Spam in User Generated Content Platforms: Developing the HaBuT Instrument to Measure User Experience

Sowmya Karunakaran[1] and Erik Brorson[2]

***Abstract*— We are in an era of user-generated content (UGC) but our understanding of the impact of UGC spam on user experience is limited. Most prior instruments to measure user experience were developed in the context of traditional spam types like web-spam or email-spam. In this paper, we develop the 15 item HaBuT scale, consisting of three sub-scales: Happiness, Burden and Trust that measures user experience with respect to UGC spam. The items in the instrument are analyzed using confirmatory factor analysis with a sample of 700 responses from internet users. This process resulted in an instrument of high reliability and validity. The instrument is a valuable tool for researchers and practitioners interested in designing, implementing, and managing systems that rely on user-generated content and to those studying the impact of UGC spam on user experience. We demonstrate a real-world application of the HaBuT scale by applying it to investigate the impact of review spam on mobile apps users. We present the study results of online experiments with 3300 participants across US, India and South Korea.**

I. INTRODUCTION

With the widespread use of user-generated content (UGC) in social media and content-based online services, spam in these sites is explosively increasing. UGC systems, often termed as Web 2.0 systems, promote online collaboration through an architecture for participation that encourages users to add value to web applications as they use them [1]. This attracts not just users but also spammers to post large volumes of content online. For example, Niu et al. found that more than half of blog posts on two of the blog sites blogspoint.com and blogstudio.com that they examined were spam [2].

UGC spam poses more challenges than traditional spam such as web spam or email spam. Firstly, unlike traditional spam, spam on UGC platforms can manifest in several forms such as fake reviews, like-baiting stories, solicitation via comments, posting irrelevant content and keyword stuffing. Secondly, UGC spammers exhibit unique non-textual patterns, such as posting activities, advertised spam link metrics, and spam hosting behaviours [3]. Thirdly, spammers use these techniques for various reasons. For example, they use keywords stuffing in order to increase their click-through rate. By inserting popular terms to the title and the content, spammers can make their posts and comments highly ranked/visible when a user searches these keywords in a UGC system [3]. Fourthly, UGC spam infiltrates legitimate websites by posting on them. Few examples of that include promotional comment in blogs, fake user profiles, unsolicited links in social bookmarking websites etc. [1]. Lastly, unlike traditional web/email spam, UGC spam not only pollutes the content contributed by normal users, resulting in bad user experiences, but also misleads users [3].

For these reasons, measuring user experience with respect to UGC spam is critical and traditional web-user experience methods do not suffice. Measurement metrics and the properties used to evaluate them vary by context and goal and therefore there are several strategies for developing and refining measurement instruments. In this paper, we focus on developing and validating an instrument to measure the impact of user generated content spam on latent constructs for user experience. The 15-item HaBuT instrument measures three constructs of user experience: Happiness, Burden and Trust. We describe the process of instrument development and demonstrate its application with an example. In the example application, we use the HaBuT instrument to identify which spam types impact user experience while reviewing mobile app reviews.

The HaBuT instrument is a useful tool to researchers and practitioners interested in designing, implementing, and managing web sites that rely on user-generated content. We argue that our work fills an important gap not addressed by existing research efforts on user experience and UGC Spam.

II. RELATED WORK

The goal of our literature review is two-fold. First, by reviewing related work we aim to summarize previous attempts to conceptualize the constructs and theories in which the construct may have proven useful as an independent or dependent variable. Second, given scale development and validation is a time-consuming and costly process, the literature review helped us determine if measures of the constructs already exist and if a fresh instrument is needed at all and thereby avoid the redundancy of developing another scale.

In this section, we discuss the need for studying the impact of UGC Spam on user experience by presenting prior research. Next, we outline work that has looked into survey instrument development in the context of online user experience.

A. UGC Spam: Taxonomy and impact on users

Prior researchers in the domain of UGC spam although primarily focus on spam detection mechanisms, few researchers have shown that UGC spam may lead to lower user engagement. Grier et al. studied UGC spam in Twitter

[1]Sowmya Karunakaran is a Researcher with Google Inc, Dublin , Ireland sowmyakaru@google.com

[2]Erik Brorson is a Master's student with Uppsala University, Sweden er.brorson@gmail.com

and found the click-through rate of Twitter spam is much higher than email spam [4]. According to Dasgupta et al. spam exposure, leads to both statistically and economically lower user engagement [5]. Fake reviews increase consumer uncertainty. The effects of reviews that are more positive and numerous are smaller on online retailing platforms that have fake product reviews [6]. Another study shows source credibility to have a strong influence on attitude but weak direct effect on behavioural intention [7]. Another study found that like-baiting stories are, on average, 15% less relevant than other stories with a comparable number of likes, comments and shares [8]. Research also shows that in the presence of spam, users will take longer time to finish a given task [9]. In summary, although there have been pockets of research that tie user experience on UGC systems, there is no universal instrument that holistically measures all elements of user experience. In this paper, we address this gap by building and testing the HaBuT instrument. In the following section, we review related work on instrument development in the broader domain of impact of web quality on user experience.

