Analysis of Recycling Capabilities of Individuals and Crowds to Encourage and Educate People to Separate Their Garbage Playfully

Pascal Lessel
DFKI GmbH
Saarbrücken, Germany
pascal.lessel@dfki.de

Maximilian Altmeyer
DFKI GmbH
Saarbrücken, Germany
maximilian.altmeyer@dfki.de

Antonio Krüger
DFKI GmbH
Saarbrücken, Germany
antonio.krueger@dfki.de

ABSTRACT
Sorting garbage is a relevant topic in many countries as it contributes to environmental protection. Empirical evidence suggests that not all people separate waste, potentially because they do not know how to do it correctly or are simply not motivated enough. We present the results of an online study (N=184) investigating people's capabilities for classifying waste, their capabilities to improve in this task over time and their current garbage separation behavior. The study confirms that the Wisdom of Crowds is applicable in this context as the crowd produces only half as many errors as the individual and feedback helps participants to improve. Based on this, we introduce the idea of a crowd classifying waste in a game, with their classification result then being used as feedback on gamified public trash cans to educate both the crowd playing the game and people using the trash can playfully.

Author Keywords
Gamification; Games With a Purpose; Wisdom of Crowds; Waste Separation; Behavior Change Support System

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION
Carefully handling waste is a relevant topic, as world cities in 2012 generated about 1.3 billion tonnes of solid waste per year, which will increase to 2.2 billion tonnes by 2025 [12]. In terms of recycling, in Germany, four (sometimes five) different trash bins for households are available that are designated to hold only a specific kind of trash, and in addition, glass containers (different containers for clear, green and brown glass) for non-returnable bottles can be found in all neighborhoods [9]. The situation is similar in other countries[1]. If the separation of garbage is done properly, it has a positive effect in terms of environmental protection, e.g. by reducing CO2 emissions [9] or greatly increasing the recovery rate of domestic waste [8]. Despite this obvious advantage of separating waste, not everyone seems to do it [7], resulting in the situation that the content of bins cannot be reused properly because of "contamination". There are different reasons why people do not separate waste correctly. Work conducted on this topic (e.g. [13, 24, 26]) suggests, among other reasons, that the number of bins and rules on what belongs in which of them makes it difficult to do it properly; that recycling seems not cost-effective enough to participate in; is inconvenient, if not every bin is easily reachable and that missing incentives result in a lack of motivation.

Figure 1. A German public trash can for waste separation.

In HCI, the topic of encouraging people to reflect on their recycling behavior (e.g. [27]) and waste in general (e.g. [10]) has been under investigation for a few years now. Some approaches investigate what happens if the trash cans are enhanced with technology. One prominent approach is the BinCam [21]. One part of this system is that paid crowd workers (recruited via Amazon Mechanical Turk (AMT)[2]) are meant to tag objects in pictures taken from a trash can according to waste categories. The performance of AMT was not good: in a random picture sample, 15 of 20 classifications were wrong [21]. As the focus of the work was not on the crowd, it remains unclear whether it was a problem of the crowd or only an issue of AMT, which has provided mixed results in the past (compare [2] but also e.g. [14]).

In our work, we want to focus on the crowd, as we see a big opportunity: Depending on the underlying design, individuals might, while classifying waste, also be affected by the task and adapt their behavior. To inform a system design, our goal

[1] `http://news.bbc.co.uk/2/hi/europe/4620041.stm`, last accessed on 05/01/2014

[2] https://www.mturk.com, last accessed on 05/01/2014

CHI 2015, April 18–23, 2015, Seoul, Republic of Korea.

http://dx.doi.org/10.1145/2702123.2702309

was to receive clearer insights on individuals' capabilities in classifying waste, on their attitude towards waste separation and whether they improve in this task over time while classifying. We decided to use an online questionnaire, in which participation was voluntary and without any monetary compensation, to reduce the chance of random answers to earn money faster [14], which could have been an issue before.

The results informed the design of a two component persuasive system: A public trash can, consisting of several bins (similar to Figure 1) for different types of waste, capable of taking pictures of newly discarded objects, and a display attached to it to show feedback. We decided to utilize a Gamification [5] approach for the trash can and try to motivate people to separate waste correctly by relying on a competitive setting. The second component is a mobile app which is designed as a Game With a Purpose [25] in which players lead a recycling company and one of the core activities they need to do is to classify waste (i.e. pictures the modified trash cans produce) to improve their revenue. In comparison to the BinCam, this app replaces the paid crowd by people playing the game and trying to improve their game score. The rationale behind using such a setting is that people not only in front of the trash can but also within the crowd might learn how to separate waste in a playful manner and will then do it in their "real life" more thoughtfully, even when not exposed to the game. We also present results of a first evaluation of this setting, showing that people appreciate this concept.

This paper contributes an understanding of individuals' capabilities in separating waste, shows how a crowd-based approach produces more reliable results than an individual alone and how feedback helps people to improve in this task. These findings are then transferred into a persuasive system design with the goal to educate people; this is presented afterwards.

RELATED WORK

Several approaches are available that investigate how a trash can can be modified. The effect of a trash can that is capable of taking pictures of its content is investigated in the BinCam [3, 4, 21]. Here, a bin was modified with sensors and whenever a new item was discarded, a picture was uploaded to a Facebook page. Besides the social pressure that other people on Facebook would see what was (potentially incorrectly) discarded, the authors also added some Gamification elements. For this, pictures were sent to Amazon Mechanical Turk and paid crowd workers counted the visible objects in terms of the waste category they belonged to. A competitive mechanism was added, in which the different can owners could see how well they were doing in comparison and could collect achievements. As stated, the authors mentioned that the quality of the crowd classification was not good and in conducted user studies (based on interviews with the owners), it was shown that the frequency of feedback was too coarse in the original approach, that people lose interest after some time and that there are privacy issues. We also utilize a trash can which is able to take pictures of the (newly) discarded object. In contrast to the BinCam approach, we aim for public trash cans with bins for several waste types to minimize the privacy issues occurring at home and to encourage waste separation. We will not use paid crowd workers to classify the pictures. Instead, we propose a game that should encourage people to classify such pictures and therefore have a potential positive effect on their own waste separation behavior.

