ORGANIZING THEIR THOUGHTS – HOW ONLINE REVIEW TEMPLATES AFFECT THE REVIEW TEXT

Martin Poniatowski
Paderborn University, martin.poniatowski@upb.de

Jürgen Neumann
Paderborn University, juergen.neumann@wiwi.uni-paderborn.de

Thomas Goerzen
Paderborn University, thomas.goerzen@upb.de

Dennis Kundisch
Paderborn University, dennis.kundisch@wiwi.uni-paderborn.de

Follow this and additional works at: https://aisel.aisnet.org/ecis2019_rp

Recommended Citation

This material is brought to you by the ECIS 2019 Proceedings at AIS Electronic Library (AISel). It has been accepted for inclusion in Research Papers by an authorized administrator of AIS Electronic Library (AISel). For more information, please contact elibrary@aisnet.org.
Organizing Their Thoughts

ORGANIZING THEIR THOUGHTS – HOW ONLINE REVIEW TEMPLATES AFFECT THE REVIEW TEXT

Research paper

Poniatowski, Martin, Paderborn University, Paderborn, Germany, martin.poniatowski@wiwi.uni-paderborn.de

Neumann, Jürgen¹, Paderborn University, Paderborn, Germany, juergen.neumann@wiwi.uni-paderborn.de

Görzen, Thomas¹, Paderborn University, Paderborn, Germany, thomas.goerzen@wiwi.uni-paderborn.de

Kundisch, Dennis¹, Paderborn University, Paderborn, Germany, dennis.kundisch@wiwi.uni-paderborn.de

Abstract

The importance of online reviews for customer decision making and their impact on sales and market efficiency is well established. One key issue for review platforms, nevertheless, is to design review systems which present helpful reviews to customers – and therefore guide reviewers towards writing reviews with high informational value but minimal cognitive effort required. In this study, we investigate whether review templates that suggest product or service features (e.g., camera usability, service in a restaurant) for the review text can successfully address this dual challenge. After having generated a review template for restaurants based on a data-driven approach, we conduct an online experiment to evaluate the influence of our template on review writing and content. In line with our hypotheses, we find that review templates suggesting product or service features can foster the creation of longer reviews covering more features. Previous literature has shown that these review properties are perceived as particularly helpful. We find that, despite writing longer reviews, the reviewers do not perceive the use of a template as requiring additional effort. Our results carry important practical implications for the design of online review systems and contribute to the literature on human interaction with review systems.

Keywords: Online Reviews, Review Template, Writing Process, Experiment, Topic Modelling.

1 Introduction

Online review systems are an important component of digital markets as they help reduce the information asymmetry between customers and sellers (Wu et al., 2015; Dellarocas, 2003). Customers have increasingly come to rely on the information they obtain from reading online reviews to make their purchase decisions (Wu et al., 2015). To achieve a high number of users, online shops and platforms offering online review systems need to ensure that past customers publish reviews with information that other customers find helpful. Online review systems can implement measures that guide reviewers towards publishing not only more reviews but also more helpful ones. To increase the number of published

¹ This work has been partially supported by the German Research Foundation (DFG) within the Collaborative Research Center 901 “On-The-Fly Computing” under the project number 160364472-SFB901.
reviews, Amazon, for example, introduced a reviewer ranking system. Research has found that the introduction of this ranking system has led reviewers to publish more reviews for products which receive either many reviews (popular products) or few reviews (niche products) (Shen et al., 2015). Review system designers can also implement measures aimed at influencing review content, typically to increase their informational value. Research found that presenting previous customers with information on how their reviewing behavior compares to that of their peers can affect review content in terms of text length (Burtch et al., 2018). Providing reviewers with templates may be another way to influence review content. Such review templates could present a structure or suggest ideas to reviewers to make review writing easier whilst supporting review quality. Many studies propose review templates as a design feature of online review systems (Kwark et al., 2014; Yin et al., 2014; Kuan et al., 2015) and review templates are already used in practice. For instance, TripAdvisor asks reviewers how they liked the room, location and services when they are about to review a hotel and Amazon asks their customers for what they have liked or disliked about their purchase (e.g., a digital camera). However, it is unclear whether such templates do influence reviewers, and if so, how. This study aims to close this gap by studying the influence of review templates on review content. Consequently, we try to answer the following research question:

*How does presenting reviewers with a review template influence the resulting review text?*

Review templates can take on different design features. For example, many popular platforms have review templates which suggest relevant product or service features to the reviewer (i.e., a feature-based review template) or ask them to comment on pros and cons. As literature has found that review texts discussing product features are perceived as more helpful (Scholz and Dorner, 2013), our study focuses on analyzing whether suggesting features in form of a feature-based review templates can lead to more helpful reviews.

To answer our research question, we design our hypotheses drawing on the Adaptive Character of Thought theory (Anderson, 1996) and the Cognitive Process Theory of Writing (Flower and Hayes, 1981). We then employ a data-driven approach to develop a feature-based review template for restaurants, and test our hypotheses using the resulting template in an online experiment. We hypothesize that using a template increases the number of product or service features discussed in the review text, and the text length, and that it reduces the perceived effort of writing. Conducting an online experiment on a crowdfunding platform, we find that participants provided with the template write significantly longer reviews covering more features. We do not find any difference in perceived effort between treatment and control group which suggests that such a template does not increase perceived effort.