B. Online User Experience Measurement Instruments

Several researchers have developed instruments to measure various elements of web quality. Aladwani and Palvia developed an instrument that captures key characteristics of web site quality from the user's perspective [10]. Their 25-item instrument measures four dimensions of web quality: specific content, content quality, appearance and technical adequacy. Yang et al. developed a five-dimension service quality instrument involving: usability, usefulness of content, adequacy of information, accessibility, and interaction [11]. Both these instruments however do not focus on potential negative user experience that may arise from additional burden due to spam content.

Suh et al. developed and validated a measure of user burden in computing systems called the User Burden Scale (UBS), which is a 20-item scale with six individual subscales representing each of the burden constructs [12]. Based on their definition of burden constructs, UGC spam can be seen to cause annoyance and lead to time, social, mental and emotional burden. Novak et al show that content relevance corresponds to greater focused attention by online users [13]. Diakopoulos and Naaman explore impact of offensive content on the quality of discourse in online news comments [14].

Rodden et al. through their HEART framework for user-centered metrics for web applications, emphasized measuring "User Happiness" [15]. According to them, Happiness refers to metrics that are attitudinal in nature and relate to subjective aspects of user experience, like satisfaction, visual appeal, likelihood to recommend, and perceived ease of use. The peculiar nature of UGC also raises credibility concerns, highlighting the need to study impact on user trust as is evident from prior research. Shelat and Egger have shown that providing content that is appropriate and useful to the target audience is a strong cue to trustworthiness [16].

TABLE I
ATTRIBUTES FROM LITERATURE BASED ON THE THREE CONSTRUCTS

Dimension	Sample Items
Happiness	Satisfaction, met expectations, ideal for task, easy to complete task, time spent on task, mental effort to complete task, conciseness, uniqueness
Burden	Offensive, annoying, irrelevant, inappropriate
Trust	Reliability, brand trustworthiness, platform truthfulness, content truthfulness, safety, usefulness, helpfulness, platform spamminess

Wiedenbeck et al built an instrument to measure online trust of websites [17]. Ayeh et al. use trust to examine credibility perceptions and user attitude towards using UGC [7].

In summary, from prior studies we inferred that Burden, Happiness and Trust capture the user experience with respect to UGC spam. Although there have been several instruments and frameworks developed to measure user's online experience, capturing the impact of spam on user experience has not been studied holistically by combining Happiness, Burden and Trust dimensions. In this paper, we develop an instrument to address this gap.

III. PROCESS OF DEVELOPING AN INSTRUMENT

As UGC based systems continue to provide online users with greater assortment and abundance of information, the need for valid and reliable instruments to measure the success of these systems is increasingly important. We use the process described by Straub 1989 for creating and validating instruments in IS research, which includes content validity, construct validity and reliability analyses [18]. This process has been widely used and accepted by several IS researchers. We discuss this in detail in the next few sections.

A. Content Validity

Content validity, involves activities such as defining the construct of interest and generating a candidate list of items from the domain of all possible items representing the construct. This step involves delimiting the domain of the construct and generating sample items representing the concept under consideration. In order to ensure content validity, the instrument needs to draw representative items from a universal pool [19]. We conducted a systematic literature review through keyword searches on three bibliographic databases, guided by procedures described by [20]. This was followed by a brainstorming session among the authors and experts using the Delphi technique [21] leading to 24 attributes. Overall, our review of the academic literature identified three dimensions of user experience with respect to spam: happiness, burden, and trust; and yielded 42 representative items. The sample items were initially assessed using a Delphi method. Two spam fighting experts and one web-user education expert participated in the process. The evaluation led to shortlisting 24 attributes. Table I summarizes the initial shortlisted attributes and dimensions.

B. Construct Validity and Reliability

In the first stage of data collection, the 24-item instrument was administered to participants. The items were measured using a seven-point scale ranging from (1) Strongly Disagree to (7) Strongly Agree with a mix of positively and negatively framed questions. We used Amazon Mechanical Turk (MTurk) to recruit participants for the first stage of the study in July 2017. MTurk is a widely used crowdsourcing internet marketplace to solicit participants for social science research. Even though MTurk has limitations with respect to sampling [22] and generalizability [23], it has been shown to be effective for this line of research [12]. MTurk participants also perform better on online attention checks than do subject pool participants [24] and in addition were shown to be superior to the samples commonly utilized in face-to-face studies [25].

