A TEST OF A COMPUTER-ADAPTIVE SURVEY USING ONLINE REVIEWS

Research paper

Sahar Sabbaghan, University of Auckland, Auckland, New Zealand, s.sabbaghan@auckland.ac.nz

Cecil Eng Huang Chua, University of Auckland, Auckland, New Zealand, aeh.chua@auckland.ac.nz

Lesley A. Gardner, University of Auckland, Auckland, New Zealand, l.gardner@auckland.ac.nz

Abstract

Traditional surveys are excellent instruments for establishing the correlational relationship between two constructs. However, they are unable to identify reasons why such correlations exist. Computer-Adaptive Surveys (CAS) are multi-dimensional instruments where questions asked of respondents depend on the previous questions asked. Assessing the validity of CAS is an underexplored research area as CAS differs from traditional surveys. Therefore, validating a CAS requires different techniques. This study attempts to validate the conclusion validity of a CAS about café customer satisfaction using online customer reviews. For our CAS to have conclusion validity, there should be a high correspondence where most respondents in CAS and online reviewers both agree that certain constructs are the cause of their dissatisfaction. We created a Computer-Adaptive Survey (CAS) of café satisfaction and used online customer reviews to assess its conclusion validity. Our research thus contributes to the measurement literature in two ways, one, we demonstrate that CAS captures the same criticisms of cafes as that in online reviews, and two, CAS captures problems about customer satisfaction at a deeper level than that found in online reviews.

Keywords: computer-adaptive, online reviews, conclusion validity.

1 Introduction

Computer-Adaptive Surveys (CAS) are most useful for the situation where there are two or more constructs which are correlated and one desires to understand the reason why those constructs are correlated. Consider a scenario where a café owner wants to know not only what aspect of customer service he could improve on, but also how he can improve (Fundin and Elg 2010; Sampson 1998; Wisner and Corney 2001). With traditional survey techniques, he would be forced to give customers a very long survey to identify the salient issues for improvement. In long surveys, respondents will suffer from a fatigue effect, and not answer questions properly (Berdie 1989; Deutskens et al. 2004).

In the past, there have been other methods for extracting customers' opinions and views, such as using text mining on online reviews (Jindal and Liu 2007; Somprasertsri and Lalitrojwong 2010; Xu et al. 2011). However, text mining has several limitations, such as the writer's intention is not always clear, as a statement can be expressed in a genuine positive sentiment in one context and in another it can be used sarcastically (Cambria et al. 2013; Grégoire et al. 2014). Another example is that while most review opinion mining studies have concentrated on sentiment analysis (thumbs up or down)(Gräbner et al. 2012; Liu 2012), the issue of analysing the depth of customer experience feedback is quite often ignored.

In CAS, the response from previous questions determine the next questions asked. CAS differs from traditional surveys in several ways. First, the items in CAS are arranged in a hierarchy, whereas traditional methods assume a "flat" set of items. Second, respondents legitimately only fill in some of the questionnaire items- unfilled questions cannot be treated as non-responses.

Assessing the validity of CAS is an underexplored research area as CAS differs from traditional surveys. Therefore, validating a CAS requires different techniques. This paper attempts to determine whether CAS does what it claims to do, i.e., whether it is a useful instrument for diagnosing root cause. Assessing the validity of CAS is an underexplored research area. As CAS differs from traditional surveys, validating a CAS requires different techniques. To do this, we have to compare CAS against external criteria that assess the same thing CAS measures. We selected online reviews as an external criterion because they are the best external evaluation basis of comparison. Online reviews are a good proxy, because the default way people assess others' satisfaction of a café today is online (Mudambi and Schuff 2010). In addition, they are recognized as having face and criterion validity, as online reviews demonstrate the relative strengths and weaknesses of a product or service (Jindal and Liu, 2007; Somprasertsri and Lalitrojwong, 2010; Xu, Liao, Li, and Song, 2011).

Our study obtained high correspondence between top-level constructs of our CAS and online reviews, but we found we could not make conclusions regarding lower level constructs. This was because online reviews did not go into as much detail about customer satisfaction as CAS did. Our research thus contributes to the measurement literature in two ways, one, we demonstrate that CAS captures the same criticisms of cafes as that in online reviews, and two, CAS captures problems about customer satisfaction at a deeper level than that found in online reviews.

The paper is constructed as follows. The next section introduces the related literature, describing CAS and its design. We then present preliminary results for CAS by comparing the results of CAS against blog and review websites. We conclude this paper with future work.

2 Computer Adaptive Surveys (CAS) in Customer Satisfaction

Customer satisfaction surveys assist organizations to understand customers' expectations so that organizations will be able to respond to, and serve customers' needs (Grigoroudis and Siskos 2010). However, it is often difficult to use traditional survey techniques to diagnose root cause. To identify root cause, it is often necessary to ask many questions. Most surveys are not designed to be very long. Therefore, they are not especially designed to be informative or diagnostic for identifying root cause (Goodman et al. 1992; Hayes 1992; Peterson and Wilson 1992). Also, in most surveys, the respondent is intended to answer the majority of questions. With a large survey, the respondent is likely to en-

counter fatigue and quit before providing critical information (Galesic and Bosnjak 2009; Groves et al. 2004; Groves 2006; Heerwegh and Loosveldt 2006; S. R. Porter et al. 2004). CAS provides a way to address this problem, because the effort to complete a CAS grows logarithmically with the length of the CAS. In contrast, effort grows linearly with the length of a traditional survey. This is because questions in a CAS are represented in a tree and the respondent navigates down one or a few branches of the tree instead of doing the whole survey.