Our research contributes to the existing literature on the relationship between the design of online reviews systems and review texts by examining how review templates affect reviewing behavior and consequently the resulting review. To the best of our knowledge, we are the first to investigate how a review template might influence reviewers.

Furthermore, our results carry substantial managerial implications for review system designers since they suggest that using a review template can guide reviewers to writing not only longer reviews but also ones that cover more features, and to do so without increasing the perceived effort involved in writing a review. Studies have shown that both length and number of features have a positive impact on perceived helpfulness (Scholz and Dorner, 2013; Pan and Zhang, 2011), which in turn affects customer decision making (Wu et al., 2015) and sales (Forman et al., 2008). Additionally, review system designers could adopt our approach to develop their own review templates.

### 2 Related Literature

Our study is related to two streams of literature. First, our analysis of how a design feature of an online review system affects review content is related to research focusing on design features and their influence on review text. Second, as the outcome variable of our paper is the review text, our research is also related to studies that analyze how aspects of review texts influence the perceived helpfulness of reviews.
First, our research is related to studies that investigate the relationship between the design of online review systems and review texts\(^2\). For instance, offering an opportunity to write reviews on a mobile device leads to review texts becoming more concrete, more emotional (measured by Language Inquiry and Word Count), and shorter (Burtch and Hong, 2014). These results are driven by the fact that the relative time between consumption and reviewing is shorter, and that the mobile phone display is smaller compared to desktop displays. Additionally, studies also show that when review systems ask previous customers via mail to provide a review, or even pay them, it negatively affects the length of a review (e.g., Kuan et al., 2015; Pan and Zhang, 2011), as these measures lead to a crowding-out effect of intrinsic motivation. Paying reviewers also has a negative impact on readability of review texts (e.g., measured by the Gunning-Fog Index) and leads to fewer product features being discussed (Khern-am-nuai et al., 2018). Another way to encourage reviewers to write longer reviews relies on social norms. If reviewers are presented with information on how many peers wrote a review, they tend to produce longer review texts (Burtch et al., 2018). We contribute to this stream of literature by being the first, to the best of our knowledge, to investigate feature-based templates as a new design feature that can influence reviewers’ writing process.

Second, the relevance of our study is based on research that scrutinizes how review texts contribute to reviews being perceived as helpful – typically measured by the number of helpful votes a review has received. As it has been shown that perceived helpfulness is positively associated with sales (Forman et al., 2008) and that reviews with a high helpfulness score play a greater role in customer decision making (Wu et al., 2015), understanding the factors impacting this measure has sparked much interest. Hong et al. (2017), for example, present an extensive review of the literature examining these factors. In the following, we focus only on aspects related to the review text.

Results from previous research suggest that length and readability have a positive impact on perceived helpfulness (e.g., Kuan et al., 2015; Pan and Zhang, 2011). Regarding text length, the product type has been found to moderate this relationship, with length having a more pronounced influence in the case of search goods (Mudambi and Schuff, 2010). Next to length, studies suggest that emotions expressed in the text also affect helpfulness. For example, anxiousness conveyed by the text leads to a more positive perception than anger (Yin et al., 2014). Also, the semantic style of the review text (e.g., basic or stylistic) and correct spelling (Scholz and Dorner, 2013) influence helpfulness (Cao et al., 2011). Regarding the content of the text, it has been found that reviews discussing several product or service features as well as those being the first to discuss rare product features are considered more helpful by readers (Scholz and Dorner, 2013). Finally, reviewers giving a neutral rating (e.g., 3 out of 5 stars) receive more helpful votes for their reviews if they structure their text by pros and cons (Schlosser, 2011). As review templates can potentially affect the content of review texts, in terms of contribution to the body of knowledge on review system design it is important to investigate the impact of, respectively, specific features, emotions conveyed, general readability, length, and structure of reviews of products or services.

\section{Theoretical Background and Hypotheses}

When consumers write a review, two cognitive processes determine the final review: (1) Thinking of the product or service features to write about and (2) Choosing the right words to convey the intended information.

First, reviewers need to recall their experiences of consumption and remember relevant product or service features, e.g., they need to recall how they liked the food of their last restaurant visit or remember what features they like or dislike about their camera. Recalling is especially important when services are to be reviewed, as a second consumption instance cannot be obtained without cost. Nevertheless,

