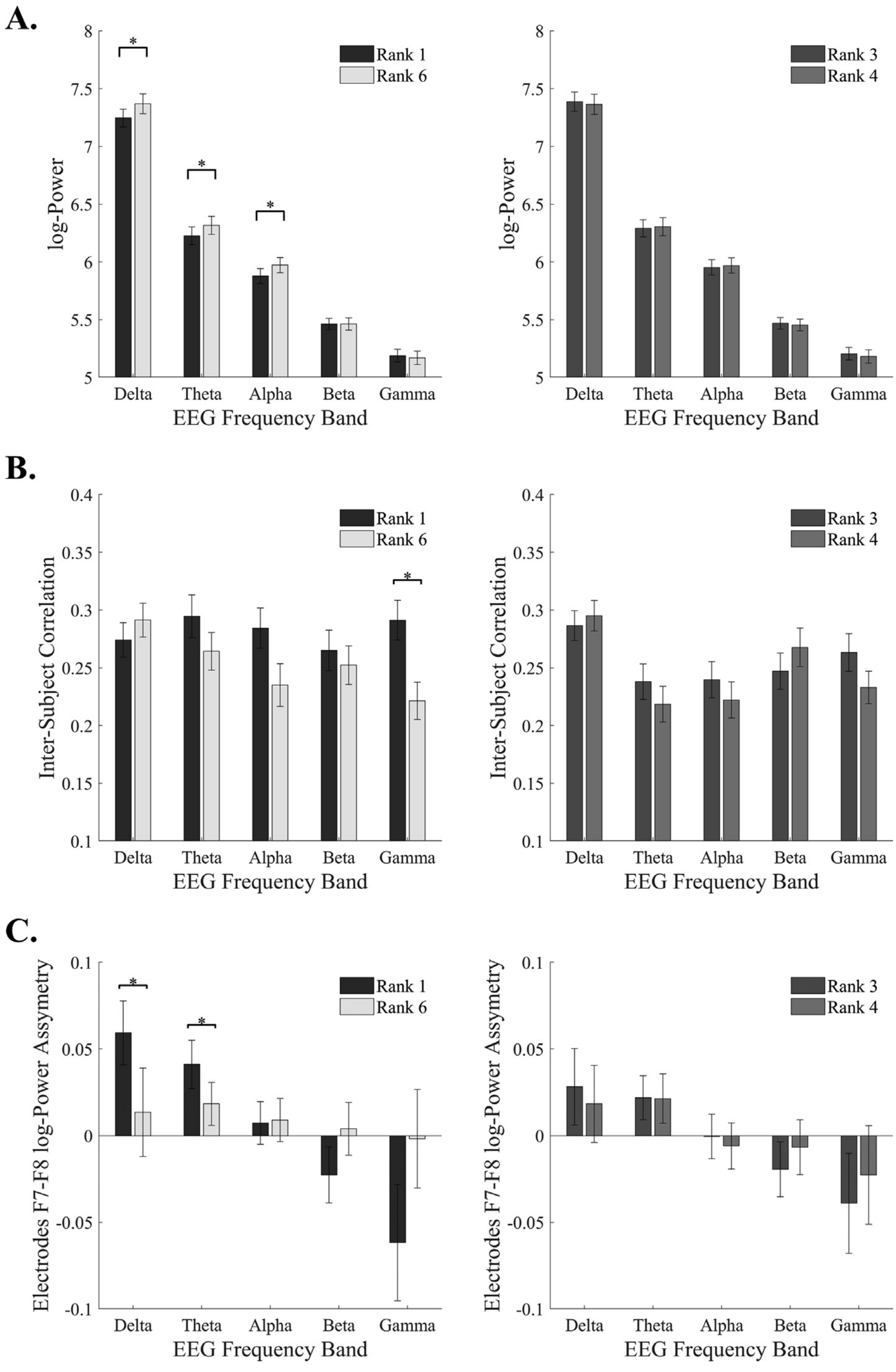

Fig. 4. *EEG Measures for Distant Rankings and Mid-Adjacent Rankings*. In the left column of each figure, we compared the EEG measures when subjects viewed product commercials which they ranked as their most favorite (ranking 1) compared to the least favorite (ranking 6). We repeated the analyses for mid-adjacent rankings (3 vs. 4), in the right columns. (A) Frequency power bands. (B) Inter-Subject correlations. (C) Hemispheric Asymmetry. Note that we found significant differences only when comparing between distant rankings (1 vs. 6) but not when comparing mid-adjacent rankings (3 vs. 4). * = $p < 0.05$, two-tailed paired t-test, corrected for multiple comparisons by Benjamini-Hochberg False Discovery Rate (Benjamini & Hochberg, 1995). Error bars indicate Standard Error of the Mean.

corrected (Benjamini & Hochberg, 1995)) in EEG activity when watching commercials for distant product rankings (rank 1 and 6), but not for close mid-ranged rankings (rank 3 and 4, p = n.s) (Fig. 4A). Furthermore, our ISC measure showed a significant difference only for the distant rankings in the Gamma band ($t(31) = -2.76$, $p < 0.05$; FDR corrected), and a marginally significant difference for the Alpha band ($t(31) = -1.82p = 0.09$; FDR corrected) (Fig. 4B). Lastly, there was a significant difference in the hemispheric asymmetry measure in the Delta and Theta power bands ($t(31) = -2.55$, $t(31) = -2.35$, $p < 0.05$; FDR corrected), and a marginally significant difference for the Gamma band ($t(31) = 2.08$ $p = 0.07$; FDR corrected), again in distant rankings, but not in the mid-range rankings (Fig. 4C). These results are similar to a previous study who found frontal brain asymmetry in gamma to be related to WTP values of 40 different products, albeit they also reported positive findings for the beta band which was not significant in our study, and none for delta, theta, or alpha (Ramsøy et al., 2018). In summary, this demonstrates the potential for some measures of the EEG recording to assist in predicting choices, but mainly for well distinct preferences, as has been shown in previous studies (Levy et al., 2011; Telpaz et al., 2015). Moreover, since each EEG measure had at least one significant difference between rank 1 and rank 6 for a power band, this demonstrates the possible advantage of using multiple EEG measures for prediction.

4.4. Reliability of measures from repeated viewings

Since subjects watched each commercial 3 times, we can investigate how their EEG measures behaved across viewings of the same commercial and inspect whether the EEG measures were reliably recaptured in every viewing of a commercial by the same subject. We found that EEG measures of one viewing, before the transformations outlined in appendix B, were mostly uncorrelated to measures of another viewing, as can be seen in Table 2. However, once we rank-ordered and centered the measures (see appendix B), our measures showed high correlation between viewings, as also shown in Table 2. This demonstrates that EEG measures fluctuated as subjects proceeded between viewings of the same commercial, but the order of EEG measures remained very similar in each viewing session. That is, EEG measures should change between repeated viewings of the same stimulus, but their order within the same viewing session of different stimuli should persist.

However, even though we found moderate to high correlations between transformed EEG measures from different viewings, we found that the addition of each repetition substantially reduced the RMSE score, as detailed in Table 1 of appendix I – "Sensitivity Analysis". This demonstrates that collecting measures repeatedly while presenting the same stimulus could reduce the measurement error, gain more accurate estimates, and assist in better prediction. Generally, the additional data is likely to assist the ML models in converging to more meaningful models, particularly for such small datasets

4.5. Prediction results

Many studies conclude at this point, showing statistics on the features they extracted from the EEG, such as correlations, *t*-tests, and model fitting measures, without preforming out of sample predictions that could shed light on the applicability and generalizability of their findings (for a review, see Hakim & Levy, 2018). However, we aim to expand this and to demonstrate our ability to perform prediction on a held-out data set. For completeness, we provide a summary of all results in appendix J ("Full Prediction Results") and all classification matrices in appendix K ("Classifications Matrices").

Table 2

Repeated Viewings Correlation. We calculated the repeated measures correlation for each EEG measure in each frequency band, between the first and second viewing of the commercials, the first and third, and finally, the second and third viewing, both before and after the transformations outlined in appendix B. Correlations were lower before transformation on the measures, across all features, while correlations were substantially higher after the transformation, showing that the measures' order was reliably replicated between viewings, but not their absolute values.