TABLE II
EFA LOADINGS (OBLIMIN ROTATION)

	Rotated Factor Pattern		
Item	Happiness	Trust	Burden
Satisfaction (H)	0.79		
Matched expectations (H)	0.82		
Ideal results (H)	0.87		
Easy to complete task (H)	0.65		
Uniqueness (H)	0.66		
Concise (H)	0.66		
Comprehensible (H)	0.66		
Helpfulness (H)	0.80		
Accuracy of information (H)	0.61		
Reliability (T)		0.71	
Platform truthfulness (T)		0.69	
Dependability (T)		0.78	
Content truthfulness (T)		0.72	
Brand trust (T)		0.86	
Safety (T)		0.74	
Offensive (B)			0.73
Annoying (B)			0.74
Irrelevant (B)			0.79
Inappropriate (B)			0.95
Proportion of variance	0.27	0.21	0.14
Cumulative variance	0.27	0.48	0.63

NOTE: The loadings below —0.2— have been suppressed

We posted a solicitation on the MTurk Human Intelligence Tasks list requesting US workers to participate in a task-based survey. Workers who responded to the survey were reimbursed USD 1.25. The task involved looking up a mock site that contained a set of business listings posted by users about restaurants in New York. 10% of these listings were spam listings, i.e. we had introduced keyword stuffing and duplicate listings. As part of the task, participants had to select restaurants from the list that met a specific criterion such as locality, rating and opening times. To ensure that the participant saw the spam, the task was designed in such a way that the participant had to encounter at least two spam listings in the process of accomplishing the task. For example, we asked participants to find the name of the restaurant on Metropolitan Avenue. There were two listings with the address Metropolitan Avenue. One of them being the genuine listing and the other a duplicate listing (spammers typically create such duplicate listings to tap in to the web traffic and clicks of confused users).

The task was followed by an experience survey. The survey took approximately 5 minutes to complete. At the end of the survey, participants could optionally provide feedback on the survey questions. We completed two studies with 75 responses each to test the survey questionnaire, leading to 150 responses in total. The participants were from 18 to 65 years of age. Approximately, 53% of the respondents were female and 47% male. In addition to gender and age we collected education and employment information and the split was as follows: High school 1%; High school graduate (includes equivalency) 10%; Some college, no degree 26%; Associate's degree 14%; Bachelor's degree 36%; Master's degree 11%; Ph.D. 2%; Others 1%; Employed Full-time 58% Employed Part-time 10% ; Self-employed 13%; Care-provider 0% ; Homemaker 6% ; Retired 4% ; Student - Undergraduate 2% ; Student - Masters 1% ; Student - Doctoral 0%.

In compliance with ethical training guidelines in our organization, we ensured that participants anonymity and privacy were respected. Firstly, all responses were anonymous, and we did not collect any personally identifiable information. Secondly, all participants received compensation for their time and effort. Thirdly, respondents had the option to exit the survey at any point and still would receive compensation for their time. Fourthly, respondents had the option to skip one or more of the demographic questions on gender, age, occupation and educational background.

We started by examining the dimensionality of the items. After performing a combination of Horns parallel analysis and Cattels scree test we found that three factors can adequately explain the variation in the data. We then performed an exploratory factor analysis for the data using polychoric correlation matrix. The three factors were extracted using the weighted least squares (WLS) fitting method. We removed items by reviewing the results of the factor analysis. Variables with communalities below 0.5 were removed one by one, and we found the final model fit is acceptable (Chisq = 50, p-val = 1, RMSEA = 0.05, [0,016; 0.062], TLI = 0.97). Table II shows the rotated factor scores.

In social sciences research, we generally expect some correlation among factors, since behaviour is rarely partitioned into units that function independently. Therefore, use of orthogonal rotation would result in a loss of valuable information if the factors are correlated, and use of oblique rotation is recommended [26]. The oblimin rotation of the factor-loading matrix that we chose permits the factors to be correlated with each other and perhaps gives a more reproducible solution. Table III gives the correlation matrix. We see a positive correlation between Happiness and Trust of 0.6 and a negative correlation between Burden and Happiness and Burden and Trust of -0.3. The negative correlation implies that lower burden value translates to higher happiness and trust.

After reduction, 19 items remained, with number of items

for subscales being 9, 4 and 6 for happiness, burden and trust respectively. We then computed internal validity coefficients of the scales using Cronbachs alpha. The alpha values for happiness, burden and trust are 0.89, 0.89 and 0.9. Next, we conducted a final review of the 19 items and the responses. We also reviewed the optional open-ended feedback responses provided by participants. Few participants had pointed out to us that reliable and dependable could have very similar meanings and to consider merging them into one item. Further participants attributed helpfulness and accuracy to traditional web quality and not to spam (which we later found to be discussed by [27]). After taking this feedback in to consideration, the final list comprised 15-items, which we further validate as described in the next section.

TABLE III
FACTOR CORRELATION MATRIX

	Happiness	Trust	Burden
Happiness	1		
Trust	0.6	1	
Burden	-0.3	-0.3	1

TABLE IV
ESTIMATED LOADINGS IN THE CFA MODEL.