\footnote{Note that there are also other factors that shape review texts, e.g., how reviewers’ cultural background impacts the extent to which emotions are expressed in review texts (Hong et al. 2016). However, in contrast with our case, these studies do not examine a design change. Moreover, to the best of our knowledge there is no literature on how the provision of a review template is associated with the characteristics of the resulting review (Gutt et al. 2019).}
remembering product or service features is relevant for both products and services. A consumer might easily remember a broken button on a camera but might forget about its long battery life. The Adaptive Character of Thought theory from cognitive science (e.g., Anderson, 1996) explains how human cognition works while dealing with a task, and how specific thoughts, ideas, and experiences are the result of production rules. Thus, if reviewers engage in the task of writing a review, certain thoughts arise that will be used to develop the review text. If there are no additional external stimuli such as pre-structured categories that activate production rules, the task is solved on the basis of a limited set of thoughts. External stimuli activate production rules automatically (Anderson, 1992) and enhance exploration of the solution space. Studies have validated this argumentation empirically. Individuals who are not given a pre-structured task do not explore the full solution space but nevertheless overestimate their performance (Gettys et al., 1987). Providing pre-structured tasks also increases brainstorming performance (Dennis et al., 1999). Additionally, a pre-structured format can increase the amount of content that is recalled from memory (Pollio and Gerow, 1968).

Therefore, a pre-structured review template that presents product or service features should help reviewers to remember and convey more experiences. We argue that a review template affects (1) the features discussed in the review text and (2) the length of the review text. If a template provides product or service features, these represent external stimuli that influence what reviewers consider for their review text. As literature has shown, such external stimuli should lead to information being provided about more features. Thus, we formulate our first hypothesis as follows:

**Hypothesis 1:** Review texts written with a feature-based review template contain information on more features when compared to those written without a template.

Following our first hypothesis, we also argue that reviewers need more characters to describe more features, which directly translates into longer reviews. Hypothesis 2 captures this argument.

**Hypothesis 2:** Review texts written with a feature-based review template are longer when compared to those written without a template.

Note that it is also be possible that the features of a review template cannibalize each other so that reviews are not longer. As such an alternative explanation exists makes it worth investigating this hypothesis.

Second, reviewers need to verbalize their experiences with the product or service in form of a review text. The Cognitive Process Theory of Writing presents a framework that describes the writing process (Flower and Hayes, 1981). According to this theory, three major processes are relevant to writing, namely, **planning**, **translating** and **reviewing**. **Planning** involves extracting information from the task environment and from long-term memory to establish a writing plan. For example, reviewers need to remember what properties of their camera, such as a long battery life, are relevant for their review or take relevant features from the review environment in form of a template. **Translating** means formulating language based on information retrieved from the writer’s memory and the task environment. During the **reviewing** process the author reads and edits the result of the translation process. Because review templates should mainly interact with the reviewer’s thoughts before they are translated into language, we focus on the planning process to derive our hypothesis. The subprocesses **generating**, **organizing** and **goal-setting** form the overall planning process. **Generating** describes the retrieval of information from memory and the evaluation of topics (i.e., processed memories) with respect to their suitability for the task. For instance, writing about a broken button might be more relevant than writing about the color of a camera. When **organizing**, authors evaluate the usefulness of each topic and order the topics deemed useful. This subprocess also involves identifying topic categories. A reviewer might decide to first write about failure-related topics, like a broken button, before focusing on taste-related aspects regarding how they like the usability of the camera’s software or camera’s color. Finally, **goal-setting** refers to the definition of criteria with which the quality of the final text can be judged (e.g., suitability for the audience).
The Cognitive Process Theory of Writing has been used in many empirical studies as a theoretical foundation to study writing processes (e.g., Olive et al., 2002). While there are contradictory results regarding which of the three major processes are the most demanding (e.g., Alves et al., 2008), it is evident that planning emerges as a demanding process (e.g., Kellogg, 2001). Since feature-based review templates would offer ideas for relevant topics, the generating subprocess should become less demanding. Additionally, review templates can also provide an order suggesting which feature to write about next. Therefore, we argue that review templates reduce the perceived effort by targeting the planning process of writing.

Hypothesis 3: Reviewers exposed to a feature-based review template perceive writing the review as less demanding (i.e., involving a lower perceived effort) compared to those written without a template.

As with the second hypothesis, there is an alternative explanation: Having to review multiple features could actually increase a reviewer’s effort compared to reviewing just one feature. At the same time, it could be cognitively less demanding when the template presents the reviewer with external stimuli.

4 Developing a Review Template

The following section is structured as follows. First, we propose a general approach for generating online review templates which can be applied to any product or service context. Second, we apply our approach to a dataset of restaurant reviews.

4.1 General Approach for Generating Online Review Templates

Before testing our hypotheses, we need to develop a review template which has the potential to facilitate the writing process. Following on from our theoretical discussion, the review template needs to provide features that represent external stimuli to the reviewer\(^3\). We propose a data-driven approach to identify these features – as illustrated in Figure 1. The central aspect of our approach is based on topic modeling to extract relevant features from the review text. Literature has shown that topic models are able to extract features on product and service quality from online review texts (e.g., Tirunillai and Tellis, 2014).

Data Gathering and Pre-Processing

We build the template based on previous online review data by gathering review texts on the product or service for which we want to develop a template (e.g., restaurants, laptops, etc.). These reviews form our initial dataset.