Correlations on Values Before Transformation									
Feature	Frontal Band Powers			Hemispheric Asymmetry			Inter-Subject Correlation		
Correlation Between Repetitions	First & Second	First & Third	Second & Third	First & Second	First & Third	Second & Third	First & Second	First & Third	Second & Third
Delta	0.14	0.12	0.09	0.14	0.03	−0.01	0.18	−0.11	−0.06
Theta	0.11	−0.03	0.06	−0.05	0.02	−0.02	0.14	0.03	0.07
Alpha	0.27	0.17	0.14	0.05	−0.04	0.09	0.05	0.13	0.19
Beta	0.22	0.18	0.11	0.17	0.11	0.13	0.1	0.15	0.12
Gamma	0.19	−0.06	−0.01	0.15	0.1	0.16	0.06	0.2	0.17
Correlations on Centered Rank-Ordered Values (After Transformation)									
Feature	Frontal Band Powers			Hemispheric Asymmetry			Inter-Subject Correlation		
Correlation Between Repetitions	First & Second	First & Third	Second & Third	First & Second	First & Third	Second & Third	First & Second	First & Third	Second & Third
Delta	0.62	0.53	0.61	0.51	0.41	0.37	0.51	0.33	0.41
Theta	0.56	0.37	0.45	0.41	0.4	0.34	0.47	0.41	0.41
Alpha	0.54	0.58	0.51	0.48	0.44	0.49	0.51	0.51	0.53
Beta	0.57	0.44	0.52	0.47	0.45	0.34	0.5	0.5	0.54
Gamma	0.57	0.47	0.43	0.5	0.41	0.36	0.52	0.62	0.6

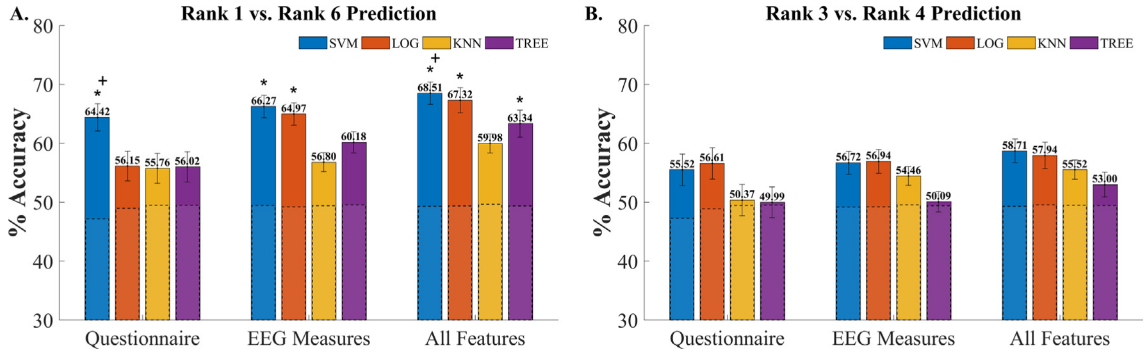

Fig. 5. Binary Prediction Accuracies. Prediction of the highest-ranking product versus the lowest (ranking 1 vs. 6) (A) yielded higher accuracies than when predicting between mid-adjacent products (ranking 3 vs. 4), across all attempted models (B). Accuracies were highest when EEG and Questionnaire features were combined, compared to each of them alone. Significance was obtained from permutation testing as explained in the methods, such that each model was tested for significance against the same model and data trained on shuffled labels (* = $p < 0.05$). Similarly, the best model was tested between "Questionnaire" and "All Features" in Fig. 5A, where the latter was found to have significantly better results (+ = $p < 0.05$). Dashed bars indicate the mean of the shuffled permutations accuracies. Error bars indicate Standard Deviation of the bootstrap distribution (Hesterberg, 2015).

4.5.1. Within-sample prediction – binary classes

We first applied several binary ML models (SVM, LOG, KNN, TREE) on the different combinations of the data types (Questionnaire alone, EEG alone, and combined) in order to predict the product rankings between the most and least preferred products (rank 1 vs. 6). We repeated the analyses for the mid-adjacent rank levels (3 vs. 4). As can be seen in Fig. 5, we successfully predicted between rank 1 and 6 (but not between rank 3 and 4), using the combination of the EEG measures (FBP, ISC, Asymmetry) on their own, reaching up to 66.27% accuracy, while the questionnaire measures on their own reached only up to 64.42% accuracy. Importantly, when we combined both the EEG and Questionnaire measures, we were able to significantly increase our prediction accuracy to 68.51% compared to the accuracy using the Questionnaire alone or the EEG alone (SVM model, permutation test $p < 0.05$, see methods for details). These results demonstrate, as hypothesized, that our neural measures did improve binary preference prediction obtained by the questionnaire measures (by 4.09 percentage points in accuracy), and therefore contained value information beyond what is captured by traditional self-reports alone.

4.5.2. Within-sample prediction – multi-class

We next wanted to examine if we can predict subjects' preference rankings of all 6 products. We used a different set of ML algorithms that can differentiate between several classes (see methods for details). We again tried to predict preferences using the Questionnaire only, the EEG measures only, and the combination of both. We used RMSE as the score for these predictions instead of calculating prediction accuracy. We reached similar results for predicting the entire range of subjects'

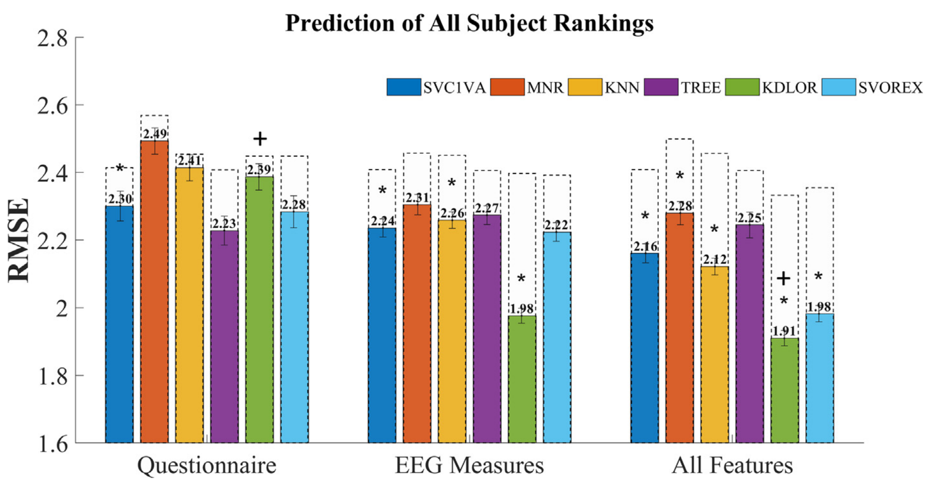

Fig. 6. Subject Rankings Prediction RMSE. The graphs show RMSE scores for prediction of all six rankings when labels are subjects' rankings. Significance was obtained from permutation testing as explained in the methods (* = $p < 0.05$), such that each model was tested for significance against the same model and data trained on shuffled labels. Similarly, the best model was tested between "Questionnaire" and "All Features", where the latter was found to have significantly better results (+ = $p < 0.05$). Dashed bars indicate the mean of shuffled permutations accuracies. Error bars indicate Standard Deviation of the bootstrap distribution (Hesterberg, 2015).

rankings as for the binary predictions. The combination of the EEG and Questionnaire measures yielded the lowest RMSE scores (1.91 for the best model), which were significantly lower than the RMSE scores of a model trained on shuffled labels (2.4 for the best model), and also significantly lower than the same models trained on Questionnaire alone (tested for SVOREX and KDLOR, which reached RMSE of 2.28 and 2.39 respectively, $p < 0.05$, all permutation tested and FDR corrected) (Fig. 6). This clearly shows that EEG measures increased prediction beyond the Questionnaire on its own, by 0.48 RMSE, which is a 20% uplift. Importantly, the lowest RMSE scores were achieved for the ordinal models SVOREX and KDLOR, which consider the ordinal rankings of the products (and not just their nominal labels).