Item	Dimension	Estimate
Satisfaction	Happiness	1
Matched expectations	Happiness	0.969
Ideal results	Happiness	0.993
Easy to complete task	Happiness	0.944
Uniqueness	Happiness	0.565
Conciseness	Happiness	0.875
Comprehensive	Happiness	0.882
Dependability	Trust	1
Safety	Trust	0.997
Brand trust	Trust	1.043
Content trust	Trust	1.408
Offensive	Burden	1
Annoying	Burden	1.155
Irrelevant	Burden	0.961
Inappropriate	Burden	1.110

NOTE: All estimates are significant with $p < 0.01$

C. Verification and Validation

We collected a sample of 700 respondents using the instrument developed. Similar to the previous step, respondents were asked to complete task-based survey via the Mechanical Turk platform. The task involved looking up a mock site that contained a set of listings posted by users about services in New York (movie theatres, security services, pest control services etc). The level of spam induced and task design was similar to the process described under the content validity and reliability section. After the respondents completed the tasks, we asked them to fill out a survey containing our items.

We used this data to fit a confirmatory factor model in order to validate our instrument. The confirmatory model fits the data well (Chisq = 324, p-value = .000: RMSEA = 0.057, [0.050, 0.063], p-value = .005, Santorra Bentler-RMSEA = 0.063, [0.06, 0.067]: TLI = 0.96). Even though the chi-square test (The null hypothesis is that the model implied covariance matrix and the empirical covariance matrix are identical, i.e we want a non-significant result) is significant which in general is true for large samples. The RMSEA, both the Satorra-Bentler adjusted and the unadjusted is below 0.08 and the TLI is close to 1. Table IV provides the summary of the CFA model. The model was fitted using the Diagonal Weighted Least Squares (DWLS) estimator to account for ordinal data. All of the items have significant loadings. The loading estimates together with their dimension and p-value are listed in Table IV. We observed a 89% inter-rater reliability using Cohen's Kappa. The final instrument is listed in Table V.

TABLE V
15-ITEM INSTRUMENT TO ASSESS UGC SPAM USER EXPERIENCE

Item Name
I was satisfied with the content
The shown content matched my expectations
The shown content was ideal for the tasks I was asked to complete
I found the tasks easy to complete
Each individual piece of content felt unique
The content was presented in a concise way
It was easy to understand the content
I believe the content on the platform helps me find what I am looking for
I trust the content on the platform
I feel safe using the content on the platform
I believe that the information I got from the content is true
I found some of the content offensive
I found some of the content annoying
I found some of the content irrelevant to my task
I found some of the content inappropriate

Each item was rated on a 7 pt agreement scale ranging from *Strongly Disagree* to *Strongly Agree*

IV. APPLICATION OF HABUT

In the previous sections, we developed the HaBuT instrument and validated it. In the following section, we demonstrate an application of the instrument to measure user experience with respect to spam on user generated reviews for apps in a mobile store.

A. Problem Background

Online reviews have become an integral part of consumer's evaluation of goods and services. Internet users rely on these reviews to make decisions and the usefulness of these reviews can play a big part in the consumer deciding to use or purchase a service or good online [28]. While researchers have focused on ways to combat bad reviews and built novel ways to combat bad reviews [29], there is no prior research that looks at which specific types of spam have the highest impact on user experience. Although automated methods enable detection of large volumes of spam, still a large fraction of manual reviews continue to happen and the algorithms need to be perfected. This becomes even more challenging with UGC content [29]. The effort required to

reduce the spam rates from 5% to 0% is much higher as compared to bringing it down from 10% to 5%. Although spam fighting teams are invested in reducing spam to 0%, it is an expensive process [30]. There is a business need for prioritization based on which spam types impact user experience negatively and invest time and resources towards building automated systems to tackle those. In summary, the questions that stand out are: For a given UGC platform which spam types impact user experience the most? Given diminishing marginal returns, should spam-detection efforts aim to reduce spam to 0% vs aim to reduce to a non-zero level that does not cause degradation in user experience?

We design an experiment and demonstrate the use of HaBuT instrument to answer these questions.

TABLE VI
EXPERIMENTAL GROUPS

Spam type	Control group 0%	Treatment groups 4%	8%
Gibberish		G4	G8
Irrelevant		I4	I8
Solicitation	C0	S4	S8
Offensive		O4	O8
Promotional		P4	P8

B. Study setup

We conduct online experiments to measure the impact of various spam types on user experience. Overall the study investigated five different types of spam that are commonly prevalent in user generated online reviews namely:

- Gibberish: e.g. *asdsad jksjfs sdhd*
- Irrelevant: e.g. *Review of a movie for a gaming app*
- Solicitation: e.g.*Follow me on twitter @xxxx*
- Abusive language: e.g.*idiotic dirty morons*
- Promotions: e.g.*Instant cash discount, register now*

Participants were asked to shortlist an online app for download after reading the reviews each app received. Figure 2 shows a snapshot of the simulated screen participants were shown for one of the tasks. Each participant completed four such tasks for different types of apps leading to each participant reviewing about 100 reviews in total for various apps. Each review approximately comprised one short sentence. Participants were randomly assigned to one of the groups shown in Table VI. Participants in the control group saw no spam across the reviews. Participants in the treatment group saw a certain percentage of spammy reviews and a specific type of spam. For example, participants in Group G8 saw eight Gibberish reviews in total, randomly spread across the 100 reviews and participants assigned to Group I4 saw four irrelevant reviews randomly spread across the 100 reviews. After completing the task, participants completed the HaBuT questionnaire. For each participant, we computed trust, burden and happiness scores based on their responses to the HaBuT questionnaire. We ran our tests across participants from US, India and South Korea with 1100 participants per country. All these three markets rank high in terms of mobile apps usage. The study was setup on Qualtrics survey platform and we used their panel services to source participants for the study. For each valid response the panel provider was paid USD 5. Similar to the steps described on study ethics in the previous section, the study was completely anonymous and we did not collect personally identifiable information.