During the pre-processing, we first sort the reviews by the count of helpfulness votes and restrict the sample only to the top 5% of reviews marked as helpful. These reviews are, according to literature, especially important to customer decision making (Hu et al., 2014; Wu et al., 2015). Therefore, they should contain relevant features that could be contained in a template. Second, using Python’s langdetect library, we remove all non-English reviews.

\(^3\) An earlier version of this approach is discussed in Poniatowski et al. (2019).
Next, following previous literature (Tirunillai and Tellis, 2014; Debortoli et al., 2016), we transform the review texts in preparation for the subsequent topic modelling. In particular, we use the Natural Language Toolkit (http://www.nltk.org/) to apply Part-of-Speech Tagging and cleanse the review texts of all words that are not adjectives, nouns, verbs or adverbs since other words would not be suitable to indicate meaningful features. We stem the remaining words with Porter’s (1980) stemming algorithm to transform words such as “desserts” to “dessert”. Additionally, we use the lexical database WordNet (https://wordnet.princeton.edu/) to lemmatize words (i.e., “better” is transformed to “good”). Based on the default list of English stop words provided by WordNet, we define a list of stop words and remove all words from the review text which are contained in that list. The list can be extended if necessary. For instance, when analyzing restaurant reviews, it may be useful to add restaurant categories (e.g., “Italian”, “Mexican”, etc.) to the list as these contain little informational value. For products, it is useful to add brand specific stop words. Finally, we remove all words which appear in more than 99% of the reviews (Tirunillai and Tellis, 2014).

**Topic Modelling**

In line with previous literature we chose latent Dirichlet allocation (LDA) to identify topics within the reviews (Tirunillai and Tellis, 2014), and the Python package scikit learn (Pedregosa et al., 2011), which provides, amongst other things, probabilistic topic modeling based on LDA. LDA is an unsupervised machine-learning method used to derive topics from a large collection of documents with textual data (here: online reviews). It assumes that each document \( D \) consists of \( T \) topics allocated by a discrete distribution, which is set priorly. The topics \( T \) are associated with words \( W \) from a discretely distributed vocabulary (Blei et al., 2003; Blei, 2012). In other words, a document \( D_i \) is defined by a probability distribution over fixed number of topics \( T \), where the probabilities for each topic add up to 1 and range between 0 and 1. Further, a topic \( T_i \) is defined by a probability distribution over a fixed set of words \( W \), where the probabilities for each word add up to 1 and range between 0 and 1 (Debortoli et al., 2016). In summary, the LDA appoints topics to the documents by assigning probabilities between 0 and 1 for each topic, implying the likelihood of a certain document being associated with a particular topic.

An important aspect with running the LDA is the a priori set number of topics. On the one hand, if the number configured is too high, the LDA identifies topics with only minimal distinctions like different writings styles and the risk of double topics with the same meaning. On the other hand, if the number of topics is set to low, the LDA might identify fewer topics than the document collection actual comprises. Literature suggests two ways to determine the number of topics, either by trying out different numbers followed by a visual examination of the results (Debortoli et al., 2016), or by computing it mathematically, e.g., by calculating the marginal log-likelihood with cross-validation (Tirunillai and Tellis, 2014). We employ the second possibility, as it does not rely on personal interpretation. In particular, we iteratively increase the number of topics starting with two topics. In each iteration, we do a fivefold cross-validation and calculate the harmonic mean of the resulting log-likelihood values. We repeat this process as long as the average log-likelihood increases and choose the topic number with the maximum log-likelihood. As a result of the LDA, all reviews are assigned a value for each topic which represents the estimated probability that the text belongs to the respective topic. Each topic contains words from the review texts which describe the associated topic. For each word, there is a probability indicating the word’s importance in describing the topic.

**Template Development**

The output of the topic modelling (i.e. LDA) is then used to extract features for the final review template. Following previous literature (Debortoli et al., 2016), we conduct a visual examination of the top ten words associated with each topic and the five reviews with the highest probability for each topic. Each author independently analyzes these words and the reviews to find meaningful features. These features are discussed among the author team to determine a final list of features. The review template will then consist of a separate text field for each feature.
4.2 Developing a Review Template for Restaurant Reviews

To investigate the impact of using a review template in an experimental setting, we choose a product or service that can be evaluated by many participants without them having to consume the product again. We decided on restaurant visits as these represent a service which a lot of people consume frequently. Consequently, we ask participants to review their last restaurant visit. This approach has the advantage that the participants have actually consumed the good without the need of creating an artificial consumption instance.

We apply the approach presented in section 4.1 to existing online reviews from the platform Yelp which have been released by the platform during its 9th Academic Dataset Challenge (https://www.yelp.com/dataset/challenge). It contains around two million reviews for all businesses categorized as restaurants in the US. During the pre-processing we add restaurant-specific stop words. First, we get a list of all restaurant categories listed on Yelp such as “Italian” or “Mexican” and add them to the list of stop words as these contain information applying only to a subgroup of restaurants. Moreover, an initial LDA revealed words representing a specific food type such as “sandwich” or “pasta”. To harmonize this information, we pre-process each word by examining whether it is a hyponym of the food categories “nutrient” or “solid food” provided by WordNet. If the word can be classified as food, we replace it with the string “food”. After restricting the sample to reviews in English voted to be most helpful, we consider about 80,000 online reviews to develop a review template.