Zlatow and Kelliher investigated public recycling bins, as almost 70% of the material contained in recycling bins is deemed unusable due to contamination [27]. They found that users need to be engaged on a deeper level. One proposal is the usage of a reward system in which a valid waste disposal is automatically detected and incentives are provided. Another proposal is to add educational facts around recycling to turn the act of recycling into an experience, which is also shown to be beneficial [26]. All these approaches try to make the act of throwing away, mostly done unconsciously, more conscious and engaging. With our approach we also aim for turning the usage of the trash can into an experience, e.g. by providing fast feedback and rewards.

Reif et al. [16] proposed Cleanly, an educational system, based on fieldwork and an online questionnaire. One of the results was that every second participant thinks that direct feedback would encourage them to participate in programs aimed at solving environmental problems. This, in general, is an important strategy in the area of encouraging behavioral change in sustainability (cf. for example [19]). Reif et al. introduces the holistic approach *trashducation*: the effort to educate people in their trash management by creating awareness of the trash they produce, decreasing the production of trash and endorsing proactive thinking about the environment. This is also relevant for us, as we also see education and motivation as important aspects of our system.

The work of McCarty and Shrum [15] makes clear that supporting recycling is an important topic, that the approaches in this context are relevant and that we need to understand why certain people engage in recycling while others do not, to better design programs that increase recycling behavior: For instance, the more people know about recycling, the higher the chance that they engage in recycling; or the more convenient recycling is made, the more people will recycle. Incentives also play a crucial role. All aspects have a direct impact on the work presented here, as we try to educate users, make recycling more convenient and provide virtual incentives.

Educating (by giving proper feedback) and motivating people to more thoughtfully recycle waste places this work in the area of sustainable HCI [18] and fits into the landscape provided by DiSalvo et al. [6]. The authors stated that finding new ways of engaging users and making them experts in sustainability is an emerging issue. Using game elements and social components provides individuals with further incentives on top of doing something reasonable [23]. Work, e.g. in the area of reducing energy consumption, has indicated positive effects of such approaches (e.g. [11]).

To our knowledge, none of these approaches analyzed the capabilities of a crowd in classifying waste and how their decisions could be utilized directly at a trash can. Our work tries to fill this gap by conducting an online questionnaire. The results informed the design of a mobile app game and a

gamified trash can with the purpose to educate and encourage people to separate waste more thoughtfully and properly.

ONLINE USER STUDY

To inform the design of our system, we first investigated how capable people are in classifying waste, and if people make mistakes in classification, what options there are to improve their performance over time. A second aspect of the user study was to receive insights into current waste separation behavior of participants. With the study we tried to find evidence for the following hypotheses:

H1 An individual is in general not capable of classifying waste without errors.

H2 Aggregating the individual classification results leads to lower error rates in comparison to only considering individual decisions.

H3 Individuals improve over time, when they receive feedback on whether or not their past decisions were correct.

H4 Visualizing the results of other participants influences individual performance in later classification tasks.

H1 is motivated by the related work showing that people have problems classifying waste correctly. **H2** is based on the idea of the Wisdom of Crowds [20], which states that a group of people is able to come to a better decision than an individual. **H3** focuses on the educational aspect. In [22] it was shown that feedback (on wrongly separated objects) based on an analysis after the garbage was picked up has a positive impact on future decisions. We want to replicate a similar effect in our setting, but by giving immediate feedback. **H4** is based on the idea that a comparison to peer decisions might have a stronger impact, than only right-or-wrong feedback alone.

Method

We implemented a gamified online questionnaire that mimics the classification process we envisioned for the crowd in our game later on. The Gamification elements were meant to encourage people to finish the questionnaire, but were used to assess potential elements for this game as well. In general, we added points if a classification was correct and subtracted points if not. We provided bonus points if an answer was given quickly and correctly. During the classification tasks (depending on the group; see below) the participants could see how many points they needed to get to the next place on the high score list and how big the distance from the previous position on the high score list was. In the end, the participants were shown the complete list, in which only nicknames and points were displayed. As garbage separation is different between countries[3], we set up the questionnaire only in German and required that participants to have lived at least three years in Germany to participate. The questionnaire was published on student mailing lists and social networks.

[3] `http://news.bbc.co.uk/2/hi/europe/4620041.stm`, last accessed on 05/01/2014

Figure 2. Example pictures for every waste category.

Behavioral Questions

After the introduction, in which we explained that the goal is to receive insights into waste separation behavior of German people, we asked the participants whether they think waste separation is easy, whether they are able to separate waste correctly and how complicated they judge waste separation to be in Germany. These questions (and all the other behavioral questions) were to be answered on a 7-point Likert scale, with labels shown on the extreme values. A set of classification tasks (see below) followed and consecutively we let participants judge their own performance in this task overall and continued with questions on the different game elements. After that, we asked questions about their waste separation behavior and collected demographic data.

Classification Task

We informed the participants that several objects would be shown, that they should classify them according to how they would be disposed of correctly in Germany and that we would assign points for their answer. The faster they provided an answer, the better for their overall score (in the case of a correct answer). For every picture they needed to decide whether the shown object belonged to domestic waste, lightweight packaging, waste paper, organic waste or none of those categories.