C. Results

We ran statistical tests to determine which spam types had significant impact on user burden, trust and happiness. We compared the mean scores, across spam rates and spam types and tested for statistical significance (Kruskal-Wallis test; p = 0.05). We corrected for multiple testing, using Bonferroni correction (adj.p = 0.017).

We observe that the instrument was useful in determining the spam type that caused the most degradation in user experience. Figure 1 shows the plot of complement of normalized mean burden scores for the five UGC spam types we studied for each country and provides a visual comparison of change in burden across spam types, spam rates and countries. From the figure, we notice that abusive words (indicated by the blue marker on the chart) had the most impact on user burden. Burden increased steeply as rates of abusive words spam increased. Promotions spam was a close second leading to significant decline in burden across all three countries. Gibberish spam had the lowest impact on user burden and did not change significantly with increase in rate of spam. We also observed cross-country differences in user experience. For example, from Table VII we observe that none of the spam types had an impact on Trust and the only exception being Gibberish spam in South Korea.

The above experiments demonstrate that the HaBuT instrument could be applied in comparative studies and for benchmarking, for example to measure cross-country differences in user experience when encountering UGC spam, to measure differences in user experience across various types of UGC spam and to measure the UGC spam tolerance thresholds of various user segments.

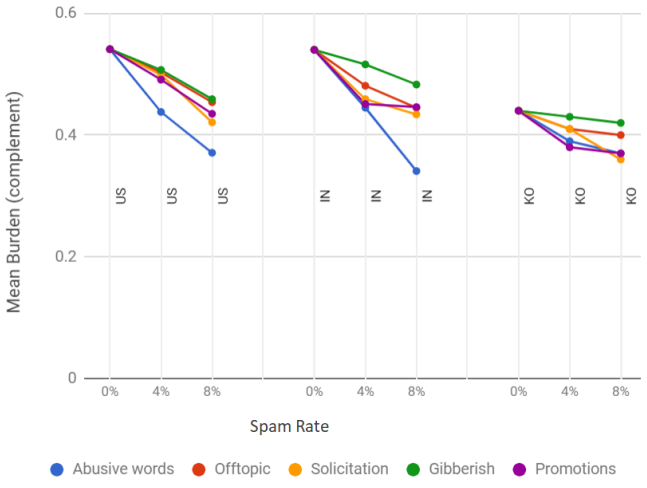

Fig. 1. Cross-country comparison of Burden scores

TABLE VII

COMPARISON OF NORMALIZED SCORES FOR MEAN HAPPINESS, BURDEN(COMPLEMENT) AND TRUST ACROSS SPAM TYPES AND SPAM RATES

Spam type	Spam rate (%)	India			South Korea			USA		
		Happiness	Burden	Trust	Happiness	Burden	Trust	Happiness	Burden	Trust
No spam	0	0.74	0.54	0.68	0.60	0.44	0.61	0.73	0.54	0.67
Abusive words	4	0.74	0.45	0.66	0.56	0.39	0.57	0.69	0.44	0.64
	8	0.72	0.34	0.67	0.57	0.37	0.59	0.71	0.37	0.65
	χ^2	0.45	39.81	0.60	3.20	8.74	2.99	1.82	29.83	2.14
	(p)	(0.797)	(<0.001*)	(0.740)	(0.202)	(0.013*)	(0.225)	(0.403)	(<0.001*)	(0.343)
Irrelevant	4	0.69	0.48	0.62	0.59	0.41	0.60	0.68	0.50	0.64
	8	0.73	0.45	0.67	0.59	0.40	0.58	0.69	0.45	0.66
	χ^2	3.57	12.32	4.21	0.94	1.65	1.14	4.72	11.36	3.40
	(p)	(0.168)	(0.002*)	(0.122)	(0.624)	(0.438)	(0.563)	(0.095)	(0.003*)	(0.182)
Solicitation	4	0.74	0.46	0.66	0.59	0.41	0.61	0.71	0.50	0.65
	8	0.71	0.43	0.65	0.63	0.36	0.62	0.70	0.42	0.67
	χ^2	0.121	14.17	1.76	4.32	9.22	0.62	1.16	19.04	1.13
	(p)	(0.547)	(0.001*)	(0.415)	(0.115)	(0.100)	(0.733)	(0.559)	(<0.001*)	(0.568)
Gibberish	4	0.74	0.52	0.66	0.60	0.43	0.58	0.70	0.51	0.64
	8	0.72	0.43	0.65	0.63	0.36	0.62	0.70	0.42	0.67
	χ^2	0.39	4.01	0.59	0.07	0.53	6.50	1.49	9.85	11.59
	(p)	(0.824)	(0.135)	(0.589)	(0.967)	(0.766)	(0.039*)	(0.476)	(0.007*)	(0.453)
Promotions	4	0.73	0.45	0.67	0.61	0.38	0.62	0.71	0.49	0.66
	8	0.74	0.45	0.66	0.61	0.37	0.61	0.70	0.44	0.64
	χ^2	0.06	13.13	0.47	0.38	8.64	0.87	1.77	19.20	1.63
	(p)	(0.970)	(0.001*)	(0.791)	(0.829)	(0.013*)	(0.874)	(0.412)	(<0.001*)	(0.443)