For finding root cause, CAS offers certain advantages over traditional surveys. Its principal advantage is that it allows the survey developer to include a large number of questions. The only questions the respondent answers are the ones most salient to the issue being addressed- in our case, the things about the café the respondent is least satisfied with. In contrast, if the same number of questions were asked on a traditional survey, the respondent is likely to encounter fatigue and quit before providing critical information (Galesic and Bosnjak 2009; Groves et al. 2004; Groves 2006; Heerwegh and Loosveldt 2006; S. Porter et al. 2004).

In CAS, respondents perform a depth-first traversal of the tree, where each stage of the traversal involves the respondent rating all items in the stage. Respondents then receive only child constructs associated with the lowest or highest rated constructs. Each respondent could traverse the CAS hierarchy in a different way. To illustrate, see Figure 1, which presents a simplified example of a café satisfaction CAS. Note that this is a simplified example that does not detail all 176 items. If food is the area the customer is least satisfied with, CAS retrieves questions about the quality of the food (i.e., preparation, portion, menu choice). If the customer is least satisfied with menu choice, CAS then retrieves questions about how the food was cooked, taste, special needs, options, and availability. CAS does not retrieve further questions on constructs the respondent rated satisfactorily. As the respondent continues to answer questions, CAS navigates deeper down the tree and questions roll down until the respondent hits one set of constructs with no children, for example, that there are insufficient vegetarian items on the menu. If most respondents agree there are insufficient vegetarian items on the menu, this would indicate lack of vegetarian items is the root cause for many customers' dissatisfaction. Of course, not every respondent would navigate the hierarchy the same way. Thus, aggregating the results from respondents allows the researcher to observe the multiple major problems across respondents. In addition, it allows managers of commercial enterprises to quickly find key issues to address.

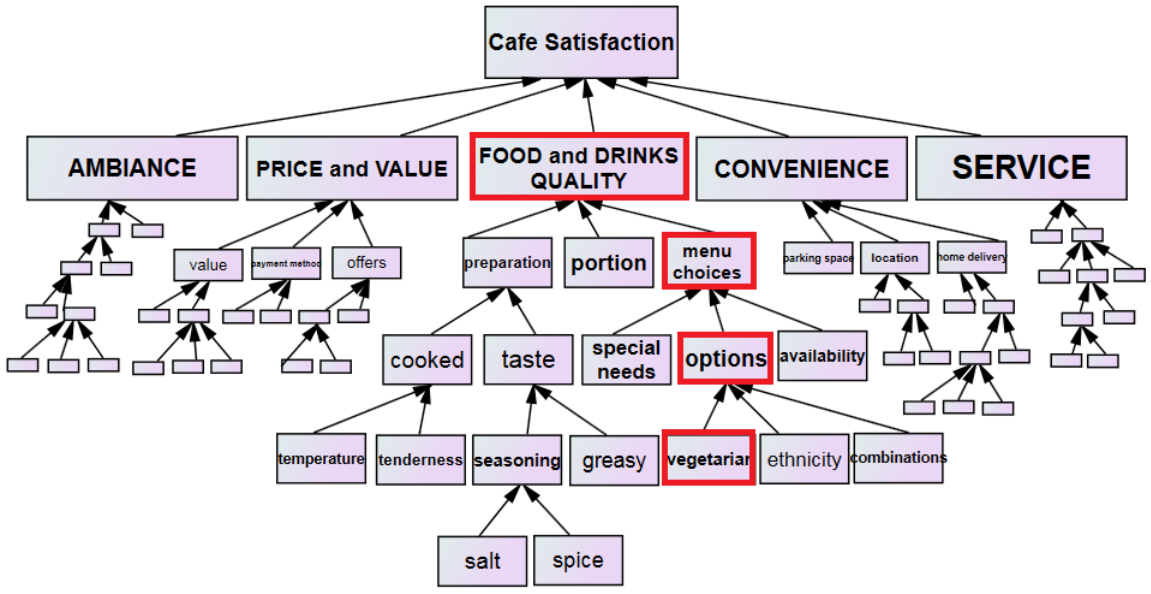

Figure 1. How CAS works for cafe satisfaction

A CAS can be thought of as a hybrid of a traditional perception survey and a Computer-Adaptive Test. Computer-Adaptive Tests are designed to efficiently assess and evaluate a respondent's ability or per-

formance by administering questions dynamically based on answers to the questions the participant answered previously (Thompson and Weiss, 2011). For example, the Graduate Management Admission Test (GMAT) asks the respondent to answer language and mathematical questions in increasing order of difficulty (Stricker, Wilder, and Bridgeman, 2006). The next question asked of a respondent depends on whether the previous questions were answered correctly. Similarly, in the Merrell and Tymms test (2007), the aim is to understand the reading ability of students to provide better feedback and implement appropriate reading practices.

CAS is comparable to CAT, but differs from CAT in several respects. One is that their goals are different; CAS aims to obtain a focused assessment of one or a few perceptual measures (e.g., which things did you like the least or most), while CAT assesses an ability or performance (Hol, Vorst, and Mellenbergh, 2008; Merrell and Tymms, 2007). Traditionally, the goal of the typical CAT is to produce a score evaluating ability or performance on a single or few constructs. In contrast, the goal of CAS is typically to identify which of many subconstructs are perceived by the respondent as most relevant to them.