4.5.3. Out-of-sample prediction – online questionnaire

As our next step, we attempted to predict the Out-of-Sample aggregate online questionnaire rankings, based on the EEG and Questionnaire data from our subject pool acquired in the lab. The aggregate online questionnaire rankings served as a proxy for the average preference rankings of the 6 product commercials at the population level. As this is what truly interests marketers and managers, we aimed to show that our EEG measures would contribute to prediction of population-level preferences, beyond traditional self-reports. Importantly, we used the exact same input measures and prediction models as we did for predicting the preferences of our subject pool. Our goal is to demonstrate that the same model that is used in the lab can be used to predict population level metrics. Again, as can be seen in Fig. 7A, we show that all models reached lower RMSE for the combined measures (1.88 RMSE for the best model) compared with those of the Questionnaire alone (2.15 RMSE for the best model), while the best model (KDLOR) was tested and found significant (0.27 improvement in RMSE, which is 12.6% uplift , $p < 0.01$). This demonstrated that the EEG measures increased the prediction accuracy obtained by the Questionnaire alone of out-of-sample online questionnaire rankings.

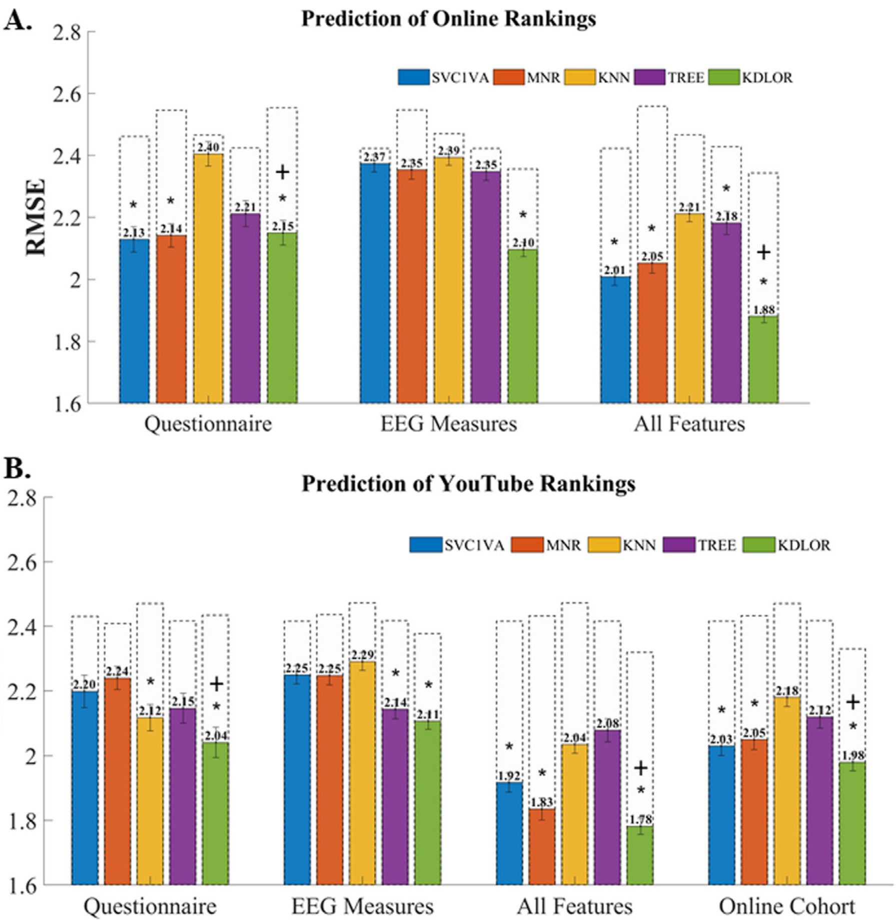

Fig. 7. Out-Of-Sample Rankings Prediction RMSE. The graphs show RMSE scores for prediction of all six rankings when labels are (A) out-of-sample Questionnaire rankings, and (B) YouTube Metrics. Significance was obtained from permutation testing as explained in the methods (* = $p < 0.05$), such that each model was tested for significance against the same model and data trained on shuffled labels. Similarly, the best model was tested between "Questionnaire", "All Features", and "Online cohort", where "All Features" was found to have significantly better results (+ = $p < 0.05$). The "Online Cohort" RMSE scores were obtained by predicting YouTube labels based on questionnaire responses collected online on the commercials and products from 172 respondents. Dashed bars indicate the mean of shuffled permutations accuracies. Error bars indicate Standard Deviation of the bootstrap distribution (Hesterberg, 2015).

4.5.4. Out-of-smple prediction – YouTube metrics

To further strengthen the claim for the potential to predict population-level preferences from EEG measures, we also attempted to predict YouTube rankings of our product commercials, based on the aggregate YouTube metrics. We obtained similar results to those of the online questionnaire. As can be seen in Fig. 7B, the RMSE significantly decreased ($p < 0.001$) for the best model (KDLOR) when we used the combined EEG and Questionnaire measures (1.78 RMSE for the best model), compared to the prediction based on the Questionnaire measures alone (2.04 RMSE for the best model), by 0.26 RMSE, which is a 12.7% uplift.

Moreover, using our in-lab subjects' measures, we were able to reach better prediction of the YouTube rankings than of the Out-of-Sample preference rankings based on the online cohort of subjects (1.78 vs. 1.88 RMSE for the best model, correspondingly, $p < 0.01$). If we assume that our subject pool is a good representation sample of the population, then it is possible that the YouTube metrics are a better proxy for the population preferences than the online questionnaires. This might be because of the aforementioned biases and disadvantages of questionnaires. In contrast, YouTube metrics are more objective and reliable, their sample pool is much larger (~several millions) than can be obtained by online questionnaires, and in general are more ecological than generalizing based on individual preferences as obtained by online questionnaires. This could be of particular interest to marketers, as they often attempt to predict a campaign's success by using self-report subjective online questionnaires.

We also attempted to predict the YouTube rankings based on the questionnaire responses made by the online cohort. The best model reached an impressive 1.98 RMSE, but it was still significantly worse than the results based on the combination of the EEG and Questionnaire responses of our subject pool (1.78 RMSE). This demonstrates that YouTube rankings could be better predicted by collecting both EEG recordings and questionnaire responses from a small in-lab subject pool, rather than collecting a large sample of online responses to a questionnaire only. Importantly, in both the YouTube and Online Questionnaire predictions, we found that the best model was KDLOR. This is probably because this model fits the task best, as it is an ordinal-type model, which considers the rank order of the labels for training and predicting. Lastly, we performed a mediation analysis (see appendix L – Mediation Analysis), which showed that EEG measures significantly contribute to prediction of preferences beyond mediation by Questionnaire features, for all three types of rankings.

4.6. Feature contribution

In addition to our prediction analyses using ML models, we wanted to assess the contribution of each feature to the prediction using a standard regression analysis (see Table 3 – "Regression Analysis"). We found similar results using the regression analysis to the ones we obtained using the ML models. That is, using both questionnaire and EEG measures yielded a significantly improved R^2 compared to a model with questionnaire data alone, for subject rankings (F Change = 5.49, $p < 0.01$), online rankings (F Change = 3.18, $p < 0.05$) and for YouTube rankings (F Change = 7.21, $p < 0.01$). Regarding questionnaire data, all three questionnaire measures were related to the predicted labels – Subject Rankings, Online Rankings, and YouTube Rankings – to varying degrees of significance. Additionally, we found that a model based on EEG data alone significantly predicted the residuals from a model that predicted rankings based on questionnaire data only (Subject Rankings: $F = 7.84$, $p < 0.01$; online rankings: $F = 6.33$, $p < 0.01$; YouTube rankings: $F = 3.72$, $p < 0.05$), which strengthens our conclusion that the EEG measures contributed distinctly from the self-report measures to prediction.

Amongst the EEG measures, we found the following measures to be the most related to our labels. Before we present them, we should clarify that a positive coefficient in the regression table means a negative relationship with value, as we sorted the rankings from highest (1) to lowest (6). As can be seen in Table 3, the frontal power of the alpha band (FBP) is positively related to rankings and therefore negatively to value, in line with a previous study linking alpha suppression to arousal or attention (Foxe & Snyder, 2011). Additionally, the beta band is marginally negatively related to rankings, showing a positive relation to value as was found in a previous study (Boksem & Smidts, 2015).