Chi-sq and p-values corresponding to Kruskal-Wallis test; * indicates significant at 0.05 level ; adj p= 0.017; N = 3300 participants.

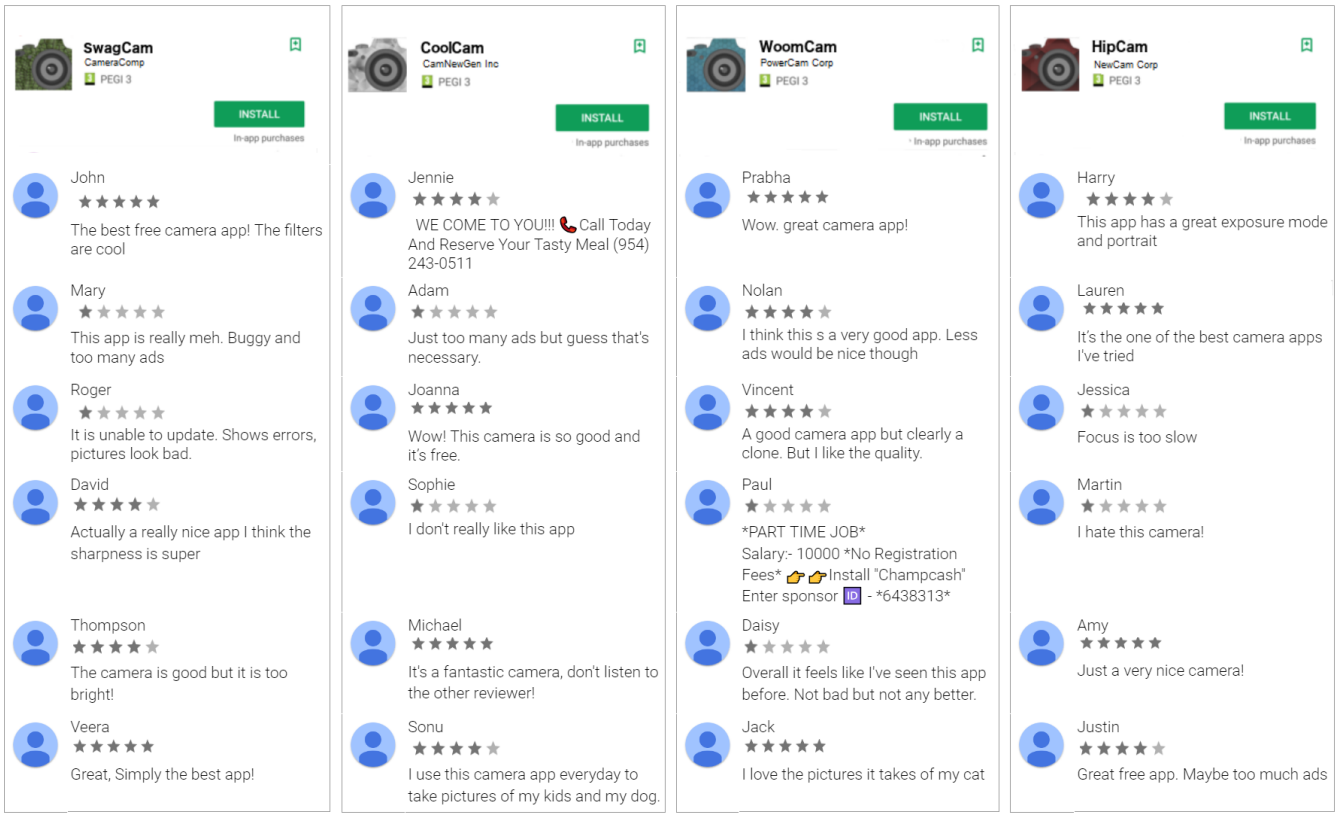

Fig. 2. Screenshot of a task where participants saw user generated reviews for four Camera Apps

V. DISCUSSION

In this section, we provide guidelines for administering the instrument and general usage guidelines. We also discuss its limitations and future directions.