These dissimilarities in goals result in structural incongruities between the two kinds of surveys. CAT relies on potentially complicated Item Response Theory (IRT) functions to determine further questions to ask respondents (Embretson and Reise, 2000; Lord, 1980; Thompson and Weiss, 2011; Thorpe and Favia, 2012). CAS, in contrast, uses an adaptive version of branching to arrange the questions. Lowest or highest scores on a set of questions causes the system to retrieve related, but more precise questions. The question structures are also different. On the GMAT, which is based on IRT, the "correct" answer adds a point to the score, while an incorrect one deducts from 0.25 to 0.20 from one's score. In contrast, items in CAS are more akin to those on traditional psychometric instruments that are designed to "load" on a construct.

Finally, initiation and termination in CAS and CAT function in specific ways. In most cases, respondents taking a particular CAT test all begin in the same way. In contrast, we could have the first 20 respondents taking a CAS begin with generic questions about the café. If we realise that most respondents are indicating issues with the service, the next 20 respondents might begin at a lower level of the hierarchy- on the service-related questions. Similarly, a CAT terminates when the CAT has enough information to perform a diagnosis, either when a fixed number of items have been answered (Babcock and Weiss, 2009; Ho and Dodd, 2012) or because further questions in the item bank provide no additional statistically meaningful information (Thompson and Weiss, 2011). In contrast, a CAS ends either when one has fully traversed a set number of branches of the hierarchy, or when the user reaches some threshold for a proxy for fatigue (e.g., user answers a certain number of questions).

3 Assessing the Credibility of CAS Results

Assessing the validity of CAS's results requires different techniques from traditional surveys for a number of reasons. One, unlike traditional surveys, a CAS cannot be used to find cause in the sense of there being an independent variable and dependent variable on the survey. In a CAS, there is an implicit dependent variable (e.g., customer satisfaction) that the survey does not ask. Second, in CAS, there are child- parent relationships, where constructs that are children to other constructs in the hierarchy should represent some dimension of the parent construct. As a result, in CAS, we can expect high correlations between a parent and one child (e.g., if a respondent says they are dissatisfied with service, there is at least one subdimension of service they are unhappy about). However, because the subdimensions are orthogonal, we expect low correlations between subdimensions (e.g., that someone is dissatisfied with the efficiency of service does not mean they are dissatisfied with the quality of service). Traditional statistical techniques cannot handle such complex correlations between items on a survey. Given the problems with traditional approaches, we argue for a new approach to assess the validity of CAS's results.

This paper solely focuses on assessing the conclusion validity of CAS. In previous work, we have assessed CAS against its equivalent traditional survey to evaluate which instrument is better at finding root cause (Sabbaghan, Gardner, and Chua, 2017b). We found that CAS has several advantages over the traditional survey, as CAS had a higher response rate, required fewer items for respondents to answer, which reduced fatigue effect, had better item discrimination in that respondents provided more extreme scores in CAS, and had a higher agreement among respondents for each item. As the goal of CAS is to measure individuals' perceptions, CAS typically identifies one or a few narrowly defined constructs that respondents as a whole have the greatest or least affiliation to. Hence for CAS to have credible results would mean that the "correct construct(s)" have been identified. Thus, the results need to be assessed against an external criterion, which is a direct and independent measure of what the CAS is designed to measure. In this study, the results of a café satisfaction CAS will be assessed by comparing the results to online reviews. Other forms of validity similarly require new techniques which have been purview of other research (Sabbaghan et al. 2016, 2017a), in which the construct validity of CAS has been assessed.

Online customer reviews are defined as "peer-generated product evaluations posted on company or third party web sites" (Mudambi and Schuff 2010). For the purpose of our study, online reviews are a good proxy for two reasons. One, they are the default way people assess others' satisfaction of a café, and they are the first thing people think of when they are looking for an independent assessment of other peoples' customer satisfaction is reviews (Mudambi and Schuff 2010). Two, online customer reviews are considered as valuable sources of information for identifying relative strengths and weaknesses of a product or service (Jindal and Liu 2007; Somprasertsri and Lalitrojwong 2010; Xu et al. 2011), as they provide an in-depth understanding of what aspect of the product or service is wrong (Barreda and Bilgihan 2013; Fu et al. 2013; Mudambi and Schuff 2010). Online customer reviews therefore have face validity and criterion validity.

In CAS, the frequency of responses to an item reflects the general level of satisfaction or dissatisfaction with that item. Items that customers do not respond to are those customers are satisfied with. If a customer satisfaction CAS is representative of customer dissatisfaction, then we would expect that the frequency of responses to an item would correspond to the frequency of that same item on online reviews. However, CAS is arranged in a hierarchy. Some items in CAS are more granular than others. Similarly, reviews can be general, or specific. If a reviewer makes a specific comment, that comment should be able to be mapped to both the specific item in CAS, as well as the item's immediate parent and all further ancestors. However, if a reviewer makes a general comment, that comment will be unable to be mapped to specific comments in CAS. Therefore:

Proposition 1. There should be a good correspondence between the CAS hierarchy and online customer reviews.

H1: There is a high degree of correspondence between the top-level constructs in CAS and online reviews.

Let x refer to a level of the CAS hierarchy, where the depth of the hierarchy ranges from 2..n.

Hxa: There is a high degree of correspondence between the xth level constructs and online reviews.

Hxb: The degree of correspondence of the xth level constructs with online reviews will be lower than the degree of correspondence of the (x-1)th level.

We use the term "correspondence" here, because what we mean is that the results of CAS should be "similar" to that of online reviews. However, the two may not be identical, as the population who perform online reviews are a specific subset of individuals who actually visit cafes. We are comparing two sets of frequencies from two samples (CAS and online reviews). If the two sets of frequencies are similar then they should have two characteristics, one, the two samples should come from a common distribution and two, there should be a high correlation between them. The first would indicate that there is an agreement in which constructs respondents are not happy with and the second is an indication of the level of agreement. As an example, if in the results of both café satisfaction CAS and online reviews, respondents agree that they are dissatisfied with the construct "price and value," then the next step would be to assess the strength of their agreement.