Hemispheric asymmetry in the beta and gamma bands are positively and negatively related to rankings, respectively. Furthermore, a sensitivity analysis on Hemispheric asymmetry features (Table 2 in appendix I – "Sensitivity Analysis"), showed that the ML models were most sensitive to these features, but less to alpha asymmetry. Most studies on asymmetry in cortical activity discuss solely the alpha band in relation to approach/avoidance tendencies (Davidson, 1998). To our knowledge, previous studies do not provide any neural basis for our findings. However, we decided to examine asymmetry in all frequency bands. This is because we did not want to perform any a-priori feature selection procedure on the frequency bands to avoid overfitting. Therefore, we decided to balance our feature selection procedure between what is known from the literature and a data-driven approach.

Lastly, inter-subject correlations (ISC) in alpha and gamma bands were negatively related to rankings, showing a positive relationship to value, as would be expected from measures linked to engagement (Barnett & Cerf, 2017; Dmochowski et al., 2014). ISC in the delta band had mostly a significant positive coefficient which reflects a negative relation with value. The delta band has been often linked to slow-wave sleep (Bersagliere, Pascual-Marqui, Tarokh, & Achermann, 2018; Finelli, Borbély, & Achermann, 2001). However, it was recently shown that coherent oscillations in the delta frequency band between parietal and frontal cortices are connected to decision-making as well (Nácher, Ledberg, Deco, & Romo, 2013). Yet, we cannot fully explain these results as previous literature on ISC in the delta band is scarce.

Table 3
Regression Analysis Assessing Feature Contribution. We used a mixed effects regression model to predict each label based on either EEG measures, Questionnaire measures, or both. We allowed random intercepts for the different subjects. There was a total of 549 samples in all models. Standard errors appear in parenthesis below each beta value. Significance is indicated by asterisks and by the colour of each cell: dark grey or *** = $p < 0.01$; grey or ** = $p < 0.05$; light grey or * = $p < 0.1$.

Variables \ Label		Subject Rankings			Online Rankings			YouTube Rankings		
		EEG Only	Quest. Only	Both	EEG Only	Quest. Only	Both	EEG Only	Quest. Only	Both
Frontal Band Powers (FBP)	Delta	0.208		0.242	0.121		0.089	0.166		0.210
		(0.267)		(0.258)	(0.274)		(0.261)	(0.271)		(0.243)
	Theta	−0.242		−0.323	−0.228		−0.263	−0.128		−0.0848
		(0.341)		(0.330)	(0.351)		(0.335)	(0.346)		(0.311)
	Alpha	0.602**		0.455*	0.349*		0.316*	0.447**		0.377*
		(0.246)		(0.239)	(0.253)		(0.242)	(0.249)		(0.225)
	Beta	−0.882*		−0.814*	−0.0736		−0.00555	−0.398*		−0.315
		(0.498)		(0.482)	(0.511)		(0.488)	(0.434)		(0.454)
	Gamma	0.100		0.453	0.177		0.0638	0.121		0.245
		(0.423)		(0.413)	(0.435)		(0.419)	(0.429)		(0.389)
Hemispheric Asymmetry	Delta	0.0718		0.0352	0.0445		0.122	0.110		−0.00811
		(0.122)		(0.119)	(0.126)		(0.120)	(0.124)		(0.112)
	Theta	0.0628		0.0587	0.0999		0.131	−0.0325		−0.0126
		(0.167)		(0.161)	(0.171)		(0.163)	(0.169)		(0.152)
	Alpha	−0.095		0.00489	−0.0985		−0.0785	−0.0484		0.0884
		(0.193)		(0.189)	(0.198)		(0.191)	(0.196)		(0.178)
	Beta	0.691***		0.653***	0.405**		0.308**	0.351**		0.321**
		(0.252)		(0.245)	(0.259)		(0.248)	(0.255)		(0.230)
	Gamma	−0.516**		−0.591**	−0.503*		−0.436*	−0.0685		−0.208
		(0.253)		(0.245)	(0.260)		(0.248)	(0.256)		(0.231)
Inter-subject Correlations (ISC)	Delta	1.444***		1.372***	0.665**		0.437**	0.312*		0.347*
		(0.605)		(0.586)	(0.621)		(0.594)	(0.613)		(0.552)
	Theta	−0.466		−0.396	−0.479		−0.366	−1.22**		−0.933*
		(0.528)		(0.511)	(0.542)		(0.518)	(0.535)		(0.481)
	Alpha	−1.112**		−1.075**	−0.642*		−0.520*	−0.827*		−0.412*
		(0.479)		(0.466)	(0.493)		(0.471)	(0.486)		(0.438)
	Beta	0.810		0.820	0.812		0.914	0.489		0.613
		(0.506)		(0.489)	(0.520)		(0.495)	(0.513)		(0.460)
	Gamma	−1.39***		−1.185**	−0.102		−0.325	−0.871*		−0.287
		(0.514)		(0.500)	(0.528)		(0.506)	(0.520)		(0.470)
Questionnaire PCA Components	Comp. 1		−0.97***	−0.92***		−0.242*	−0.238*		−0.73***	−0.68 ***
			(0.160)	(0.161)		(0.159)	(0.163)		(0.151)	(0.151)
	Comp. 2		0.585**	0.617**		1.027**	1.041**		0.603**	0.546**
			(0.288)	(0.287)		(0.287)	(0.290)		(0.270)	(0.270)
	Comp. 3		0.397	0.328		1.980**	1.912**		1.392***	1.296**
			(0.332)	(0.335)		(0.331)	(0.339)		(0.310)	(0.315)
Adjusted R2 (Within)		0.121	0.111	0.153***	0.087	0.112	0.135**	0.137	0.192	0.279***
BIC		2223.858	2140.156	2105.917	2153.590	2137.555	2119.278	2218.481	2157.953	2138.627

Collectively, these results suggest that a combination of engagement (enhanced ISC), possible approach/avoidance tendencies (asymmetry in activations), and preference and purchase intent (alpha and beta band activation), together contributed to the relationship between the EEG signal and predicted labels (see a review for details on all measures: Hakim & Levy, 2018).

To further strengthen these findings, we examined the relationship between each feature and the predicted label via partial dependence plots (PDPs). We used our best model for prediction of all six rankings (KDLOR) to generate the PDPs. These plots allow to inspect each feature's contribution to the predictions when using our ML model. Moreover, PDPs can show that the ML model could have used the features differently compared to how the features are used in the regression analysis, due to the differences between the two modelling approaches (see appendix B for more information). In our case, the PDPs generally outlined similar feature contributions for prediction of subject rankings as the regression model. Yet, they showed that the ML model utilized some of the features differently than the regression for population-level predictions. The PDPs can be found, along with detailed analysis, in appendix M ("Partial Dependence Plots").

Moreover, to better understand our prediction results, we wanted to explore if any of the features were more closely related to specific YouTube Metrics. Therefore, we also performed a regression analysis on each YouTube metric separately (see appendix N, "YouTube Metrics Analysis"). While questionnaire measures explained all YouTube metrics well (aside from comments), EEG measures mostly explained likes and views. Alpha ISC was related to likes and views, while Delta and Theta ISC were related to dislikes. Asymmetry in beta and gamma were related to likes, while beta was also related to views and gamma also related to comments. Importantly, even when regressing each YouTube metric individually, the combination of EEG and questionnaire data provided the best model.

However, although we found relationships between EEG measures and all our prediction labels and also with specific YouTube metrics, we prefer to refrain from making any reverse-inference claims on these results, but only claim that EEG, regardless of the interpretation of its measures, can contribute to preference prediction.

5. Discussion

In the current study, we show that using a combination of various types of features from EEG and advanced ML models increase the predictive power of subjects' future choices obtained using traditional self-reports alone. Moreover, we demonstrate that we can use EEG signals to predict aggregated rank-ordered preferences of an out-of-sample cohort of subjects and the commercials' overall success at the population level using the same EEG measures and prediction algorithms. Importantly, we show that we can utilize a simple, affordable, and hence highly applicable, EEG headset with only eight wet electrodes. This demonstrates the applicability of our approach for the marketing industry. Lastly, we highlight and implement solutions to several important prediction challenges that arise when using a ML approach that, if not accounted for, may inflate prediction accuracies, cause overfitting, and could lead to biased and inaccurate conclusions (see appendix B).