Following the marginal likelihood approach with fivefold cross-validation, we conclude that the optimal number of topics is 17. As a result of our visual examination and discussion among the authors, the most frequently mentioned dimensions used to describe a topic were: food, service, place, atmosphere/ambiance, and value for price. For instance, the topic with the words “tast flavour top slice make red crust serv fresh fri” was labelled “food”. Other topics, such as the word list “went, cousin, sammi, ami, kay, obnoxi, hookah, costco, aisl, rachel” did not yield a suitable service feature. To validate whether these features actually appear in reviews, we gave a random sample of the top 5% of helpful reviews consisting of 300 reviews to two human coders, who independently assessed the features that are considered in the reviews (Lombard et al., 2002). As a result, we find that, on average between the two coders, at least one of the dimensions is present in about 99% of the sample and thus the dimensions capture the majority of the reviews’ content. Excluding the most obvious dimension “food” reveals that about 91% include at least one of the remaining dimensions.

5 Research Method

To test our hypotheses, we design an experiment which investigates the effect of a review template on different dependent variables such as the coverage of features, length of the review, and the perceived effort to write the review. To investigate this potential influence, we designed an experiment on the commercial crowdworking platform FigureEight (https://www.figure-eight.com). Although this platform offers access to millions of potential contributors distributed worldwide, we limited the access to our task to participants in the US because (1) the Yelp reviews we use to develop the review template were all written for US-based restaurants and (2) because we want to avoid cultural differences since individuals from different countries may perceive the experience of restaurant visits differently. The task itself looked as follows: Having developed our review template on existing Yelp reviews about restaurant reviews, we asked the crowd to remember their last restaurant visit and write a review about it. After asking several demographic questions about participants’ age and sex, we asked each participant to rate their last restaurant visit. More specifically, we first ask them to give it a rating on a five star rating scale and then to write a review about their experience. To answer our research question, we designed one control and two different treatment conditions (see Figure 2: Visualization of the Experiment Design), in which we varied the use and the form of the template.

Note that words were stemmed and lemmatized which yields “tast” for a word like „taste“. 

---

4 Note that words were stemmed and lemmatized which yields “tast” for a word like „taste“.
Participants in the control condition had to write a free text review without using a review template. In contrast, participants in our first treatment condition were exposed to our review template, consisting of five separate features. We further designed a second treatment condition where we displayed the review template as well as specific sample words for each feature of the review template. Figure 3 illustrates the different perspectives for each condition. Participants were randomly assigned to one of the three experimental conditions.

Once participants had written the review, we further asked them to rate their perceived effort on a five-point scale between (1= very difficult; 5= very easy) and measured the actual time it took participants to write the review.

With this design, we are pursuing several objectives: First, we are able to analyze the content of each review in terms of covered features to test our first hypothesis. Second, by comparing the length of the reviews written in the different experimental conditions we are able to investigate whether use of a review template stimulates the writing process, leading to longer reviews, in line with our second hypothesis. Third, by analyzing the reported perceived ease of writing as well as time spent on the task we are able to investigate whether the review templates reduce the perceived cognitive effort to write a review as well as whether individuals spent more time on solving the task, all of which may have important managerial implications.
In sum, 298 persons (112 in the control condition, 98 in Treatment 1, and 88 in Treatment 2) participated in our experiment. Of these, 148 female participants took part in our experiment while 28 participants did not indicate their gender. The average reported age of participants in our crowd was 35.8 years, excluding 28 participants who did not state their age. Each participant received a fixed wage of 0.20$. We did not offer an additional incentive for writing long(er) reviews. Participants were also allowed to write nothing and would still get paid. Participants were not able to assign themselves to the task multiple times, however, and since they could use different worker accounts, we further checked the IP addresses of each participant and excluded 32 duplicates.

6 Experiment Results

In the following, we report the results of our experiment. This section is separated into three sub-sections: in the first we report the results connected to our first two hypotheses, in the second we examine those related to our third hypothesis, and in third we present further analyses undertaken, e.g. robustness checks.

6.1 Review Content

To answer the question whether the use of a review template positively influences both the number of features that appear within a review and the number of characters written by the reviewer, we first had to determine how many features appeared in each review. To do that, we recruited two human coders who first checked whether the content of the reviews makes sense in general and next, examined the coverage of the derived features (i.e., food, service, place, atmosphere/ambiance and value for price) for each review independently. Based on the human coders’ assessment, we calculate the interrater reliability (Lombard et al., 2002) for each feature, using the percentage match as well as two conservative standard measures for evaluating interrater reliability (Lombard et al., 2002). Results (see Table 1) indicate acceptable interrater agreements (above 90% agreement; Cohen’s $\kappa > 0.8$; Krippendorff’s $\alpha > 0.8$) for all features, except for “place”. The reason for this result could be that this feature seems to be ambiguous, since it has been derived from reviews considering its location (e.g., inside a shopping mall or in the reviewer’s hometown) as well as the restaurant’s accessibility (e.g. parking spots in front of the restaurant). One possible solution to avoid ambiguity could be by naming the feature “Place/Location”.