4 Methodology

We developed a CAS of café satisfaction comprising. 175 survey questions. Most customer satisfaction surveys comprise 30 items or less- because of a lack of respondent patience, often only a single question is asked per construct (Heerwegh and Loosveldt 2006). There are five overarching constructs in our survey: (1) convenience, (2) service quality, (3) quality of food and drink, (4) price and value, and (5) ambiance. The remainder of this section describes how the sample, instrument, data collection and analysis were conducted.

We selected the 5 cafes in the university campus that had the most available reviews on blog and review sites. We obtained online reviews for the 5 cafés from 4 popular customer review sites, Zomato, Yelp, Four Square and Trip Advisor (Sadhu et al. 2016). The target population was customers of the 5 cafes- i.e., university students. Our sample was students from the Information Systems and Operation Management (ISOM) Department and Economics Department. We were limited to only two departments due to conditions imposed by our ethics committee. As an incentive to participate, respondents were entered into lucky draw worth 20 dollars. There were approximately 5700 undergraduates and 120 post graduates in both departments. An invitation to participate in our study was disseminated through the university student learning management system. We obtained a total of 275 responses to the invitation, of which 163 respondents were female and 112 were male. 260 respondents were undergraduates and 15 respondents were post graduates.

The item bank was developed as follows. First, we synthesized existing café satisfaction surveys (Gulluce Caglar et al. 2014; Hwang and Zhao 2010; Kim et al. 2005; Liang and Zhang 2009; Pizam and Ellis 1999; Pratten 2004; Ryu and (Shawn) Jang 2008; Shanka and Taylor 2005). In addition, the first author trawled Internet café forums to identify common complaints. New items were developed based on those complaints. Here, principles from grounded theory (Strauss and Corbin 1994) specifically, axial coding, guided us. Hence, approximately 400 items were collected. Items across the surveys and from the forums were then compared and duplicates were discarded. Fewer than 300 items remained after this step and two independent raters blinded to the study's purpose went through the items and marked items which were either vague or repetitive. Approximately 60 items were dropped. Next, we rearranged and reorganized the questions into a hierarchy. We assessed the instrument for construct and content validity and items which did not "fit in" the hierarchy were dropped, leaving only 175 items. The construct validity method used has been published elsewhere.

4.1 Data Collection

The data was collected in the following manner. First, respondents were presented a consent form and agreed with it. They then filled out some demographic details. They next chose one of the 5 mentioned cafés they wished to assess. Out of 306 respondents, 31 quit without providing a reason. To test for non-response bias, we conducted a wave analysis (Lankford et al. 1995; Rogelberg and Stanton 2007) and collected our sample in two rounds. In the first round, 170 responses were collected. In the second, 105 responses were collected. A non-response bias test was performed by running an independent sample t-test on each item against the two waves. We found the mean scores were not significantly different on all 175 items and thus there was no evidence for non-response bias.

Reviews from review sites were transformed into "opinion sentences." An opinion sentence "contains one or more product features and one or more opinion words" (Hu and Liu 2004, p. 172). First, we divided each online review into opinion sentences using the methodology of Hu and Liu, which is as follows. For each review, we identify a product/service feature and identify the associated opinion word(s) (such as an adjective). If there was enough context for it to have meaning on its own, then we accepted it as an opinion sentence. If there was not enough context than we combined two or more sentences and assessed comprehensibility. The average opinion sentence length was 14 words, which

we deemed reasonable as a sentence length is normally between 8-20 words (Smith 1961). We obtained a total of 441 opinion sentences.

Second, we employed two independent raters blinded to the purpose of the study to identify each opinion sentence as either negative or positive. Since the aim of the survey is to explore dissatisfaction, positive opinion sentences were discarded. Rater 1 recorded 244 negative opinion sentences while rater 2 recorded 235 negative opinion sentences. The inter-rater reliability was significant, and kappa was 0.833, above the recommended threshold of 0.7 (Landis and Koch 1977). All negative opinion sentences that raters disagreed on were dropped, leaving 225 negative opinion sentences. Café 1-5 had a final total of 66, 67, 12, 67, and 15 negative opinion sentences respectively.

Third, the raters independently mapped each negative opinion sentence to every construct in the CAS. The raters first started with the five top-level constructs (service, convenience, ambiance, food and drink quality, and price and value) and mapped each negative opinion sentence to one or more of the constructs, as seen in Table 1. We allowed raters to find no mappings, but there was no instance of such in our study. As an example, the comment "THE MOST OVER PRICED COFFEE I have ever had." was recorded in "Price and Value" by both raters. The ratings were assessed for inter-rated reliability. The level of agreement among the two independent raters for each top-level construct (service, convenience, ambiance, food and drink quality, and price and value) was calculated using Kappa, and had the following results, 0.820, 1, 0.747, 0.686, and 0.728. The agreement for the construct "convenience," was 1, because there were no negative ratings assigned by either rater to the construct. For the purposes of this study, the assignments by the two raters were analysed separately- no attempt was made to reconcile different categorizations by the raters.

Once the raters had completed the top-level mapping, they continued to map the lower levels, one level at a time. Each parent construct has at least two child constructs. For example, our ambiance construct has 38 children which are arranged in three levels, while "price and value" has 18 children of which three are located in the second level. Again, raters mapped every negative opinion sentence to every construct. Each mapping exercise was restricted to only the direct children of one construct. Thus, when raters mapped the children of ambiance, they did not consider the children for price. We encountered sample size problems doing this as in both CAS and online reviews, the volume of data decreased from higher levels to lower levels. Hence, the number of respondents and online reviews per category decreases as we move to lower levels. Hence, we could only fully assess the top-level constructs for all cafés, second level of "food and drinks quality" and only one branch of "price and value" for cafés 1,2, and 4.