A. Guidelines for Use

The instrument has 15 questions, with each question evaluated on a 7 point scale (1 = Strongly disagree to 7 = Strongly Agree) leading to a potential maximum score of 105. Note that Burden has a negative correlation with Happiness and Trust. Hence, while computing the overall user experience scores, researchers should use the complement of Burden scores. We recommend that survey respondents be required to answer every question, to allow comparison of scores across users. Researchers may also choose to administer only the sub-scales relevant to their system since each sub-scale on our instrument had a good alpha value.In addition to using the HaBuT scale based self-reported measures, where possible, researchers should consider including behavioural metrics such as task completion times or task success.

By understanding users experiences in the context of UGC spam, UGC platform designers can add provisions for users to intuitively report and provide feedback on issues that affect their experience most negatively. While it is good to improve the overall user experience by dealing with all types of UGC spam on their platform, the cost of spam management (manual reviews, developing ML based spam detection,) might be quite high [9]. We believe that the instrument can be useful in helping UGC product managers to determine different types of trade-offs in their feature list (e.g., options for flagging, reporting, providing elaborate feedback).

The internet currently hosts thousands of UGC platforms, this trend is only likely to grow in the future, and the scales might be used to assess the quality of a given UGC platform. By using this instrument, user experience researchers and online abuse experts will be able to determine the spam types for which user experience deteriorates rapidly even with low levels of spam content. Researchers may also use the instrument to determine the point of inflection (i.e. the point at which the user experience deteriorates significantly) of user experience by administering this instrument to groups of users who had different levels of exposure to spam. This evaluation may provide a fast and early feedback to firms that are looking to invest in spam fighting efforts.

B. Limitations

This study is only a first step toward understanding the influence of UGC spam on user experience. Other aspects of validity still need to be tested. For one, we did not conduct a test-retest validation due to the difficulty of following up with online, anonymous participants. There are also limitations with numerical scales and subjective measures. Our testing and validation steps improved the internal validity, and use of several groups of subjects improved the external validity and generalizability of the instrument. Instruments, however, are always subject to further improvement. Although user experience with respect to spam content is inherently subjective, researchers can use the scale in combination with other more objective tests such as task completion rates and task success to gain an overall complete picture of user experience. Finally, although the user population on Mechanical Turk, which we used for our tests of validity, is relatively diverse for a US Internet sample, the reviews for MTurk as a survey platform are mixed. For example, the respondents are considered WEIRD (Western, educated, industrialized, rich, and democratic) [22]. Another evaluation shows that the generalizability of MTurk findings to a broader audience might be limited due to issues such as feature comprehension [23]. Hence, it is necessary to test the instrument across different populations and different cultures. We address this partly by testing it with users from three culturally countries as described in Section 4.

C. Future Research Directions

In addition to the tests described in the previous sections, as part of future research, we plan to conduct testing with participants across varying demographics, especially relating to age, gender, education level, cultural background, and technological expertise. This will ensure that the scale is widely applicable across all populations. Future research can also evaluate the user experience across different task types. Our study used tasks that required participants to find specific items (for example,"you want to find a reputable security company near your new home on Lafayette Street). Future studies could examine the instrument with exploratory search tasks. While the current study utilized UGC platform with an international and informational content focus, UGC based applications with a domestic and transaction focus should also be examined with the instrument.

VI. CONCLUSION

As UGC in social media continues to rise, the spam in these platforms is also prone to rise. This unmanageable rise in UGC content allows spammers to show spam content at striking positions such as at the front page on the UGC site, top listing of a comment stream or as a top listing on a map leading to bad user experience. The unique characteristics of UGC Spam and the lack of an accepted instrument for measuring user satisfaction led us to this research. Past research on UGC spam primarily focused on spam detection techniques. In this paper, we describe the design and validation of a scale for measuring the impact of spam on user experience in UGC based systems. The results of our two-stage approach to instrument development led us to three dimensions of user experience with respect to UGC spam: Happiness, Burden and Trust and an instrument of high reliability. The instrument is valuable to researchers and practitioners interested in designing, implementing, and managing online systems that rely on user-generated content. We believe the paper makes three major contributions. Firstly, it inspires the field to not just be mindful of spam and the potential of trolling when designing for UGC platforms, but to be proactive in identifying and tracking the impact on user experience. Secondly, the paper articulates the process

of creating, validating and applying an instrument, demonstrating that this methodology yields far more reliable and valid findings from self-report measures compared to ad-hoc questionnaires. Thirdly, the paper provides a practical application of the HaBuT instrument by demonstrating its use to measure cross-country user experience.

REFERENCES

[1] P. Hayati, V. Potdar, A. Talevski, N. Firoozeh, S. Sarenche, and E. A. Yeganeh, "Definition of spam 2.0: New spamming boom," in *Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on*, pp. 580–584, IEEE, 2010.

[2] Y. Niu, H. Chen, F. Hsu, Y.-M. Wang, and M. Ma, "A quantitative study of forum spamming using context-based analysis.," in *NDSS*, 2007.

[3] E. Tan, L. Guo, S. Chen, X. Zhang, and Y. Zhao, "Spammer behavior analysis and detection in user generated content on social networks," in *Distributed Computing Systems (ICDCS), 2012 IEEE 32nd International Conference on*, pp. 305–314, IEEE, 2012.