To acquire a ground truth for the pictures, we consulted different official material on the topic of how to separate garbage in Germany correctly and selected only pictures for which we found consensus in the material. In addition, every picture was rated by three judges beforehand in terms of its potential difficulty level and in case of incongruity, a discussion solved it (only in five cases was such a discussion necessary). The general selection of pictures was guided by the goal to have eight pictures with different representatives per difficulty category (three easy pictures, three medium pictures and two difficult pictures per category), resulting in $8 \times 5 = 40$ pictures. An example of an easy classification task was an apple (organic waste), one for a medium task was a non-returnable can (lightweight packaging), and a supermarket bill (which consists of thermal paper which should be disposed of as domestic waste) was assessed as hard to classify. Unlike the BinCam [21], we planned to use computer graphics algorithms in our system later on to extract one disposed object at a time; hence, every picture contained only one object. Figure 2 shows examples of the different waste categories.

Figure 3. Example classification task of the questionnaire.

Participants could skip a classification completely and could add a comment to each of them. In addition, after they received feedback (depending on their group; see below), they also had the chance to comment on it. For every decision, participants also had to indicate how confident they were in it on a 7-point Likert scale. To simplify this, we used a grid layout in which the participants could assign a classification answer and a confidence with only one click (cf. Figure 3). We also measured the time per answer — an unusually long time might be an indicator that the participant used outside information. The order of the pictures was randomized and participants were assigned to one of five groups randomly (but equally) at the start. We varied the feedback participants received after a classification, depending on the group:

- **No feedback** (NF): Participants do not receive any feedback after their classification. They only see their overall score after all classification task and the high score list. This condition serves as a baseline.
- **Ground truth feedback** (GTF_{Only}): Participants always see whether their decision was correct or incorrect and what would be the correct answer. All gamification elements are available, i.e. they can see their points, how many points they are from the next position on the high score list, and their current placement on it.
- **Ground truth feedback with explanation** ($GTF_{Explanation}$): Same as GTF_{Only}. In addition, an explanation of how the ground truth decision was given by providing a short statement, and a reference to an official document was shown.
- **Ground truth feedback with same crowd decision** ($GTF_{CrowdSingle}$): Same as GTF_{Only}. In addition, they see how many people decided in the same way, by seeing a percentage (e.g. 12% had the same opinion).
- **Ground truth feedback with crowd decisions** ($GTF_{CrowdAll}$): Same as GTF_{Only}. In addition, they see how the crowd decided, by seeing a percentage per classification option.

Retest of false classifications

To check if any improvements were achieved due to the feedback, we asked the participants whether they wanted to improve their score by classifying a few more pictures before seeing the high score list ("bonus run"). These pictures were selected based on the errors made. The feedback group assignment remained the same for this task. In addition, we asked the participants, whether they wanted to receive an additional invitation to a follow-up study. One week after completion of the questionnaire, they received an e-mail with a new link. The questionnaire consisted of two questions (*"Did you consider the topic of waste separation more during the last week?"* and *"Do you think that you disposed of and sorted waste more thoughtfully during the last week?"*) and a set of classification tasks split in two chunks. The selection of the objects was again guided by the errors a participant made in the original questionnaire. In the first chunk, instead of a picture, we showed only the name of the object (everything else in the task remained the same). This should have minimized picture recognition effects. In the second chunk, we used the same pictures again. The order of text/pictures in both chunks was randomized and no feedback was provided.

Results

The questionnaire was accessible for four weeks and 184 people completed the questionnaire (78 female, 93 male; 13 did not specify a gender). The age distribution was skewed younger (29 <21, 124 to 21-30, 12 to 31-40, 12 to 41-50, 7 >50), due to the way we promoted the questionnaire, but which is not a problem in our case, as this fits our later target group (see system design below). Concerning the main backgrounds, 87 participants reported being students, 59 being employed and 15 working in an apprenticeship. The number of people the participants lived with was well distributed, while most (48) reported living together with 2 other persons.

Waste Separation Behavior

We asked several questions before and after the classification task to assess people's attitude towards waste separation. In general, people reported thinking waste separation is easy (M=5.11, SD=1.28, Mdn=5), that they are able to separate waste correctly (M=5.22, SD=1.27, Mdn=5), that they are in general eco-sensitive (M=5.07, SD=1.39, Mdn=5), that they separate to the best of their knowledge (M=5.42, SD=1.58, Mdn=6) and that waste separation is important to them (M=5.03, SD=1.42, Mdn=5). Participants (54) who disagreed (selecting four or less on the Likert scale) were shown potential reasons: 17% selected that it is too complicated, 30% that it is too much effort, 39% that there is too little space for waste separation at home, 31% that waste separation is useless and 26% stated that incentives are missing. In contrast, of the participants who agreed (130) 89% want to save resources and protect the environment, 35% have financial reasons and 65% also stated that it is their responsibility to protect the environment. In both cases free text answers could be provided, but were inconclusive. The participants disagreed with the statement that they seek information, if they are unsure how to dispose waste correctly (M=3.17, SD=1.58, Mdn=3) and with the statement that waste separation in Germany is difficult (M=3.46, SD=1.56, Mdn=3). After the classification task, we asked again whether they are able to separate waste and observed a small change in the self-assessment in respect to their capabilities in waste separation (M=4.9, SD=1.17, Mdn=5, $t(183)=3.35$, $p<.01$, $d=-0.25$).

Performance in the Classification Tasks

We deleted answers in the classification tasks that took longer than three times the average answering time (M=7s, SD=104s) as an explanation for this might be that participants had utilized external material. In the end, 7236 classifications in the main run were considered. Skipped classifications were counted as wrong classifications. On average an individual made M=23.37% errors (SD=7.61%, Mdn=22.5%), which supports **H1** — an individual is not able to classify waste without errors in general. We analyzed the errors per waste category and with respect to our a priori assigned difficulty level. Table 1 shows that the a priori levels fit in general (with organic waste as the only exception), i.e. medium and hard labels produced more errors than easy-rated objects and showed that some objects are harder to dispose of correctly. We analyzed the kinds of errors made more deeply, i.e. if an error was made, we checked which other category was selected. The corresponding confusion matrix is shown in Table 1 and together with the difficulty levels shows that the domestic waste objects produced the most errors. Another aspect which is also of interest is the relationship between assigned confidence value and the decision. We found that many participants utilized only the extreme values (a reason might be the time aspect to achieve more points; see limitation section). We filtered and used the remaining 88 answers for analysis. It showed that if people decided wrongly, they provided on average a lower confidence score than if they decided correctly ($M_{incorrect}$=4.89, $SD_{incorrect}$=1.00, $M_{correct}$=5.68, $SD_{correct}$=1.07, t(87)=-11.24, p<.01, d=-0.79).