4.2 Analysis

At this point in our methodology, we have (1) Likert scale scores of every item selected by respondents which are from 1 to 5, (2) frequency of responses to the items, and (3) frequency of opinion sentences that mapped to each item. We analysed the data in the following way as shown in Table 1. First, for each respondent, for every level, we gave the construct(s) with the lowest score a count of 1 and the rest a zero. This mimics the CAS process, which only is concerned with the construct a respondent is least satisfied with. As an example, if in one level a respondent gave five constructs the scores 2, 2,3, 4 and 5, then we would count the constructs as, 1, 1, 0, 0 and 0. Second, for every level, for the child constructs of the same parents, we calculated the sum of the counts. We continued this for each construct.

Café	Construct	Service	Convenience	Ambiance	Food &Drinks	Price & Value
Café 1	CAS	5	1	16	20	45
	Rater 1	13	0	3	24	36
	Rater 2	10	0	3	18	37
Café 2	CAS	8	15	10	15	40
	Rater 1	7	0	9	28	35
	Rater 2	7	0	8	25	35
Café 3	CAS	8	7	33	13	39
	Rater 1	2	0	2	7	5
	Rater 2	1	0	2	7	4
Café 4	CAS	11	4	16	20	38
	Rater 1	15	1	12	26	16
	Rater 2	14	1	11	24	18
Café 5	CAS	5	5	19	10	22
	Rater 1	1	0	3	6	4
	Rater 2	1	0	3	6	6

Table 1. Frequency of Low Scores for CAS and Negative Opinion Sentences for Top-Level Constructs

Third, we wanted to assess whether the frequency of responses to the items from the café satisfaction CAS for every level and the frequency of opinion sentences which mapped to each item come from a common distribution. Traditionally, the comparison of frequencies is done with goodness-of-fit tests. However, such tests examine whether two distributions are identical, not whether the distributions are "similar." Most goodness-of-fit tests (e.g., chi-square-based) do not work well with our data, because they require a large number of categories (Cheung and Rensvold 2002). Our CAS has five top-level constructs, which produces a low degree of freedom (df=4). The number of subcategories for the lower levels range from two to nine. Thus, to test whether the two samples come from a common distribution, we employ a Kolmogorov-Smirnov two sample test (known as a D-statistic) (Massey 1951; Pettitt and Stephens 1977; Pratt and Gibbons 1981). The null hypothesis regarding the distributional form is rejected (i.e., the two distributions are not identical) if the D-statistic, is greater than a critical value. Thus, our hypothesis is supported if the p-value is greater than 0.05 (i.e., the null hypothesis is not rejected). The two sample K-S test is ideal for our case, because it is robust to a smaller numbers of categories (even those as small as two) (Klotz 1967; Pratt and Gibbons 1981). Also, Kolmogorov-Smirnov two sample tests are robust to very small sample sizes (even as small as 10) (Birnbaum 1952; Massey 1951; Pettitt and Stephens 1977). We allow a Kolmogorov-Smirnov two sample test to be performed on the frequency of responses from top-level constructs, second-level constructs of "food and drinks quality", and third-level constructs of "price and value" which are child constructs of the second-level parent construct "value". The second-level parent construct "value" was the only branch with mapped negative opinion sentences. In addition, we recognize that failing to reject the null hypothesis is an unusual statistical approach. However, it is commonly employed for goodness-of-fit testing (Cochran 1952).

Given that practices for goodness of fit testing deviate from traditional inferential statistics, we also attempted to assess the degree of correspondence between the two distributions with a more traditional measure. Hence, as a second test of correspondence, we performed Pearson Correlation tests between our constructs and the negative opinion sentences (Cohen 1988; Kline 1998). Pearson's r measures the strength and direction (decreasing or increasing, depending on the sign) of a linear relationship between two variables. A high, positive r would suggest the distributions are similar. These tests were only performed when the number of opinion sentences for each level of the hierarchy was at least 5 x the number of constructs in the hierarchy. For example, if a parent construct had 3 children, then there had to be at least 15 opinion sentences mapped to the children for us to perform the analysis. 5x the

number of cells is a commonly accepted guideline for the minimum sample size used for many statistical analyses (Bentler and Chou 1987; Straub and Gefen 2004; Thomas and Watson 2002).

It should be noted that a Pearson r test of the data suffers from its own limitations. Notably, a Pearson's r test is traditionally used to measure correspondence between two interval values. In our case, we are using it to compare two aggregated groups of frequencies. The key difference is that significance in the traditional Pearson r test is calculated based on the sample size. In our case, statistical significance is calculated based on the number of categories, and hence is independent of the sample size. This makes a calculation of statistical significance meaningless. Essentially, we recognize there are limitations in the statistical tests we employ for this study, but would highlight there do not appear to be better alternatives.

5 Results

In regards to our hypothesis for top-level parent constructs (H1), we expected that our café satisfaction CAS would have a high degree of correspondence. As demonstrated in Table 2, the D-statistic is lower than the D-critical for all 5 cafes, and hence the p-values are all above 0.05. Hence, the sample from CAS and the sample from the online reviews appear to come from a common distribution.