[4] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@ spam: the underground on 140 characters or less," in *Proceedings of the 17th ACM conference on Computer and communications security*, pp. 27–37, ACM, 2010.

[5] A. Dasgupta, K. Punera, J. M. Rao, X. Wang, J. Rao, and X. Wang, "Impact of spam exposure on user engagement.," in *USENIX Security Symposium*, 2012.

[6] Y. Zhao, S. Yang, V. Narayan, and Y. Zhao, "Modeling consumer learning from online product reviews," *Marketing Science*, vol. 32, no. 1, pp. 153–169, 2013.

[7] J. K. Ayeh, N. Au, and R. Law, "do we believe in tripadvisor? examining credibility perceptions and online travelers attitude toward using user-generated content," *Journal of Travel Research*, vol. 52, no. 4, 2013.

[8] E. Owens and C. Turitzin, "News feed fyi: Cleaning up news feed spam," *Facebook newsroom*, 2014.

[9] F. Ridzuan, V. Potdar, A. Talevski, and W. F. Smyth, "Key parameters in identifying cost of spam 2.0," in *IEEE Advanced Information Networking and Applications*, IEEE, 2010.

[10] A. M. Aladwani and P. C. Palvia, "Developing and validating an instrument for measuring user-perceived web quality," *Information & management*, vol. 39, no. 6, pp. 467–476, 2002.

[11] Z. Yang, S. Cai, Z. Zhou, and N. Zhou, "Development and validation of an instrument to measure user perceived service quality of information presenting web portals," *Information & Management*, vol. 42, no. 4, pp. 575–589, 2005.

[12] H. Suh, N. Shahriaree, E. B. Hekler, and J. A. Kientz, "Developing and validating the user burden scale: A tool for assessing user burden in computing systems," in *CHI*, pp. 3988–3999, ACM, 2016.

[13] T. P. Novak, D. L. Hoffman, and Y.-F. Yung, "Measuring the customer experience in online environments: A structural modeling approach," *Marketing science*, vol. 19, no. 1, pp. 22–42, 2000.

[14] N. Diakopoulos and M. Naaman, "Towards quality discourse in online news comments," in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pp. 133–142, ACM, 2011.

[15] K. Rodden, H. Hutchinson, and X. Fu, "Measuring the user experience on a large scale: user-centered metrics for web applications," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2395–2398, ACM, 2010.

[16] B. Shelat and F. N. Egger, "What makes people trust online gambling sites?," in *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, pp. 852–853, ACM, 2002.

[17] C. L. Corritore, R. P. Marble, S. Wiedenbeck, B. Kracher, and A. Chandran, "Measuring online trust of websites: Credibility, perceived ease of use, and risk," *AMCIS 2005 Proceedings*, p. 370, 2005.

[18] D. W. Straub, "Validating instruments in mis research," *MIS quarterly*, pp. 147–169, 1989.

[19] L. J. Cronbach, "Test validation," *Educational measurement*, 1971.

[20] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering–a systematic literature review," *Information and software technology*, vol. 51, no. 1, pp. 7–15, 2009.

[21] C.-C. Hsu and B. A. Sandford, "The delphi technique: making sense of consensus," *Practical assessment, research & evaluation*, vol. 12, no. 10, pp. 1–8, 2007.

[22] M. G. Keith and P. D. Harms, "Is mechanical turk the answer to our sampling woes?," *Industrial and Organizational Psychology*, vol. 9, no. 1, pp. 162–167, 2016.

[23] S. Schnorf, A. Sedley, M. Ortlieb, and A. Woodruff, "A comparison of six sample providers regarding online privacy benchmarks," in *SOUPS Workshop on Privacy Personas and Segmentation*, 2014.

[24] D. J. Hauser and N. Schwarz, "Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants," *Behavior research methods*, vol. 48, no. 1, pp. 400–407, 2016.

[25] K. Casler, L. Bickel, and E. Hackett, "Separate but equal? a comparison of participants and data gathered via amazons mturk, social media, and face-to-face behavioral testing," *Computers in Human Behavior*, vol. 29, no. 6, pp. 2156–2160, 2013.

[26] A. B. Costello and J. W. Osborne, "Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis," *Practical assessment, research & evaluation*, 2005.

[27] E. T. Loiacono, R. T. Watson, D. L. Goodhue, *et al.*, "Webqual: A measure of website quality," *Marketing theory and applications*, vol. 13, no. 3, pp. 432–438, 2002.

[28] K. Z. Zhang, S. J. Zhao, C. M. Cheung, and M. K. Lee, "Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model," *Decision Support Systems*, vol. 67, 2014.

[29] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *Journal of Big Data*, vol. 2, no. 1, p. 23, 2015.

[30] S. Xie, Q. Hu, J. Zhang, and S. Y. Philip, "An effective and economic bi-level approach to ranking and rating spam detection," in *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pp. 1–10, IEEE, 2015.