5.1. Preference predictions

For some of our prediction attempts, we were able to show significant prediction accuracies when we used the EEG measures on their own. We showed similar results when using only the features from the questionnaire. However, when we wanted to predict subjects' rank-ordered preferences or population-level preferences, it was the composite of *both* EEG and questionnaire features that had the highest significant predictive accuracy. These findings are strengthened by the mediation analysis in appendix L, and by the regression of the residuals (results Section 4.6 'Feature Contribution'), showing unique explained variance for both the EEG measures and questionnaire for all levels of prediction. These results support our hypothesis that each type of feature captures a different aspect of the valuation process and the combination gives rise to a better prediction. Hence, we emphasize that neuroscientific tools should be combined with self-reports for better prediction.

Our results provide multiple evidence that EEG measures do contribute to prediction in addition to self-reports. First, we conducted a mediation analysis (appendix L), which showed that the two data sources share some information, yet both contribute distinctly to explain rankings, beyond their shared information. Second, our regression results in Table 3 demonstrate that several EEG measures significantly contributed to a model that also included the questionnaire features, and that the combined data yielded a regression model that was significantly better than a model with the questionnaire data alone. Third, we show that a model based on EEG data only could significantly predict the residuals from a model that predicted rankings based on questionnaire data only.

We also showed that the EEG measures alone were better in predicting within-sample rankings, than the questionnaire alone. On the other hand, the questionnaire alone outperformed the EEG measures for population metrics predictions. A possible explanation for this is that EEG measures of different subjects were more indicative of their own personal preferences than the questionnaire they filled out, while the questionnaire on its own was more predictive of a similar online questionnaire and YouTube metrics. This result could suggest that EEG measures contain hidden information regarding the valuation process in individual subjects, which questionnaire responses might not capture. Questionnaires may be more rigid and biased, and contain the disadvantages mentioned in the introduction.

In contrast to findings from previous studies (Venkatraman et al., 2015), we show that EEG measures increased prediction accuracy that was obtained using self-reports elicited by a questionnaire. This was due to three main innovations explored in this paper. First, we used multiple complex EEG measures, which gave prediction models broader information regarding subjects' neural activity and improved their ability to capture complex and varied cognitive aspects of subjects' evaluations of product commercials. This conclusion is strengthened by our analysis demonstrating how different EEG measures are diversely related to different YouTube metrics. Second, we searched beyond standard regression models, and used state-of-the-art ML prediction models. This allowed us to attempt various modelling approaches which could better suit our prediction task. Third, we accounted for possible errors in the analysis process that cleaned our predictions, prevented overfitting, and strengthened their interpretation.

We hoped for higher prediction results than our maximal success rate of 68.5%. However, this success rate is well within the higher range of other similar out-of-sample prediction attempts (Hakim & Levy, 2018). However, based on a cost analysis (see below), even a moderate increase in prediction rates could have a huge impact on the bottom line. A gain of 12–13% in error reduction when predicting the success of campaigns, for only a small fraction of the marketing budget, can be highly beneficial. Moreover, we think that this limitation is an important point for managers and all practitioners to know about. That is, neuromarketing tools could be beneficial but they are limited. This by itself could reduce the overclaims frequently found in the industry regarding the ability to "read our thoughts". There are several reasons for our (and others) moderate prediction rates. First, EEG signals are complex and notoriously noisy. Second, manually searching which features are the most predictive are akin to searching for a needle in a haystack. There are endless possibilities both in the frequency and in the time domains. Hence, due to the moderate results shown in the current study and in other studies (see Hakim & Levy, 2018 for a review), we surmise that manual feature extraction from EEG signals is inherently limited in its prediction power. Therefore, in future studies, we plan to move to an automatic feature extraction approach and explore whether we could break the current ceiling in the field and reach higher prediction rates.

It should be noted that our EEG measures included 15 features, and we did not perform any feature-selection procedure to select specific features but used all measures for all EEG-based predictions. In contrast, we did select only the three strongest features from the questionnaire, based on an unsupervised procedure that did not involve data leakage (PCA). Regardless, we report results on test sets that were not involved in any feature processing, thus equalizing the playing field for both types of data. Moreover, in order to avoid possible critiques on unfairly lowering prediction results from the questionnaire, we chose the analysis procedure which yielded the best results, though it is possible that further exploration and optimization of the questionnaire responses would improve upon these results.

Yet, this is a main point regarding our critique of questionnaires. One could always claim that the questionnaire used was not good enough, and this caused the EEG measures to increase prediction accuracy. Theoretically, it is possible that continually refining questionnaires could result in an ultimate version that overcomes all their limitations and unearths the precise, unbiased, and genuine preferences of individuals. However, we believe this to be true only in theory, as the ‘‘perfect" questionnaire is inherently unattainable. Ideally, there could exist a perfect questionnaire which captures individual and population preferences precisely in a particular marketing setting, but it may take an immaterial amount of time and effort calibrating such questions for each setting, while the literature is already beginning to converge on powerful, general, and objective neural predictors of preference.

5.2. Modelling challenges

Applying ML approaches on EEG data can lead to several challenges, which we aimed to identify and address. EEG activity can differ widely between subjects and trials, requiring appropriate normalization within and between subjects for these models to be used correctly. The multi-leveled nature of experimental datasets also demanded careful handling to avoid leakage, as did any transformations on the data which could be considered peaking and possibly yield falsely inflated prediction rates (see appendix B for details).

Importantly, our modeling approach allowed us to avoid the problems of multiple comparisons, by relying on previous literature for feature creation while utilizing all measures for all frequency bands in our predictions. Moreover, we examined ecological and market-relevant stimuli, without the need to subjectively judge commercials or specific scenes within them. This was done by averaging band powers during the entire commercial (FBP, Asymmetry), or preforming analysis that is based on the entire commercial timeline (ISC).

In addition, our results show that in binary predictions, the SVM model performed best, likely due to its greater flexibility and lack of assumptions. For prediction of ordinal rankings (that is, for predicting the entire range of rankings 1–6), whether they be on the individual or population level, using ML models that are geared for ordinal prediction, SVOREX and KDLOR, were mostly better than nominal models, but it is possible that other ordinal models may have also worked well. The models TREE, KNN, MNR and SVC1VA are all nominal models, while KNN is a simplistic model, and MNR assumes linearity, and therefore, did not produce the best results for ordinal rankings' prediction. SVOREX and KDLOR have an additional advantage, they both employ a kernel trick to incorporate non-linearity, which is recommended for most problems where the linearity of mappings is unknown. However, the models differ in some key aspects. Models based on discriminant analysis, like KDLOR, are generative models, they assume that the data is normally distributed, and focus on all datapoints for prediction making them more prone to outliers. These models are superior when their assumptions are met. In contrast, SVM family of models are discriminative models, they make less assumptions, are optimized over a subset of the data that lies on the separating margin (support vectors) and are generally very flexible. In conclusion, The best model for preference prediction will likely be the model that most accurately considers the dependent variable (ordinal versus nominal or continuous), and that has its assumption met by the data (or otherwise lacks assumptions).

Another important conclusion is that we must examine the between subject variation in preferences and how they might influence meaningful predictions. For within-sample prediction, if subjects were highly correlated with one another, such that most like and dislike the same products, then predictions become closer to predicting the product commercial identity rather than subject-specific ranking of that commercial. That is, the EEG measures could be reflective of the commercial's objective characteristic, such as its length, saliency, brightness, motility and others, rather than individual preferences, and there would be no way to tell apart the two possible interpretations of the EEG measures' predictability. Importantly, in the current study, we made sure that the preference correlation across subjects was relatively low.

For out-of-sample predictions it is even harder to tell these interpretations apart. First, it is easier to predict an overall population product ranking than subject specific choice. When attempting to predict subject-specific preferences, the rank of a product can change for every subject. On the other hand, when attempting to predict population-level metrics the rank of a product remains the same across all subjects – the averaged rank allotted to the product by the population. Then, if subjects' rankings correlate well with the population rankings, i.e. everyone likes and dislikes the same products, the EEG measures only need to capture a pattern of response to the objective attributes of the product commercial, rather than any value-related signals, in order to predict population rankings successfully.

This interpretation is particularly strong for small sets of products, but its relevance decreases the more products are included. We chose to use a small set of stimuli with multiple repetitions for two main reasons. First, to improve the signal to noise ratio to increase our prediction power. Second, to have a research design, which is close to what is commonly done in a marketing research in the industry. However, we acknowledge that this limits the generalizability of our results for real world applications. Another limitation in the current study is the size of the dataset. Although many neuromarketing studies include pools of around 30 subjects (Hakim & Levy, 2018), which is similar to our sample size, ML models benefit from larger datasets. Therefore, our findings are limited in their robustness and generalizability due to the relatively small sample size.