Café	Rater	Alpha	D-statistic	D-critical	*p* for K-S	Sample size for CAS	Sample size for Reviews	r	df	*p* for r
café 1	R1	0.05	0.139	0.217	0.430	87	66	0.862	4	0.060
	R2	0.05	0.121	0.227	0.672	87	56	0.919	4	0.0275
Café 2	R 1	0.05	0.155	0.216	0.297	88	66	0.919	4	0.0275
	R 2	0.05	0.149	0.220	0.371	88	62	0.734	4	0.792
Café 3	R1	0.05	0.172	0.354	0.775	90	16	---	-	---
	R2	0.05	0.208	0.374	0.618	90	14	---	-	---
Café 4	R 1	0.05	0.198	0.212	0.079	89	70	0.53	4	0.358
	R 2	0.05	0.162	0.214	0.239	89	68	0.657	4	0.228
café 5	R 1	0.05	0.189	0.385	0.762	61	14	---	-	---
	R2	0.05	0.225	0.366	0.486	61	16	---	-	---

Table 2. Degree of Correspondence for the top-level constructs

Our Pearson correlation tests produce similar results as demonstrated in Table 3. Cafés 3 and 5 each had less than 25 negative opinion statements (5x5) mapped to the top-level constructs, hence they had an insufficient sample size and were omitted. Observe how in the table, the Pearson r produces very high correlations (all r> 0.6). Cohen (1988) notes that an r >0.5 is a strong correlation. This suggest support for our hypothesis of correspondence between top-level constructs and online reviews for cafes 1, 2 and 4. Note the statistic r is a measure of effect size, and is sample-size neutral. The statistic p for the Pearson r is not significant, because the p-value is calculated based on the number of categories, not on the sample size. The degrees of freedom is 4, because there are 5 categories. The Pearson r test of significance does not take into account that we really had 53 respondents for each café.

Café	Construct	Rater	Sample size for CAS	Sample size for Reviews	D-statistical	D-critical	*p* for K-S	r	df
Café 1	" food & drinks quality"	R 1	22	20	0.423	0.4	0.033	0.502	2
		R 2	20	20	0.4	0.409	0.059	0.458	2
Café 2	"food & drinks quality"	R 1	15	26	0.259	0.419	0.508	0.769	2
		R 2	15	24	0.267	0.425	0.462	0.712	2
Café 4	"food & drinks quality"	R 1	20	34	0.121	0.367	0.988	0.923	2
		R 2	20	31	0.102	0.373	0.999	0.953	2
Café 1	"value"	R 1	87	40	0.257	0.26	0.054	0.457	2
		R 2	87	29	0.494	0.282	0.001	0.526	2
Café 2	"value"	R 1	84	38	0.26	0.258	0.040	0.324	2
		R 2	84	26	0.5	0.3	0.001	0.363	2
Café 4	"value"	R 1	71	26	0.263	0.302	0.119	0.662	2
		R 2	71	26	0.302	0.303	0.05	0.694	2

Table 3. Degree of Correspondence for "food and drinks quality" and "value"

In regards to our hypothesis for lower level parent constructs, we expected that our café satisfaction CAS would have a lower degree of correspondence than the higher level parent constructs. For the overarching construct "food and drinks quality" as seen in Table 3 above, for cafes, 2 and 4 the K-S p value results suggest almost all the CAS constructs and online reviews come from the same distribution except for rater 1 in café 1. While, for cafés 2 and 4, results suggest a high correlation between CAS and online reviews, for café 1 Pearson r is just moderate. This suggests that the construct "food and drinks quality" and its attendant subconstructs capture the essence of dissatisfaction associated with food and drink quality.

For the overarching construct "price and value", a K-S two sample test was not suitable as only one of the three second-level constructs (i.e., value) had data from online reviews. We performed a K-S two sample test and Pearson Correlation on the subconstructs of value. As Table 3 demonstrates, the results are mixed for café 1, not supported for café 2, and supported for café 4.

6 Discussion and Conclusion

Our results demonstrate that our top-level constructs from CAS have a high degree of correspondence with online reviews. However, we could not adequately validate the lower level constructs due to a lack of negative opinion sentences that mapped to those constructs. There are several reasons for this. First, some online reviews raised problems only at the level of the top-level constructs. For example, a reviewer might say, "I really don't like their coffee or any of their desserts." Notice how the reviewer complains about the poor quality of the food and drink but does not break down why he or she is dissatisfied with the food and drink. Second, some online reviews only fully cover a single complaint about a café in depth. As an example, a reviewer might say "The cheesecake tasted great the first few bites but towards the end the layer of milk chocolate and the rich cheesecake becomes overwhelmingly sweet and rich". In this quote, the reviewer has only complained in depth about the cheesecake, and nothing else. It is possible the reviewer is dissatisfied with other elements of the café, but chooses not to talk about it.

Why is this? Most online reviews are posted on the company or third party website. These environments often limit the number of words in the review. As an example the average word count of Amazon reviews is 181.5 words (Wu and Huberman 2008). That word count impacts review quality is well documented, as longer reviews generally increase the helpfulness of the review (Mudambi and Schuff 2010). Longer reviews are found to be more useful. As an example, the usefulness of reviews are ex-

pected to increase by 19.9% when the reviewers write 83 words more than the average length of reviews (135 words) (López and Farzan 2014). Simply put, the average online review is of low quality. This suggests that CAS is actually better than an analysis of online reviews for diagnosing problems with café satisfaction. However, as we are not able to fully assess the conclusion validity of CAS, future work to develop techniques for assessing CAS conclusion validity is required.

References

Babcock, B., and Weiss, D. J. (2009). "Termination Criteria in Computerized Adaptive Tests : Variable-Length CATs Are Not Biased," in *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Barreda, A., and Bilgihan, A. (2013). "An Analysis of User-Generated Content for Hotel Experiences". *Journal of Hospitality and Tourism Technology*, 4(3), 263–280.