<table>
<thead>
<tr>
<th>Feature</th>
<th>%-Agreement</th>
<th>Cohen’s $\kappa$</th>
<th>Krippendorff’s $\alpha$</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
<td>96.3%</td>
<td>0.914</td>
<td>0.914</td>
<td>298</td>
</tr>
<tr>
<td>Service</td>
<td>96.6%</td>
<td>0.924</td>
<td>0.924</td>
<td>298</td>
</tr>
<tr>
<td>Place</td>
<td>64.4%</td>
<td>0.203</td>
<td>0.073</td>
<td>298</td>
</tr>
<tr>
<td>Atmosphere/Ambiance</td>
<td>92.6%</td>
<td>0.852</td>
<td>0.852</td>
<td>298</td>
</tr>
<tr>
<td>Value for Price</td>
<td>96.4%</td>
<td>0.926</td>
<td>0.926</td>
<td>298</td>
</tr>
</tbody>
</table>

*Table 1: Interrater Reliability*

However, based on the overall high interrater agreement, we conclude that the features have been assessed correctly, allowing us to calculate the feature coverage for each experimental group. To test our first hypothesis, we compare the number of dimensions covered by reviews written in each experimental condition. Results of a comparison using Mann-Whitney tests (Table 2) show that the number of features covered in the reviews written with the use of the review template are significantly higher based on results from both human coders, and consistent for both treatment conditions. Thus, we find support for Hypothesis 1.: as results indicate that reviews written with a feature-based review template contain information on more features compared to those written without a template.

---

5 Distributions are not completely uniform amongst the conditions due to removing observations with duplicated IPs.
Control Group | 1st Treatment Group | 2nd Treatment Group
---|---|---
Human Coder 1 | 2.392 (0.086) | 4.130 (0.090) | 3.783 (0.112)
p-Value | <0.01 | <0.01 | Human Coder 2 | 2.347 (0.090) | 4.800 (0.068) | 4.875 (0.068)
p-Value | <0.01 | <0.01

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Error</th>
<th>Std. Deviation</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>112</td>
<td>143.804</td>
<td>80</td>
<td>15.457</td>
<td>163.583</td>
<td>0.089</td>
</tr>
<tr>
<td>1st Treatment</td>
<td>98</td>
<td>194.439</td>
<td>119.5</td>
<td>21.428</td>
<td>212.129</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>112</td>
<td>143.804</td>
<td>80</td>
<td>15.457</td>
<td>163.583</td>
<td>0.672</td>
</tr>
<tr>
<td>2nd Treatment</td>
<td>88</td>
<td>162.682</td>
<td>73</td>
<td>23.231</td>
<td>217.923</td>
<td></td>
</tr>
</tbody>
</table>

Results (see Table 3) show that reviews written with the support of a feature-based template (1st Treatment) were statistically significantly longer by on average of about 50 characters (p-value = 0.089) than those written without the template (Control). Comparing the control and second treatment groups reveals no significant (p-value = 0.672) divergence for the review length. Although participants in the second treatment condition did not write longer reviews, results of our first comparison (Control vs. 1st Treatment) partially support our second hypothesis. The reason why the review length of the second treatment group does not differ significantly from those written without a template (Control) could be as follows: Due to the additional sample words given in the feature fields (see Figure 3) in the 2nd Treatment, participants might be less motivated to use their own words to write longer reviews, as they could use the sample words provided instead of writing longer sentences. A first analysis indicates that the occurrence of the sample words in the review texts is higher for the second than the first treatment. The differences between these two conditions are statistically significant for the features “place”, “atmosphere/ambiance” and “value for price” (p-value < 0.01), and statistically significant for the other two features (p-value < 0.05). This supports the notion that participants in the second condition tend to write their review using the sample words provided. Another explanation for shorter reviews within this group could be that since we did not provide any instructions with regards to the review form, participants might assume that reviews consisting of a few words instead of complete sentences are expected. Further analysis of the number of words (approximated by the number of blanks) and number of sentences (approximated by the number of periods) supports this assumption. It reveals that participants of the control group used on average about 26.56 words, while participants of treatment groups one and two used on average 35.73 and 29.91 words, respectively, to write a review. Furthermore, the average number of used periods is about 1.92 for the control condition, 2.38 for participants of treatment one, and 1.83 for treatment two."
This supports the assumption that participants might write shorter reviews consisting of few words rather than formulate complete sentences.