	P	L	D	OR	O
Easy	5.51%	4.2%	24.05%	1.65%	8.14%
Medium	11.59%	11.01%	45.23%	21.06%	24.2%
Difficulty	25.36%	31.06%	87.77%	20.11%	67.97%
Overall error	12.75%	13.47%	48%	13.5%	29.12%

	P	L	D	OR	O	S
P	**87.29%**	4.79%	6.15%	0.55%	0.9	0.34%
L	0.83%	**86.63%**	9.57%	0%	2.83%	0.14%
D	22.36%	12.9%	**51.83%**	1.72%	10.62%	0.55%
OR	1.32%	0.13%	11.24%	**86.54%**	0.55%	0.21%
O	0.07%	11.61%	17.2%	0.0%	**70.84%**	0.27%

Table 1. Top: Errors per waste category and a priori assigned difficulty level, aggregated over all participants. Bottom: Confusion matrix showing the aggregated classification results (P=paper waste, L=lightweight packaging, D=domestic waste, OR=organic waste, O=other).

Assessing Improvements Over Time

The participants had the chance to improve over time:

- During the run: After removing outliers, we compared the error rate in the GTF-conditions (M=22.6%, SD=6%) with the NF-condition (M=25.18%, SD=8.8%), but a t-test showed no significant difference (p=0.5).

- Bonus run: 123 participants completed the bonus run. We expected an improvement in all GTF-conditions, based on the recognition of the pictures and remembering the correct answers. As the option to receive bonus points was not articulated before the classification task and answering the behavioral questions served as distractor, a better performance indicates that participants had seen the correct result and could also remember it also later on. The average error rate of the users was M=80%, SD=21% in the NF-condition and M=12%, SD=13.1% in the GTF-conditions. A one-way ANOVA showed a significant effect between condition and error rate (Welch's F(4,77.54)=68.07, p<.01, est. ω=0.83). The Games-Howell post hoc procedure was used since the homogeneity of variance assumption was not met, and revealed that every GTF-condition is significantly different from the NF-condition (every comparison with p<.01). We also checked whether the crowd-feedback had any impact on the way users performed. The rationale behind this was that users might be more likely memorize wrong classifications, if they saw how the crowd decided. However, none of the comparisons showed a significant effect.

- Follow-up run: As the bonus point run could still have only a short-term effect, we wanted to measure whether we could find evidence for improvements after a week. As a limitation, only 36 participants (NF:12, GTF:5, $GTF_{Explanation}$:3, $GTF_{CrowdSingle}$:8, $GTF_{CrowdAll}$:8) took part in this second study, limiting its expressiveness. We asked whether participants sought information about waste separation between the two questionnaires (M=2.86, SD=1.84, Mdn=2) and whether they separated waste more conscientiously (M=2.92, SD=1.86, Mdn=3). As outlined, we had two chunks of the classification task. The average error rate of users previously in the NF-condition in chunk 1 (chunk 2) was M=54.7%, SD=11.4% (M=54.3%, SD=14.3%) and of users previously in the GTF-conditions was M=33%, SD=18.2% (M=23.8%, SD=6.6%). Because of the low number of participants in $GTF_{Explanation}$, we excluded this condition in the one-way ANOVA analysis of chunk 1 and 2; additionally, one participant completed only chunk 1. We found a significant effect between group and error rate in each chunk (F(3,29)=5.61, p<.01, ω=0.54 and F(3,28)=26.3, p<.01, ω=0.84). Gabriel's post hoc procedure revealed that every GTF-condition is significantly different from the NF-condition (for chunk 1 with p<.05, for chunk 2 with p<.01), indicating that feedback was indeed helpful for subsequent classifications, even after a week. The error rates over all conditions in chunk 2 (M=34.2%, SD=17.6%) compared to chunk 1 (M=38.5%, SD=16.4%) were significantly lower (t(34)=3.71, p<.01, d=0.27) indicating that people indeed utilized recall effects, but in both chunks feedback had improved the user performance, providing evidence supporting **H3**. Again, no effect could be found by considering the crowd conditions. Hence, **H4** cannot be supported with the results made.

Crowd Performance

We tested different aggregation methods to see if a crowd-based approach produces more reliable results than individuals in this setting. We used different aggregation algorithms: a standard majority voting (MV); a weighted majority voting taking the provided confidence scores into accounts as weights (MMVC) and a weighted majority voting considering the percentage of correct decisions as weights to identify experts (WMVE). The problems with the confidence values (see above) resulted in a strict condition, in which we set the

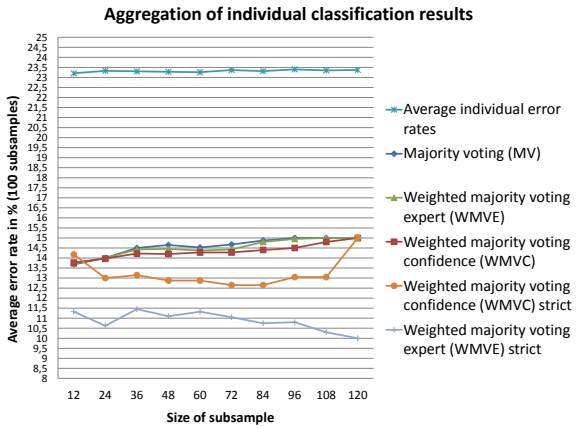

Figure 4. Crowd error rate in relation to aggregation algorithm and averaged individual error rates for different subsamples.

weights to 0 for participants who apparently did not use the confidence value properly. We also integrated a strict value for the expert rating (setting the weight to 0 for participants having a higher than average error rate). To receive more reliable results we took random subsamples of the dataset and repeated the subsample selection 100 times. Figure 4 shows the aggregated results in terms of error rates. For a better direct comparison, we also averaged over the individual (aggregated) error rate in the corresponding subsample. The result supports **H2**: the Wisdom of Crowds is applicable in this scenario, as the crowd produces a better result, independent of the specific aggregation mechanism and even at a low sample size. In our case, the strict expert metric worked best.