Bentler, P. M., and Chou, C.-P. (1987). "Practical Issues in Structural Modeling," *Sociological Methods & Research*, 16(1), 78–117.

Berdie, D. R. (1989). "Reassessing the Value of High Response Rates to Mail Surveys," *Marketing Research*, 1(September), 52–65.

Birnbaum, Z. W. (1952). "Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size," *Journal of the American Statistical Association*, 47(259), 425–441.

Cambria, E., Schuller, B. B., Xia, Y., and Havasi, C. (2013). "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intelligent Systems*, 28(2), 15–21.

Cheung, G. W., and Rensvold, R. B. (2002). "Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance," *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233–255.

Cochran, W. G. (1952). "The χ^2 Test of Goodness of Fit," *The Annals of Mathematical Statistics*, 23(3), 315–345.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Deutskens, E., de Ruyter, K., Wetzels, M., and Oosterveld, P. (2004). "Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study," *Marketing Letters*, 15(1), 21–36.

Embretson, S. E., and Reise, S. P. (2000). "Item Response Theory for Psychologists," *Quality of Life Research*, 4(3), 1–371.

Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J.,and Sadeh, N. (2013). "Why People Hate your App". In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (1276-1284).

Fundin, A., and Elg, M. (2010). "Continuous Learning Using Dissatisfaction Feedback in New Product Development Contexts," *International Journal of Quality & Reliability Management*, 27(8), 860–877.

Galesic, M., and Bosnjak, M. (2009). "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey," *Public Opinion Quarterly*, 73(2), 349–360.

Goodman, J. A., Broetzmann, S. M., and Adamson, C. (1992). "Ineffective - That's the Problem with Customer Satisfaction Surveys," *Quality Progress*, 25(5), 35–38.

Gräbner, D., Zanker, M., Fliedl, G., and Fuchs, M. (2012). "Classification of Customer Reviews Based on Sentiment Analysis," in *19th Conference on Information and Communication Technologies in Tourism (ENTER), Springer, Helsingborg, Sweden, 2012*.

Grégoire, Y., Salle, A., and Tripp, T. M. (2014). "Managing Social Media Crises with Your Customers: The Good, the Bad, and the Ugly," *Business Horizons*, 58(2), 173–182.

Grigoroudis, E., and Siskos, Y. 2010. *Customer Satisfaction Evaluation: Methods for Measuring and Implementing Service Quality*, (F. S. Hillier, ed.), New York: Springer.

Groves, R. M. (2006). "Nonresponse Rates and Nonresponse Bias in Household Surveys," *Public Opinion Quarterly*, 70(5), 646–675.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2004). *Survey Methodology Statistics*, Wiley Series in Survey Methodology, (J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, B.-E. Pennell, and T. W. Smith, eds.), (Vol. 2nd), Wiley-Interscience.

Gulluce Caglar, A., Saglik, E., OZHAN, Ç. K., and Kaya, U. (2014). "Service Quality and Customer Satisfaction Relationship: A Research in Erzurum Ataturk University Refectory," *American International Journal of Contemporary Research*, 4(1), 100–117.

Hayes, B. E. (1992). *Measuring Customer Satisfaction*, Milwaukee, WI: ASQC Quality Press.

Heerwegh, D., and Loosveldt, G. (2006). "An Experimental Study on the Effects of Personalization, Survey Length Statements, Progress Indicators, and Survey Sponsor Logos in Web Surveys," *Journal of Official Statistics*, 22(2), 191–210.

Ho, T.-H., and Dodd, B. G. (2012). "Item Selection and Ability Estimation Procedures for a Mixed-format Adaptive Test," *Applied Measurement in Education*, 25(4), 305–326.

Hol, a. M., Vorst, H. C. M., and Mellenbergh, G. J. (2008). "Computerized Adaptive Testing of Personality Traits," *Journal of Psychology*, 216(1), 12–21.

Hu, M., and Liu, B. (2004). "Mining and Summarizing Customer Reviews," *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168-177.

Hwang, J., and Zhao, J. (2010). "Factors Influencing Customer Satisfaction or Dissatisfaction in Restaurant Business Using Answertree Methodology," *Journal of Quality Assurance in Hospitality & Tourism*, 11(2), 93–110.

Jindal, N., and Liu, B. (2007). "Analyzing and Detecting Review Spam," *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 547–552.

Kim, Ay. S., Moreo, P. J., and Yeh, R. J. M. (2005). "Customers' Satisfaction Factors Regarding University Food Court Service," *Journal of Foodservice Business Research*, 7(4), 97–110.

Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling*, (Third.), New York: The Guilford Press.

Klotz, J. (1967). "Asymptotic Efficiency of the Two Sample Kolmogorov-Smirnov T," *Journal of the American Statistical Association*, 62(319), 932–938.

Landis, J. R., and Koch, G. G. (1977). "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, 33(1), 159–174.

Lankford, S. V, Buxton, B. P., Hetzler, R., and Little, J. R. (1995). "Response Bias and Wave Analysis of Mailed Questionnaires in Tourism Impact Assessments," *Journal of Travel Research*, 33(4), 8–13.

Lee, J., Park, D., and Han, I. (2008). "The Effect of Negative Online Consumer Reviews on Product Attitude: An Information Processing View," *Electronic Commerce Research and Applications*, 7(3), Elsevier B.V., 341–352.

Liang, X., and Zhang, S. (2009). "Investigation of Customer Satisfaction in Student Food Service An Example of Student Cafeteria in NHH," *International Journal of Quality and Service Sciences*, 1(1), 113–124.

Liu, B. (2012). "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, 5(1), 1–167.