Therefore, future studies should focus on strengthening supporting evidence for powerful neural predictors using larger quantities of stimuli and more subjects. Once neural predictors of value are adequately established, then they can be utilized on real-world problems that interest marketers and managers, which usually involve only a small set of closely similar stimuli.

This issue presents a limitation on predicting population-level preferences in general based on a small sampled group, and it is not an issue particular to EEG. Our study showed the contribution that EEG has for out-of-sample prediction, that could be replicated for any marketing decision – be it choosing from different versions of a certain commercial, differing brands of a product type, between varying product categories, or otherwise. However, the application to other marketing or managerial decisions could be limited, because we achieved less accuracy when predicting between closely related preferences (rank 3 vs. rank 4) in comparison to distant preferences (rank 1 vs. rank 6). The demand in the industry is to predict between closely related stimuli, such as several versions of the same commercial that only differ in a few parameters. Therefore, we conclude that this remains a challenge for future studies, as other neuromarketing studies have yet to demonstrate an ability to predict between very similar stimuli.

Moreover, the research community could benefit mainly from our results regarding within-sample predictions. These showed that substantial success can be achieved by using multiple types of measures of EEG recordings that are based on previous literature. This suggests that the neural correlates of value, as measured by EEG, are probably a combination of various measures, which reflect different cognitive aspects of the valuation process, and are not restricted to a specific unique signal, such as the BOLD signal measured in fMRI (Bartra et al., 2013; Levy & Glimcher, 2012). Researchers should also be incentivized to discover novel ways to process the EEG signal and develop new measures to capture aspects of value information independent of existing measures, to further enhance EEG-based prediction.

5.3. Interpretability and the reverse inference problem

We would like to emphasize that although the regression analysis we conducted provides insight on how our approach could improve predictions, we think that understanding which EEG features were predictive is less important to managers than simply gaining predictive power. While the former may interest academics, who investigate brain processes underlying preferences, we also aimed this study to industry practitioners. Thus, our main goal was to demonstrate the ability to predict preferences, rather than explain which are the best features of the EEG signal for prediction.

Many researchers claim that if they find a frequency band that is predictive of preferences, then that would tell us something about the cognitive processes involved with preference creation. However, we believe this logic to be problematic due to the issue of "reverse inference" (Poldrack, 2006). For example, say that the Alpha band power was shown to be correlated with attention processes in previous studies. This does not mean that when we find that the Alpha band is a good predictor of preferences it is because attention was involved. Hence, we do not wish to make strong cognitive claims regarding the predictive contribution of our features in order to avoid the reverse inference problem.

To our understanding, managers desire explainable predictions in specific cases, when the research question is whether a given stimulus would be more memorable, or a specific part of the commercial more engaging, etc. However, there are often cases where managers wish to test and choose which option is more preferred by consumers, as the options are already after production and no more changes can be made. It is in these cases where our research could be most useful, although, we do agree that providing explanations is important during pre-production and stimuli creation.

5.4. Cost-benefit analysis

In order to entice the industry to adopt our suggested methodology, we provide an analysis into the costs and potential gains of using EEG alongside questionnaires. The cost of an EEG device ranges from several thousands of dollars for simpler and less reliable devices, through $10 K for the device we used in the current study, and up to $50 K for the most sophisticated and electrode-dense devices. While the device can be used multiple times for many years, some consumables need to be replenished (electrodes, gel for wet electrodes, etc.), for a cost of about $5 per subject. Other costs include participants fee, which in our design was $15 per subject for one hour. This brings us to a total cost of $20 per hour of recording and questionnaire response, excluding the initial cost of acquiring the headset. Of course, this does not take into account labor cost of the experimenter, which can be on the order of a few thousands of dollars per experiment (for our example let's assume a labor cost of $5000). The initial cost can be cut if one hires a neuromarketing company that already owns the necessary materials and devices, but probably the cost per subject would be higher to account for their general overhead and profits.

Thus, an experiment similar to our design with 30 subjects, would cost approximately $5600 (excluding initial cost) to gain both questionnaire and EEG data. Based on our results, from this combined data, one can gain a 4.09 percentage points increase in prediction accuracy between the highest and lowest preferences (64.42% accuracy without EEG measures to 68.51% with them). This also reflects a decrease of 12.7% in prediction error of the YouTube popularity (a decrease of 0.26 in RMSE acquired for the questionnaire-based prediction only [2.04] to prediction based on all features [1.78]), 12.6% decrease for Online responses, and 20.1% decrease for prediction error of within-subject preferences.

Using only online questionnaires (on mTurk, for example) can be a cheap alternative, because costs are around $200 per experiment for approximately 100 responses (depending on the length of the questionnaire, number of tested stimuli, targeting, etc.). However, by adding the EEG data, marketers and managers can reduce the prediction error from commercial responses to population preferences by 12–13%, at the additional cost of only $5600 per experiment, an additional $5400 compared to the mTurk option. Hence, if we assume a marketing budget of $1M, where the difference in cost between the two methods is only 0.54% of the budget, yet it can improve estimates of success by 12–13% at the population level, and up to 20% at the individual level, we can conclude that adding EEG measures can be both beneficial and worthwhile.

5.5. Conclusions

Although considerable knowledge regarding consumer preferences can be attained using traditional marketing tools, such as questionnaires and focus groups, many marketing strategies are still unsuccessful in aiding managers and marketers to make business and marketing decisions (Hamel & Prahalad, 1994; Martin, 1995; Ovans, 1998). The relevant neuroscientific literature has found substantial evidence that links the EEG measures we used to consumer preferences. Moreover, we apply novel data science methodology and techniques that align with industry standard prediction modeling. Hence, our research expands the manager's toolbox, with a cost-effective tool, EEG, with which they can access customers' cortical activity to execute managerial decisions empirically, possibly better than what could be accomplished with traditional self-reports. Practitioners can utilize our findings and algorithm to predict responses to new company products, to modifications in existing products, to specific tactics or to various marketing campaign options and branding decisions, before substantial investment in media spending. By achieving accurate marketing performance predictions, as our population prediction results imply, managers could drastically decrease failures or uncertainties in their strategy, increase their marketing effectiveness, broaden their audience, improve brand image, and maximize their return on investment.

Acknowledgements

We want to thank all Dino Levy's lab members for their help and guidance in many discussions. We thank Ma'ayan Hakim for translations of the marketing questionnaire. We also thank Yoav Zeevi, a Ph.D. student of Professor Yoav Benjamini, for his guidance with statistical tests performed throughout the paper. DJL thanks the Israel Science Foundation (grant #954/17) and The Henry Crown Institute of Business Research in Israel for their support of this study.

Author contributions statement

AH performed the data analysis, created and optimized the prediction models and modelling challenges. DJL designed the experiment. TS and MSS conducted the experimental procedure. SK conducted the experiment, performed EEG preprocessing and manual artifact rejection. DF provided valuable consultation and guidance. AH and DL wrote the manuscript. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare no competing financial interests.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijresmar.2020.10.005.

References

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology, 8*, 456. https://doi.org/10.3389/fpsyg.2017.00456.

Barnett, S. B., & Cerf, M. (2017). A ticket for your thoughts: Method for predicting content recall and sales using neural similarity of moviegoers. *Journal of Consumer Research, 44*(1), 160–181. https://doi.org/10.1093/jcr/ucw083.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage, 76*, 412–427. https://doi.org/10.1016/J.NEUROIMAGE.2013.02.063.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological), 57*, 289–300. https://doi.org/10.2307/2346101.

Bersagliere, A., Pascual-Marqui, R. D., Tarokh, L., & Achermann, P. (2018). Mapping slow waves by EEG topography and source localization: Effects of sleep deprivation. *Brain Topography, 31*(2), 257–269. https://doi.org/10.1007/s10548-017-0595-6.

Bhavsar, R., Sun, Y., Helian, N., Davey, N., Mayor, D., & Steffert, T. (2018). The correlation between EEG signals as measured in different positions on scalp varying with distance. *Procedia Computer Science, 123*, 92–97. https://doi.org/10.1016/J.PROCS.2018.01.015.