### 6.2 Writing Effort

To examine whether the writing process of a review using a featured-based review template is less demanding (i.e. decreases the perceived effort of writing a review) compared to writing it without a template (Hypothesis 3) we gather the perceived difficulty to write the review as well as how long it took to complete it: First we analyze the perceived difficulty of writing a review (i.e., perceived effort) by asking participants to assess the perceived difficulty of writing the review (“How difficult was it to write this review?”) on a scale from 1 (“very difficult”) to 5 (“very easy”). Table 4 below shows the results of comparing this data between control and treatment groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Error</th>
<th>Std. Deviation</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>107</td>
<td>4.29</td>
<td>5</td>
<td>0.094</td>
<td>0.971</td>
<td>0.330</td>
</tr>
<tr>
<td>1st Treatment</td>
<td>93</td>
<td>4.11</td>
<td>4</td>
<td>0.120</td>
<td>1.160</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>107</td>
<td>4.29</td>
<td>5</td>
<td>0.094</td>
<td>0.971</td>
<td>0.133</td>
</tr>
<tr>
<td>2nd Treatment</td>
<td>86</td>
<td>4.05</td>
<td>4</td>
<td>0.124</td>
<td>1.147</td>
<td></td>
</tr>
</tbody>
</table>

**Table 4: Perceived Difficulty by Experimental Group**

Based on the Mann-Whitney test, we could not determine any significant differences in the perceived difficulty between the control group and the treatment groups. Considering these results, we cannot conclude that providing a review template reduces or increases perceived effort. Therefore, we analyze the duration as a dependent variable in the following.

The time a participant took to write a review is measured in seconds. Table 5 presents the durations of the treatment groups in comparison with the control group.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Error</th>
<th>Std. Deviation</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>112</td>
<td>98.74</td>
<td>51.10</td>
<td>13.055</td>
<td>138.160</td>
<td>0.014</td>
</tr>
<tr>
<td>1st Treatment</td>
<td>97</td>
<td>120.63</td>
<td>92.847</td>
<td>11.629</td>
<td>114.533</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>112</td>
<td>98.74</td>
<td>51.10</td>
<td>13.055</td>
<td>138.160</td>
<td>0.802</td>
</tr>
<tr>
<td>2nd Treatment</td>
<td>87</td>
<td>111.76</td>
<td>57.319</td>
<td>14.883</td>
<td>138.820</td>
<td></td>
</tr>
</tbody>
</table>

**Table 5: Review Duration of Experimental Groups**

Comparing the means of the time participants took to write the review reveals significant differences. Compared to the control group, participants in the first treatment condition take significantly longer (p-value = 0.014). Regarding the second treatment condition, the results indicate no significant differences. Conducting a correlational analysis regarding the number of used sample words and duration indicates that the number of used sample words is not correlated with the reviewing duration. Thus, this yields a first indication that providing sample words does not affect reviewing duration. Therefore, we assume that missing instructions regarding the review form and the resulting lack of complete sentences could be the reason for this statistically insignificant result.

Combining these results with our insights on the perceived effort, we can conclude that a feature-based template without sample words increases the actual effort in terms of time taken, albeit that participants do not perceive it as such. Thus, our third Hypothesis 3 is partially supported.
6.3 Further Analyses

We took measures to check the robustness of our results. Since one might argue that our results could be driven by outliers, e.g., extremely long reviews, we apply the three-sigma rule (Pukelsheim, 1994) to identify outliers for exclusion. According to the rule, all observations with values above three standard deviations from the mean can be considered as outliers. Thus, we drop all observations with reviews longer than three standard deviations from the mean (i.e., longer than 760 characters) and repeat our analysis. Additionally, we drop all observations where participants wrote nothing. After restricting the sample in this way our results remain qualitatively unchanged. As the sample size shrinks in this case, the results are not statistically significant with regard to review length. Still, comparing review lengths shows that reviews of the first treatment condition are longer on average. The results regarding the coverage of features remain statistically significant even after implementing this restriction.

Moreover, to enhance our findings regarding the impact of review templates on reviewers, we further investigate other descriptive measures that could be of interest. For instance, it could be interesting for researchers and review system designers to know which condition led to the fewest reviews with no text, or to particularly long reviews. Google Maps Reviews for example has recently implemented incentives to increase the number of reviews with a textual component and thus decrease those that only contain a rating but no text. Therefore, knowing whether templates could achieve such a goal is relevant to practice. Table 6 presents these additional measures. For each condition we examine the number of reviews with no content, the number of reviews contained in the top 10% of longest reviews, and the numbers of reviewers who found writing the review either “very easy” or “very difficult”. The fewest reviews with no content are found in the first treatment group (16%). This indicates that the template helped participants to begin writing and to finish the task. Additionally, the first treatment group contains the highest number of particularly long reviews (i.e., contained in the top 10% of the longest reviews in our sample). In line with our experimental results, in every condition there are about 50% of participants who found writing the review “very easy” and only a few participants out of every group who found it “very difficult”. As explained above, given the longer reviews and higher feature coverage, the differences in perceived effort remain insignificant when writing with or without a template, while both the actual effort and length of output increase. Moreover, we introduce an additional measure based on the perceived difficulty as well as the review length for writing effort, calculated by dividing the review length in characters by perceived difficulty to write the review. For example, in case a participant writes a review of 165 characters (i.e. the average review length) and perceives writing the review as very easy, the measure is $\frac{165}{5} = 33$. Based on comparing these measures, it seems that writing the review is relatively easier for participants from the control group compared to the other participants. Additionally, we calculated the Flesch-Kincaid Grade Level score to measure the readability of the resulted reviews, following previous literature (e.g., Singh et al. (2017)). This reveals that the reviews written by the treatment groups with regard to readability differ significantly from the control group (treatment one group: p-value < 0.05 and treatment two group: p-value < 0.01) and, based on the Flesch-Kincaid Grade Level, are easier to read.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Control</th>
<th>1st Treatment</th>
<th>2nd Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviews with zero length</td>
<td>26 (23%)</td>
<td>16 (16%)</td>
<td>30 (34%)</td>
</tr>
<tr>
<td>Top 10% of reviews in terms of length</td>
<td>7 (6%)</td>
<td>14 (14%)</td>
<td>7 (8%)</td>
</tr>
<tr>
<td>Number of reviewers who found review writing „very easy“</td>
<td>62 (55%)</td>
<td>46 (47%)</td>
<td>40 (45%)</td>
</tr>
<tr>
<td>Number of reviewers who found review writing „very difficult“</td>
<td>1 (1%)</td>
<td>6 (6%)</td>
<td>5 (6%)</td>
</tr>
<tr>
<td>Review length divided by perceived difficulty</td>
<td>36.059</td>
<td>52.966</td>
<td>42.817</td>
</tr>
<tr>
<td>Flesch-Kincaid Grade Level score</td>
<td>1.546</td>
<td>0.530</td>
<td>0.014</td>
</tr>
</tbody>
</table>