Perception of Feedback and Gamification Elements

For the assessment of the perception of feedback (namely true/false feedback, seeing the correct answer, justification for the ground truth answer and how the crowd has decided) and Gamification elements (seeing points, position on high score list, competition in general) we asked whether the element motivated them to classify more pictures (A), or if the participants belong to a condition in which the element was not used (B), whether it would have motivated them. The feedback elements used were seen as motivating (Mdn=5 for the different aspects in A, Mdn=6 in B); the only exception was the crowd feedback (A:n=74, M=3.5, SD=2.03, Mdn=3, B:n=110, M=3.72, SD=1.84, Mdn=4), with no significant difference between $GTF_{CrowdSingle}$ and $GTF_{CrowdAll}$. An explanation could be that such feedback is not interesting when the correct answer is also shown. Here, a follow-up study needs to investigate this further. The Gamification elements were also perceived positively (median 4 for seeing points, other 5). In all but one case participants not exposed to the corresponding element provided slightly higher scores. Always seeing the current position on the high score list is the only aspect which was perceived better by participants exposed to it (A:n=148, M=4.18, SD=2.16, Mdn=5, B:n=36, M=3.5, SD=1.76, Mdn=4). Participants were also able to add further aspects that would motivate them to classify more waste, but this was in general inconclusive.

Discussion

The study showed that individuals make errors in recycling. We provided the five major German categories and on average an individual disposes of one in five items incorrectly. The individuals classified themselves as eco-aware and capable of separating waste correctly but were not particularly interested in finding the information on how to dispose of an object correctly if unsure. In this sense, we found evidence that supports **H1** serving as additional basis to motivate the search for proper user assistance in this task, i.e. providing a system which gives immediate feedback. The conducted analysis showed that the Wisdom of Crowds is applicable in this context, as the crowd performed better than the individual (**H2**): depending on the aggregation algorithm, the crowd produces roughly only half as many errors as an individual on average. An explanation can be found in the work of Surowiecki [20], as the five rules of collective intelligence are also applicable here. In contrast to the BinCam [21], the crowd performance in our study was much better. Reasons for that might be that congruency of nationality and waste disposal rules were ensured in our case and that only one object should be classified at a time instead of counting all elements in a bin, which might be hard, as objects might be partly covered by objects above them. Our crowd was not paid and motivated to achieve a game score, which might convince them to decide more thoroughly [14]. We found evidence supporting **H3**; participants used the feedback shown and produced fewer errors not only for the same pictures but also for text-only representations after one week. Additionally, the gamification and feedback elements were perceived positively. This opens up the option for our game scenario in which not only the user in front of the trash can can improve through the feedback, but also the crowd providing it. With this setup though, we could not find supporting evidence for **H4**.

The results indicate that feedback helps people even when they are unaware of being retested. Thus a system offering a feedback mechanism and integrating it into the everyday activity of throwing something away can have a positive effect. Our study showed that an individual alone is not capable of providing sufficient feedback, but the error of a (potentially small) group of people is significantly lower. It seems possible to deploy a system that utilizes crowd feedback, but it should also integrate functionalities to educate the crowd explicitly as well. Instead of generating feedback and educating the crowd at the trash cans, we argue for using an additional system to make feedback accessible faster: an option to connect a crowd separately, also receiving feedback on their classification and with further educational options (e.g. external resources) to improve themselves, which is motivated by Gamification/game elements (as shown in our study). Integrating a Crowd as a "tool" for providing feedback at the point of interest while potentially being self-educated (and thus behaviorally influenced) is an interesting aspect for Sustainable HCI that might also be of relevance

Limitations

The study itself has limitations. First, we only investigated the waste separation behavior of German people. It remains open how this differs in other countries, which is a threat to

external validity. Second, it can not be completely ruled out that the high amount of positive answers to behavior questions were influenced by social desirability or acquiescence bias. As the focus of our study was different (cf. hypotheses), this is only a minor limitation. Third, the survey used Gamification elements to motivate participants to do the (potentially) not-so-interesting task of classifying several waste pictures. As a negative side effect, people might have misjudged the time constraint and skipped through the feedback, instead of reading it thoroughly. If this is true, the effects in the feedback conditions might currently be underestimated. Forth, the selection of pictures includes uncommon objects, for the sake of the survey but the situation at a real trash can might be different. Considering that uncommon (in general, a priori labeled as medium and difficult) objects were classified worse, the performance of the crowd in a real-life situation is still questionable and will be investigated with our prototype. Fifth, the age distribution was skewed younger, limiting the expressiveness. Nonetheless, the target group for our mobile app, matches this distribution[4].

SYSTEM DESIGN

Based on our study we propose a system design consisting of two components: a modified public trash can and a mobile app. While the trash can produces images from discarded objects, the mobile app is able to present them to a crowd. Both components have the goal to get people engaged in sorting their waste: the trash can by displaying whether the object was disposed of correctly and building a competition around it and the mobile app by motivating people to improve their fictive "recycling company". A crucial component is the classification of waste. While an automated approach would be interesting it would not have helped in terms of the learning effect for the crowd, which is why it can only be seen as complementary and was not followed at this point in time.