Birmingham, P. & Wilkinson, D. (2003). Using research instruments: A guide for researchers. Retrieved from https://content.taylorfrancis.com/books/download?dac=C2012-0-02781-8&isbn=9781134483464&format=googlePreviewPdf.
Boksem, M. A. S., & Smidts, A. (2015). Brain responses to movie trailers predict individual preferences for movies and their population-wide commercial success. *Journal of Marketing Research, 52*(4), 482–492. https://doi.org/10.1509/jmr.13.0572.
Braeutigam, S., Rose, S. P. R., Swithenby, S. J., & Ambler, T. (2004). The distributed neuronal systems supporting choice-making in real-life situations: Differences between men and women when choosing groceries detected using magnetoencephalography. *European Journal of Neuroscience, 20*(1), 293–302. https://doi.org/10.1111/j.1460-9568.2004.03467.x.
Buchanan, B., & Henderson, P. W. (1992). Assessing the bias of preference, detection, and identification measures of discrimination ability in product design. *Marketing Science, 11*(1), 64–75. https://doi.org/10.1287/mksc.11.1.64.
Cartocci, G., Cherubino, P., Rossi, D., Modica, E., Maglione, A. G., Di Flumeri, G., & Babiloni, F. (2016). Gender and age related effects while watching TV advertisements: An EEG study. *Computational Intelligence and Neuroscience, 2016*, 1–11. https://doi.org/10.1155/2016/3795325.
Chan, H.-Y., Smidts, A., Schoots, V. C., Dietvorst, R. C., & Boksem, M. A. S. (2019). Neural similarity at temporal lobe and cerebellum predicts out-of-sample preference and recall for video stimuli. *NeuroImage, 197*, 391–401. https://doi.org/10.1016/J.NEUROIMAGE.2019.04.076.
Christoforou, C., Papadopoulos, T. C., Constantinidou, F., & Theodorou, M. (2017). Your brain on the movies: A computational approach for predicting box-office performance from viewer's brain responses to movie trailers. *Frontiers in Neuroinformatics, 11*, 72. https://doi.org/10.3389/fninf.2017.00072.
Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods, 250*, 126–136. https://doi.org/10.1016/j.jneumeth.2015.01.010.
Davidson, R. J. (1998). Anterior electrophysiological asymmetries, emotion, and depression: Conceptual and methodological conundrums. *Psychophysiology, 35*(5), 607–614. https://doi.org/10.1017/S0048577298000134.
Day, G. S. (1975). The Threats to marketing research. *Journal of Marketing Research, 12*(4), 462. https://doi.org/10.2307/3151099.
Dmochowski, J. P., Bezdek, M. A., Abelson, B. P., Johnson, J. S., Schumacher, E. H., & Parra, L. C. (2014). Audience preferences are predicted by temporal reliability of neural processing. *Nature Communications, 5*, 1–9. https://doi.org/10.1038/ncomms5567.
Dmochowski, J. P., Sajda, P., Dias, J., & Parra, L. C. (2012). Correlated components of ongoing EEG point to emotionally laden attention – a possible marker of engagement?. *Frontiers in Human Neuroscience, 6*(May), 1–9. https://doi.org/10.3389/fnhum.2012.00112.
Finelli, L. A., Borbély, A. A., & Achermann, P. (2001). Functional topography of the human nonREM sleep electroencephalogram. *European Journal of Neuroscience, 13*(12), 2282–2290. https://doi.org/10.1046/j.0953-816X.2001.01597.x.
Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research, 20*(2), 303. https://doi.org/10.1086/209351.
Foley, D. (1972). Considerations of sample and feature size, in. *IEEE Transactions on Information Theory, 18*(5), 618–626. https://doi.org/10.1109/TIT.1972.1054863.
Foxe, J. J., & Snyder, A. C. (2011). The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. *Frontiers in Psychology, 2*(JUL), 154. https://doi.org/10.3389/fpsyg.2011.00154.
Fuentemilla, L., Cucurell, D., Marco-Pallarés, J., Guitart-Masip, M., Morís, J., & Rodríguez-Fornells, A. (2013). Electrophysiological correlates of anticipating improbable but desired events. *NeuroImage, 78*, 135–144. https://doi.org/10.1016/j.neuroimage.2013.03.062.
Genevsky, A., & Knutson, B. (2018). Neuroforecasting aggregate choice. *Current Directions in Psychological Science, 27*(2), 110–115.
Gibbs, A. (1997). Focus Groups. Retrieved April 27, 2018, from Social Research Update website: http://sru.soc.surrey.ac.uk/SRU19.html.
Golland, P. & Fischl, B. (2003). Permutation Tests for Classification: Towards Statistical Significance in Image-Based Studies. https://doi.org/10.1007/978-3-540-45087-0_28.
Good, P. (2013). Permutation tests: A practical guide to resampling methods for testing hypotheses Retrieved from https://www.google.com/books?hl=en&lr=&id=pK3hBwAAQBAJ&oi=fnd&pg=PR5&dq=good+2000+permutation+testing&ots=qWfHdGEFuz&sig=TUjItFtefE5gak0_dbegyPKS72c.
Griffin, A., & Hauser, J. R. (1993). The voice of the customer. *Marketing Science, 12*(1), 1–27. https://doi.org/10.1287/mksc.12.1.1.
Guixeres, J., Bigné, E., Azofra, J. M. A., Raya, M. A., Granero, A. C., Hurtado, F. F., & Ornedo, V. N. (2017). Consumer neuroscience-based metrics predict recall, liking and viewing rates in online advertising. *Frontiers in Psychology, 8*(OCT), 1–14. https://doi.org/10.3389/fpsyg.2017.01808.
Gutierrez, P. A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., & Hervas-Martinez, C. (2016). Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering, 28*(1), 127–146. https://doi.org/10.1109/TKDE.2015.2457911.
Hakim, A. & Levy, D. J. (2018). A gateway to consumers' minds : Achievements, caveats, and prospects of electroencephalography-based prediction in neuromarketing. (May), 1–21. https://doi.org/10.1002/wcs.1485.
Hamel, G., & Prahalad, C. K. (1994). Competing for the future. Retrieved from https://books.google.co.il/books/about/Competing_for_the_Future.html?id=PlML2w8dhJ0C&redir_esc=y.
Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Natural vision. *Science, 303*(MARCH), 1634–1640. https://doi.org/10.1126/science.1089506.
Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician, 69*(4), 371–386. https://doi.org/10.1080/00031305.2015.1089789.
Hsu, M., & Yoon, C. (2015). The neuroscience of consumer choice. *Current Opinion in Behavioral Sciences, 5*, 116–121. https://doi.org/10.1016/J.COBEHA.2015.09.005.
Hua, J., Xiong, Z., Lowey, J., Suh, E., & Dougherty, E. R. (2005). Optimal number of features as a function of sample size for various classification rules. *Bioinformatics, 21*(8), 1509–1515 https://doi.org/10.1093/bioinformatics/bti171.
Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition, 15*(4), 673–692. https://doi.org/10.1016/J.CONCOG.2006.09.004.
Khushaba, R. N., Greenacre, L., Kodagoda, S., Louviere, J., Burke, S., & Dissanayake, G. (2012). Choice modeling and the brain: A study on the Electroencephalogram (EEG) of preferences. *Expert Systems with Applications, 39*(16), 12378–12388. https://doi.org/10.1016/j.eswa.2012.04.084.
Khushaba, R. N., Wise, C., Kodagoda, S., Louviere, J., Kahn, B. E., & Townsend, C. (2013). Consumer neuroscience: Assessing the brain response to marketing stimuli using electroencephalogram (EEG) and eye tracking. *Expert Systems with Applications, 40*(9), 3803–3812. https://doi.org/10.1016/j.eswa.2012.12.095.
Koelstra, S., Mühl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., ... Patras, I. (2012). DEAP: A database for emotion analysis; Using physiological signals. *IEEE Transactions on Affective Computing, 3*(1), 18–31. https://doi.org/10.1109/T-AFFC.2011.15.
Kong, W., Zhao, X., Hu, S., Vecchiato, G., & Babiloni, F. (2013). Electronic evaluation for video commercials by impression index. *Cognitive Neurodynamics, 7*(6), 531–535. https://doi.org/10.1007/s11571-013-9255-z.
Lau, T. M., Gwin, J. T., & Ferris, D. P. (2012). How many electrodes are really needed for EEG-based mobile brain imaging?. *Journal of Behavioral and Brain Science, 2*, 387–393. https://doi.org/10.4236/jbbs.2012.23044.
Laurence, J. A. & Gerhold, M. M. (2016). Saving the day : The relationship between emotion and purchase intent in television advertising. 1–12.
Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology, 22*(6), 1027–1038. https://doi.org/10.1016/J.CONB.2012.06.001.
Levy, I., Lazzaro, S. C., Rutledge, R. B., & Glimcher, P. W. (2011). Choice from non-choice: Predicting consumer preferences from blood oxygenation level-dependent signals obtained during passive viewing. *Journal of Neuroscience, 31*(1), 118–125. https://doi.org/10.1523/JNEUROSCI.3214-10.2011.
Luck, S. J. (2014). An introduction to the event-related potential technique (Second edi). Retrieved from http://www.worldcat.org/title/introduction-to-the-event-related-potential-technique/oclc/861671073.
MacKenzie, S. B., & Podsakoff, P. M. (2012). Common method bias in marketing: Causes, mechanisms, and procedural remedies. *Journal of Retailing, 88*(4), 542–555. https://doi.org/10.1016/J.JRETAI.2012.08.001.
Martin, J. (1995). Ignore Your Customer Retrieved from. *Managing: Ideas & Solutions, 131*(8), 121.