Table 6: Further Measures
7 Conclusion

As online reviews have become an inherent part of online shopping and third-party platforms, they present a valuable source of information for customers. Online review systems rely on customers to contribute reviews of high informational value so that future customers can learn from them. Therefore, nudging customers towards contributing this content, for instance by showing the customer’s reviewing behavior in relation to other reviewers (Burtch et al., 2018), has received increasing interest from practice and research. However, little is known on how the design of online review systems can support reviewers in their writing process. In this study, we try to narrow this gap by analyzing review templates as one such method which has been proposed by previous literature (Kwark et al., 2014; Yin et al., 2014; Kuan et al., 2015). In particular, we focus on templates that provide reviewers with product or service features, which they can consider while writing their reviews (i.e., feature-based review templates).

Following the Adaptive Character of Thought theory (e.g., Anderson, 1996), we hypothesize that feature-based templates help reviewers to (1) cover more features in their texts and (2) write longer reviews. Our hypothesis is based on the notion that the template provides external stimuli to the reviewer which would otherwise not be considered. Drawing on insights from the Cognitive Process Theory of Writing (Flower and Hayes, 1981), we argue that, as review templates provide a structure for the review, planning and organizing the review text should involve less cognitive effort. We propose an approach to developing such a feature-based template based on existing online review data, apply the approach to restaurant data from the online review platform Yelp and use the resulting template in an experiment we conduct on an online crowdfunding platform. Our results reveal that participants, when presented with separate text fields for service features, write significantly longer reviews that contain more topics and take a longer time to write, compared to when being presented with a blank text field. Interestingly, though, reviewers do not report different levels of perceived effort across these conditions, i.e., longer reviews with more topics come at no additional costs in terms of perceived effort.

To the best of our knowledge, this study is the first to examine the impact of online review templates on the content and writing process of online reviews and contributes to the existing literature on the relationship between the design of online review systems and review texts. Our results provide substantial implications for review system designers, since they suggest that implementing a feature-based review template can be beneficial for reviews in terms of length and number of covered topics. These metrics are positively associated with perceived helpfulness (Pan and Zhang, 2011) which has been shown to have a positive influence on sales (Forman et al., 2008) and customer decision making (Wu et al., 2015). Review system designers can also adopt our approach to generate a feature-based review template for their own review system. It is important to make practitioners aware of the influence that review templates can – and do – have on reviewing behavior.

Naturally, there are limitations to our study. Since we are running an experiment to obtain these results, it is essential to evaluate the external validity of our study by investigating review templates in the field with existing online review systems. Especially the usage of a commercial crowdfunding platform and, therefore, a financial incentive – however fixed – could have biased the review content. To avoid the risk of biased results, a field experiment could be conducted. Moreover, we investigate the influence of two specific variants of review templates, whereas other types of templates might yield different results. Similarly, we only focus on restaurants as the subject of reviews.

Our findings open up several avenues for future research. The related literature we present above has pointed to other metrics that can affect perceived helpfulness, such as emotions conveyed in the text. Whether review templates affect these metrics still needs to be analyzed. Moreover, to analyze how templates affect review diversity, we plan to determine an average review based on our dataset and calculate the diversity of the reviews produced in our experiment by determining the distance (e.g., cosine distance) to the average review. Finally, a follow-up experiment to investigate whether the reviews written in our study with the help of a template are perceived as more helpful is important to further understand the impact on review readers. In line with this, further measures for review helpfulness such as emotions and readability (e.g., Flesch-Kincaid Grade Level score) should be included in the analysis. In such an experiment, different presentation styles of the template (e.g., presenting plain text or features as subtitles in the review) could also be studied.
8 References