Scenario

Alice walks through campus and encounters one of the trash cans belonging to the "Trash Game". She inspects the display and learns that her faculty has received many points in this week but is not in first place. She hopes that more people will sort their waste correctly to improve their faculty's score. Later, during a lecture, she receives a new notification on her smartphone generated by the "Trash Game" mobile app. Her fictive recycling company, WasteGoneProperly, has received a new task and needs to handle a new object. Because Alice plans to improve her company with new upgrades, she needs to gain more in-game money, so she immediately reacts to the job, as this provides her with bonus cash. She knows that the trash cans produces these pictures and provides feedback based on the opinion of the crowd. She particularly likes that, as this helps people to learn how to do it correctly. Within the game she reacts to the new picture by stating how her company would handle this kind of product. Only then does she see how other players have classified the garbage. Unfortunately, the majority of the crowd decided differently. Alice wonders about this and later in the day she decides to dive into this topic. In the meantime, certain companies have received the task to provide evidence showing that their decision was indeed correct, and Alice is able to read their statements. Some arguments and references are accepted by the community, so that she remembers how to classify the item and to do it correctly in the future.

[4]cf. `http://goo.gl/WQhaCO`, last accessed on 05/01/2014

Figure 5. Left: Exterior/interior of the prototype. Right: Steps of extraction algorithm: (From top to bottom) Before insertions, after insertions, recognized differences, red rectangle showing extracted picture.

Modified Trash Can

Unlike the BinCam-approach [21] we focus on public trash cans, as people in Germany are already familiar with trash cans consisting of multiple bins in public (cf. Figure 1), potentially decreasing privacy issues. These bins are color coded, and printed pictures show prominent representatives of the corresponding categories. The design process of our trash can was guided by the realization of requirements: the can needs to recognize newly discarded objects; a picture should be generated that shows only the new object and should be made available online (in the mobile app); it should be able to receive crowd feedback and display it in a playful way by using Gamification elements.

Hardware design

We created a hardware prototype as a proof-of-concept. The basis is a wooden frame consisting of three different chambers and a smartphone attached on the ceiling of each (see Figure 5, left). The purpose of these devices is the detection of new objects and taking pictures of them. A Raspberry Pi is used to do the data handling between server and smartphones and displays the results on the trash can's display. For the persons using the trash can, the basic use stays the same — it is sufficient to throw the waste in one of the bins. The only difference visible is that this insertion is recognized and direct feedback on the display is shown.

Software design

We decided to create a competitive environment with the trash cans of the "Trash Game". Each can is associated to a group. Depending on the area of use, this could be, for example, different faculties, or divisions in a company. The smartphones are always checking for differences in the RGB pixels taken by the camera (which is more robust in comparison against changes in the lighting conditions). If a difference is recognized, a picture is taken after two seconds (to allow time for the object to fall). Subsequently, this picture is compared to the last picture taken to extract only the newly discarded object. For that, we identify areas that have changed and to reduce errors (as underlying objects might also have changed their position), we do a template matching and compare pixel arrangements of these areas with the pixel areas in the last picture and discard similar areas (see Figure 5, right). Informal tests indicated that the algorithm is robust enough to produce pictures similar to the ones used in our study. These pictures are sent to a server, which distribute them to the mobile app to get the crowd engaged. Our study has shown that even a small number of people are able to classify waste better than an individual. Currently, we consider the crowd votes after 30 seconds and if there are enough votes (20 is our current threshold) we display this as final result on the trash can's display (otherwise it is shown as preliminary, with more time being needed to come to a consensus). The display shows this counter and dynamically visualizes the distribution of votes the crowd made. If a person does not want to wait until the time is up she also has the chance to scan a barcode and is redirected to a web page showing the results afterwards.

After a correct (or incorrect) disposal (i.e. a (dis)agreement between crowd and chosen bin) the trash can receives (loses) points. The points are always clearly visible. In addition, each trash can shows the CO2 savings/production from the correctly/incorrectly separated waste inside it and the mobile game is advertised as well. Different screens are shown if nothing is currently added: A high score list shows the points of the predecessor and the successor only for the current week (so that the high score is not discouraging because the distances are too big); or the last additions are shown with the crowd voting or a screen showing common disposal hints, and the advantages of waste separation being done correctly. All the screens are chosen to get people interested in the trash can itself and to provide them with gamified feedback (to get them engaged in waste separation) as well as providing them with feedback that might improve the learning rate. Overall, we follow the design suggestions of [17].

Mobile App

The goal is to get the crowd engaged in the classification of waste, as their decisions can be used to decide if an object was sorted correctly (as shown with a low error rate). To get the individuals engaged we decided to design the mobile app as a game fitting into Games with a Purpose [25]. The prototype is implemented as a web app to ensure device-independence. By receiving feedback, people are able to get better at subsequent classifications of the same or similar objects, as shown in our study. Hence, the main idea we follow with this app is

Figure 6. Landscape classification screen of the app after a decision.

that not only are people at the location of the trash can educated, but the crowd itself also has the chance to improve.

The player is head of a recycling company that tries to become the market leader by quickly and correctly disposing of specific objects. Occasionally, players receive tasks and the faster they react to them, the more money is generated, which can be spent for improving the company. This improves the market value of the company, but alters in-game mechanics as well (e.g. receiving more tasks or receiving more money per solved task). Currently, we distinguish three types of tasks, the *classification tasks*, *evidence tasks* and *knowledge tasks*. With different tasks we want to achieve a more diverse game play but also improve the data quality. A profile in the background is used which counts the times the player was correct and tracks demographic data, most importantly his country, as this moderates which tasks he receives: only pictures from bins located in the same country should be classified.