McDaniel, S. W., & Kolari, J. W. (1987). Marketing strategy implications of the miles and snow strategic typology. *Journal of Marketing, 51*(4), 19. https://doi.org/10.2307/1251245.

McDaniel, S. W., Verille, P., & Madden, C. S. (1985). The threats to marketing research: An empirical reappraisal. *Journal of Marketing Research, 22*(1), 74. https://doi.org/10.2307/3151552.

Musch, J., & Reips, U.-D. (2000). A brief history of web experimenting. *Psychological Experiments on the Internet, 61–87*. https://doi.org/10.1016/B978-012099980-4/50004-6.

Nácher, V., Ledberg, A., Deco, G., & Romo, R. (2013). Coherent delta-band oscillations between cortical areas correlate with decision making. *Proceedings of the National Academy of Sciences of the United States of America, 110*(37), 15085–15090. https://doi.org/10.1073/pnas.1314681110.

Neeley, S. M. & Cronley, M. L. (2004). When research participants Don'T tell it like it is: pinpointing the effects of social desirability bias using self Vs. indirect-questioning. ACR North American Advances, NA-31. Retrieved from http://www.acrwebsite.org/volumes/8930/volumes/v31/NA-31.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping, 15*(1), 1–25. https://doi.org/10.1002/hbm.1058.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231.

Ohme, R., Reykowska, D., Wiener, D., & Choromanska, A. (2009). Analysis of neurophysiological reactions to advertising stimuli by means of EEG and galvanic skin response measures. *Journal of Neuroscience, Psychology, and Economics, 2*(1), 21–31. https://doi.org/10.1037/a0015462.

Ohme, R., Reykowska, D., Wiener, D., & Choromanska, A. (2010). Application of frontal EEG asymmetry to advertising research. *Journal of Economic Psychology, 31*(5), 785–793. https://doi.org/10.1016/j.joep.2010.03.008.

Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance Retrieved from. *Journal of Machine Learning Research, 11*(Jun), 1833–1863.

Ovans, A. (1998). The customer doesn't always know best. *Market Research, 7*(3), 12–14.

Plassmann, H., Venkatraman, V., Huettel, S., & Yoon, C. (2015). Consumer neuroscience: Applications, challenges, and possible solutions. *Journal of Marketing Research, 52*(4), 427–435. https://doi.org/10.1509/jmr.14.0048.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data?. *Trends in Cognitive Sciences, 10*(2), 59–63. https://doi.org/10.1016/j.tics.2005.12.004.

Ramsøy, T. Z., Skov, M., Christensen, M. K., & Stahlhut, C. (2018). Frontal brain asymmetry and willingness to pay. *Frontiers in Neuroscience, 12*(MAR). https://doi.org/10.3389/fnins.2018.00138.

Ravaja, N., Somervuori, O., & Salminen, M. (2013). Predicting purchase decision: The role of hemispheric asymmetry over the frontal cortex. *Journal of Neuroscience, Psychology, and Economics, 6*(1), 1–13. https://doi.org/10.1037/a0029949.

San Martin, R., Appelbaum, L. G., Pearson, J. M., Huettel, S. A., & Woldorff, M. G. (2013). Rapid brain responses independently predict gain maximization and loss minimization during economic decision making. *Journal of Neuroscience, 33*(16), 7011–7019. https://doi.org/10.1523/JNEUROSCI.4242-12.2013.

Shestyuk, A. Y., Kasinathan, K., Karapoondinott, V., Knight, R. T., & Gurumoorthy, R. (2019). Individual EEG measures of attention, memory, and motivation predict population level TV viewership and Twitter engagement. *PLoS ONE, 14*(3), 1–27. https://doi.org/10.1371/journal.pone.0214507.

Smidts, A., Hsu, M., Sanfey, A. G., Boksem, M. A. S., Ebstein, R. B., Huettel, S. A., ... Yoon, C. (2014). Advancing consumer neuroscience. *Marketing Letters*. https://doi.org/10.1007/s11002-014-9306-1.

Smith, M. E., & Gevins, A. (2004). Attention and brain activity while watching television: Components of viewer engagement. *Media Psychology, 6*(3), 285–305. https://doi.org/10.1207/s1532785xmep0603_3.

Smithson, J. (2000). Using and analysing focus groups: Limitations and possibilities. *International Journal of Social Research Methodology, 3*(2), 103–119. https://doi.org/10.1080/136455700405172.

Song, J., Davey, C., Poulsen, C., Luu, P., Turovets, S., Anderson, E., ... Tucker, D. (2015). EEG source localization: Sensor density and head surface coverage. *Journal of Neuroscience Methods, 256*, 9–21. https://doi.org/10.1016/J.JNEUMETH.2015.08.015.

Sutton, S. K., & Davidson, R. J. (2000). Prefrontal brain electrical asymmetry predicts the evaluation of affective stimuli. *Neuropsychologia, 38*(13), 1723–1733. https://doi.org/10.1016/S0028-3932(00)00076-2.

Telpaz, A., Webb, R., & Levy, D. J. (2015). Using EEG to predict consumers' future choices. *Journal of Marketing Research, 52*(4), 511–529. https://doi.org/10.1509/jmr.13.0564.

Vecchiato, G., Astolfi, L., Fallani, F. D. V., Cincotti, F., Mattia, D., Salinari, S., ... Babiloni, F. (2010). Changes in brain activity during the observation of TV commercials by using EEG, GSR and HR measurements. *Brain Topography*. https://doi.org/10.1007/s10548-009-0127-0.

Vecchiato, G., Toppi, J., Astolfi, L., De Vico Fallani, F., Cincotti, F., Mattia, D., ... Babiloni, F. (2011). Spectral EEG frontal asymmetries correlate with the experienced pleasantness of TV commercial advertisements. *Medical and Biological Engineering and Computing, 49*(5), 579–583. https://doi.org/10.1007/s11517-011-0747-x.

Venkatraman, V., Dimoka, A., Pavlou, P. A., Vo, K., Hampton, W., Bollinger, B., ... Winer, R. S. (2015). Predicting advertising success beyond traditional measures: New insights from neurophysiological methods and market response modeling. *Journal of Marketing Research, 52*(4), 436–452. https://doi.org/10.1509/jmr.13.0593.

Wei, Z., Wu, C., Wang, X., Supratak, A., Wang, P., & Guo, Y. (2018). Using support vector machine on EEG for advertisement impact assessment. *Frontiers in Neuroscience*, *12*(MAR). https://doi.org/10.3389/fnins.2018.00076.

Yadava, M., Kumar, P., Saini, R., Roy, P. P., & Prosad Dogra, D. (2017). Analysis of EEG signals and its application to neuromarketing. *Multimedia Tools and Applications, 76*(18), 19087–19111. https://doi.org/10.1007/s11042-017-4580-6.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393.