López, C., and Farzan, R. (2014). "Analysis of Local Online Review Systems as Digital Word-of-Mouth," in *Proceedings of the 23rd International Conference on World Wide Web*, 457–462.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems Applied Psychological Measurement*, (Vol. 5), Hillsdale, N.J.: L. Erlbaum Associates.

Massey, F. J. J. (1951). "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, 46(253), 68–78.

Merrell, C., and Tymms, P. (2007). "Identifying Reading Problems with Computer-Adaptive Assessments," *Journal of Computer Assisted Learning*, (23), 27–35.

Mudambi, S. M., and Schuff, D. (2010). "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com," *MIS Quarterly*, 34(1), 185–200.

Peterson, R. A., and Wilson, W. R. (1992). "Measuring Customer Satisfaction: Fact and Artifact," *Journal of the Academy of Marketing Science*, 20(1), 61–71.

Pettitt, A. N., and Stephens, M. A. (1977). "The Kolmogorov-Smirnov Goodness-of-Fit Statistic with Discrete and Grouped Data," *Technometrics*, 19(2), 205–210.

Pizam, A., and Ellis, T. (1999). "Customer Satisfaction and its Measurement in Hospitality Enterprises," *International Journal of Contemporary Hospitality Management*, 11(7), 326–339.

Porter, S. R., Whitcomb, M. E., and Weitzer, W. H. (2004). "Multiple Surveys of Students and Survey Fatigue," *New Directions for Institutional Research*, 2004(121), 63–73.

Pratt, J. W., and Gibbons, J. D. (1981). "Kolmogorov-Smirnov Two-Sample Tests," in *Concepts of Nonparametric Theory*, New York: Springer, 318–344.

Pratten, J. D. (2004). "Customer Satisfaction and Waiting Staff," *International Journal of Contemporary Hospitality Management*, 16(6), 385–388.

Rogelberg, S. G., and Stanton, J. M. (2007). "Introduction: Understanding and Dealing with Organizational Survey Nonresponse," *Organizational Research Methods*, 10(2), 195–209.

Ryu, K., and Jang, S. (2008). "DINESCAPE: A Scale for Customers' Perception of Dining Environments," *Journal of Foodservice Business Research*, 11(1), 2–22.

Sabbaghan, S., Gardner, L., and Chua, C. E. H. (2016). "A Q-Sorting Methodology for Computer-Adaptive Surveys," *ECIS 2016 Proceedings*.

Sabbaghan, S., Gardner, L., and Chua, C. E. H. (2017a). "A Threshold for a Q-Sorting Methodology for Computer-Adaptive Surveys," In *ECIS 2017 Proceedings*.

Sabbaghan, S., Gardner, L., and Chua, C.E.H., (2017b). "Computer-Adaptive Surveys (CAS) as a Means of Answering Questions of Why," *In PACIS 2017 Proceedings.*

Sadhu, D., John, S., Kulkarni, R., and Tiwari, A. (2016). "Mining and Predicting Reviews to Micro-reviews and Detection of Manipulated Reviews for Hotel Websites," *International Journal of Emerging Technology and Computer Science*, 1(2), 28–31.

Sampson, S. E. (1998). "Gathering Customer Feedback via the Internet: Instruments and Prospects," *Industrial Management & Data Systems*, 98(2), 71–82.

Schläfke, M., Silvi, R., and Möller, K. (2012). "A Framework for Business Analytics in Performance Management," *International Journal of Productivity and Performance Management*, *62*(1), 110–122.

Shanka, T., and Taylor, R. (2005). "Assessment of University Campus Café Service: The Students' Perceptions," *Asia Pacific Journal of Tourism Research*, 10(March 2015), 329–340.

Shin, C. D., Chien, Y., and Way, W. D. (2012). "A Comparison of Three Content Balancing Methods for Fixed and Variable Length Computerized Adaptive Tests.,"

Smith, E. A. (1961). "Devereux Readability Index," *The Journal of Educational Research*, 54(8), 298–303.

Somprasertsri, G., and Lalitrojwong, P. (2010). "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization," *Journal of Universal Computer Science*, 16(6), 938–955.

Straub, D., and Gefen, D. (2004). "Validation Guidelines for IS Positivist Research," *Communications of the Association for Information Systems*, (13), 380–427.

Strauss, A., and Corbin, J. (1994). "Grounded Theory Methodology," *Handbook of Qualitative Research*, 273–285.

Stricker, L. J., Wilder, G. Z., and Bridgeman, B. (2006). "Test Takers' Attitudes and Beliefs About the Graduate Management Admission Test," *International Journal of Testing*, 6(3), 255–268.

Thomas, D. M., and Watson, R. T. (2002). "Q-Sorting and MIS Research: A Primer," *Communications of the Association for Information Systems*, (8), 141–156.

Thompson, N., and Weiss, D. (2011). "A Framework for the Development of Computerized Adaptive Tests," *Practical Assessment, Research & Education*, 16(1), 1–9.

Thorpe, G. L., and Favia, A. (2012). "Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications," *Psychology Faculty Scholarship*, (20).

Wisner, J. D., and Corney, W. J. (2001). "Comparing Practices for Capturing Bank Customer Feedback-Internet versus Traditional Banking," *Benchmarking: An International Journal*, 8(3), 240–250.

Wu, F., and Huberman, B. A. (2008). "How Public Opinion Forms," in *In International Workshop on Internet and Network Economics*, 334–341.

Xu, K., Liao, S. S., Li, J., and Song, Y. (2011). "Mining Comparative Opinions from Customer Reviews for Competitive Intelligence," *Decision Support Systems*, 50(4), 743–754.