Classification tasks are triggered if a picture of a new object is taken by a trash can. Players receive a notification that a new classification task is available for them. By accepting it, they see the picture and the classification options (similar to the online questionnaire). In addition, they can provide a confidence value to their vote (cf. Figure 6). Only after they have selected an option they can see how the other companies have decided, and they will be informed about the final decision for this object. If they belong to the majority, they receive virtual money. If not, it depends on how big the difference in the voting is. If it is small (indicating that the crowd was not completely sure), no money is lost; otherwise, because the company needs to recycle the object in a more expensive way, some money is subtracted. The reason for showing the crowd distribution lies in the fact that people who see the distribution might be motivated to engage either in the seminars or in the evidence task (e.g. if they see that the vote was only slightly in favor of the other option). We used the strict expert aggregation, as this performed best in our study.

To support a deeper engagement with sorting questions, we integrated evidence tasks. Several companies are selected for this special task randomly, but can also participate in this task voluntarily. The main goal is to provide evidence (e.g. official links and/or reasonable explanations) for a previously classified object that either supports the crowd opinion or contradicts it. Similar to the Stack Exchange network[5], all players

[5]E.g. stackoverflow, `http://stackoverflow.com/`, last accessed on 05/01/2014

who participated in the crowd voting can see this discussion and up- and down-vote the different contributions, and finally accept an answer. Involved players receive a notification and potentially a (virtual) refund. In addition the company which provided the accepted evidence receives a reward that positively affects their market value. This task was created to account for crowd errors, which we also saw in our online survey, especially for uncommon objects.

The third kind of task is a task in which predefined questions are provided and players either contribute answers (e.g. "what is this object called?"or "do these pictures show the same objects?") or assess already-given answers (e.g. "Are these valid names for this object"?), which improves a knowledge base in the background. This can later serve to change the game mechanics and to inform ML-algorithms.

Besides these tasks the player has the option to attend "courses" in which commonly wrong classifications are presented with proper explanations. Attending a course provides the player a virtual certificate and bonus points, if he classifies such an object correctly in the future. To keep people engaged, we ensure that specific pictures that were already classified by the crowd (but not by the player) are given to the players even if nothing is discarded at the moment.

Preliminary Evaluation

To get an impression on how our prototypes are perceived, we presented them to university employees, students and visitors at the cafeteria foyer around lunchtime. They experienced the process by throwing an object (we provided several) into one of the bins and could also vote with the mobile app. The crowd feedback was provided in a Wizard of Oz-style experiment in which we select which feedback is shown (potentially adjusted by the participant vote). The selection was based on the results from our online study, e.g. if a supermarket bill was inserted, we presented the crowd classifications for this object, i.e. we used real values. While showing the process, we also explained the concepts of the prototypes and answered questions. Subsequently, the participants were provided with a questionnaire, consisting of several questions to be answered on a 7-point Likert scale and free-text questions.

35 people participated in our evaluation (12 female, 23 male). Questions concerning their waste separation behavior were answered similarly to our online study: Participants think that waste separation is easy (M=5.11, SD=1.23, Mdn=5), that they do it correctly (M=5.17, SD=0.923, Mdn=5) and to the best of their knowledge (M=5.6, SD=1.5, Mdn=6). They also think that waste separation in Germany is not complicated (M=3.83, SD=1.89, Mdn=4) and they do not seek more information if unsure how to separate waste correctly (M=2.71, SD=1.7, Mdn=2). Concerning the game concept, participants liked the trash can (M=5.77, SD=1.6, Mdn=6) and the mobile app (M=5.68, SD=1.57, Mdn=6). If they had the chance to decide to dispose of their waste in a normal or the augmented trash can, they would use our prototype to receive feedback (M=5.44, SD=1.66, Mdn=6). Eleven participants said this was because they liked the idea and four stated that it helps the environment. Three participants stated that the waiting time is a problem. We elaborated on this further and learned that 8 would wait, 14 would wait if feedback is provided quickly, 6 would not wait at all and 2 reported that they would wait only if unsure. Only one participant reported that he would not consider the mobile web page as an alternative if the waiting time is too long. Here, a solution could be to integrate approaches similar to [1], but with only virtual incentives. In general, the crowd feedback is judged useful (M=5.6, SD=1.33, Mdn=6) even though it could be wrong (which was also demonstrated). On the other hand, the responses were mixed for the question whether participants would also let the crowd classify their waste at home (M=3.71, SD=1.86, Mdn=4), mostly because of privacy issues, as the free-text answers showed (13), which is a replication of the results of the BinCam [21]. Concerning additional features, participants wished for further graphical elements, a variable design for different age groups, a mechanical component in which the waste is stored until the crowd has decided, and a way to achieve bonus points. Interestingly, they are undecided whether to play the mobile game (M=3.94, SD=1.9, Mdn=4). Here, a reason for this could be that only a part of the app was presented and the other functions were only explained. The answers showed that people focused more on the classification part inside the app and have not considered that these tasks are integrated into the game play. We will elaborate on this aspect further as a next step. In contrast, they liked that they could have an influence on the feedback the trash can shows (M=5.48, SD=1.56, Mdn=6). Concerning the functions, participants suggested a social media connection to compete with friends, collaboration with real recycling companies which serve as experts providing a more reliable ground truth, and that challenges should be integrated.

Discussion

Besides the goal to improve the waste separation capabilities of people at the trash can and within the crowd, the options provided by our setup might be interesting for recycling companies. The crowd provides a more reliable information on the content disposed in the bins; more specifically, through the knowledge tasks, we might also achieve a complete classification of waste inside the bin, which enables companies to better decide what should be done with the contents themselves. Moreover, with such an approach it could also be possible to learn from errors and share this knowledge with policy makers to help them to improve their own (educational) publications. In addition, it could be of interest for companies to deploy courses or publications within our app to improve crowd performance. The preliminary evaluation showed that both components are perceived well, even though participants were reluctant to play the mobile game. As a next step, we will focus on this part to receive deeper insights.

CONCLUSION

In this paper we presented results of a gamified online questionnaire which assessed people's capabilities in waste separation and found that people make errors in this task. In contrast, we demonstrated that crowd decisions in this setting produce only half as many errors as individuals. Based on the findings, we presented a novel system – the Trash Game, consisting of a trash can which is able to take photos of newly

discarded objects and show feedback on whether or not the waste was separated correctly. This decision is made by utilizing a crowd which is motivated to participate by playing a mobile game in which every player is head of a recycling company and, amongst others, has the task to classify waste correctly. The core idea of both components is to achieve a learning effect, in which people in front of the trash can and within the crowd improve in recycling and will do it correctly, even if not exposed to the game components. The next step is a long-term evaluation of the mobile app. Assessing the impact on the crowd's real life will be a focus, as well seeing how large the effects are that such an approach can produce.

ACKNOWLEDGEMENTS

This work has been supported by the Cluster of Excellence on "Multimodal Computing and Interaction" funded by the German Science Foundation (DFG).

REFERENCES

1. Bernstein, M. S., Brandt, J., Miller, R. C., and Karger, D. R. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proc. UIST 2011*, ACM (2011), 33–42.
2. Buhrmester, M., Kwang, T., and Gosling, S. D. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science 6*, 1 (2011), 3–5.
3. Comber, R., and Thieme, A. Designing Beyond Habit: Opening Space for Improved Recycling and Food Waste Behaviors Through Processes of Persuasion, Social Influence and Aversive Affect. *Personal Ubiquitous Computing 17*, 6 (Aug. 2013), 1197–1210.
4. Comber, R., Thieme, A., Rafiev, A., Taylor, N., Krämer, N., and Olivier, P. BinCam: Designing for Engagement with Facebook for Behavior Change. In *Proc. INTERACT 2013*, Springer Berlin Heidelberg (2013), 99–115.
5. Deterding, S., Dixon, D., Khaled, R., and Nacke, L. From Game Design Elements to Gamefulness: Defining "Gamification". In *Proc. MindTrek 2011*, ACM (2011), 9–15.
6. DiSalvo, C., Sengers, P., and Brynjarsdóttir, H. Mapping the Landscape of Sustainable HCI. In *Proc. CHI 2010*, ACM (2010), 1975–1984.
7. Environment Bureau Hong Kong. Blueprint for Sustainable Use of Resources. `http://goo.gl/JfSUZW`, May 2013. [last accessed 05/01/2014].
8. Environmental Protection Department. Programme on Source Separation of Domestic Waste. `http://goo.gl/JgxbHK`, May 2010. [last accessed 05/01/2014].
9. Federal Ministry for the Environment, Nature Conservation and Nuclear Safety. Waste Management in Germany 2013: Facts, Data, Graphics, December 2012.
10. Greengard, S. Tracking Garbage. *Commun. ACM 53*, 3 (Mar. 2010), 19–20.
11. Gustafsson, A., Katzeff, C., and Bang, M. Evaluation of a Pervasive Game for Domestic Energy Engagement Among Teenagers. *Comput. Entertain. 7*, 4 (Jan. 2010), 54:1–54:19.
12. Hoornweg, D., and Bhada-Tata, P. What a Waste: A Global Review of Solid Waste Management. *Urban Development Series*, 15 (March 2012), 1–116.
13. Hornik, J., Cherian, J., Madansky, M., and Narayana, C. Determinants of recycling behavior: A synthesis of research results. *The Journal of Socio-Economics 24*, 1 (1995), 105 – 127.
14. Ipeirotis, P. G., Provost, F., and Wang, J. Quality Management on Amazon Mechanical Turk. In *Proc. HCOMP 2010*, ACM (2010), 64–67.
15. McCarty, J. A., and Shrum, L. The recycling of solid wastes: Personal values, value orientations, and attitudes about recycling as antecedents of recycling behavior. *Journal of Business Research 30*, 1 (1994), 53 – 62.
16. Reif, I., Alt, F., Hincapié Ramos, J. D., Poteriaykina, K., and Wagner, J. Cleanly: Trashducation Urban System. In *Ext. Abstract CHI 2010*, ACM (2010), 3511–3516.
17. Schell, J. *The Art of Game Design: A Book of Lenses*. Morgan Kaufmann Publishers Inc., 2008.
18. Silberman, M. S., Nathan, L., Knowles, B., Bendor, R., Clear, A., Håkansson, M., Dillahunt, T., and Mankoff, J. Next Steps for Sustainable HCI. *Interactions 21*, 5 (Sept. 2014), 66–69.
19. Strengers, Y. A. Designing Eco-Feedback Systems for Everyday Life. In *Proc. CHI 2011*, ACM (2011), 2135–2144.
20. Surowiecki, J. *The Wisdom of Crowds*. Anchor, 2005.
21. Thieme, A., Comber, R., Miebach, J., Weeden, J., Krämer, N., Lawson, S., and Olivier, P. "We've Bin Watching You": Designing for Reflection and Social Persuasion to Promote Sustainable Lifestyles. In *Proc. CHI 2012*, ACM (2012), 2337–2346.
22. Timlett, R., and Williams, I. Public participation and recycling performance in England: A comparison of tools for behaviour change. *Resources, Conservation and Recycling 52*, 4 (2008), 622 – 634.
23. Tscheligi, M., and Reitberger, W. Persuasion as an Ingredient of Societal Interfaces. *Interactions 14*, 5 (Sept. 2007), 41–43.
24. Vining, J., and Ebreo, A. What Makes a Recycler?: A Comparison of Recyclers and Nonrecyclers. *Environment and Behavior 22*, 1 (1990), 55–73.
25. von Ahn, L., and Dabbish, L. Designing Games with a Purpose. *Commun. ACM 51*, 8 (Aug. 2008), 58–67.
26. Werner, C. M., and Makela, E. Motivations and Behaviors that Support Recycling. *Journal of Environmental Psychology 18*, 4 (1998), 373 – 386.
27. Zlatow, M., and Kelliher, A. Increasing Recycling Behaviors Through User-Centered Design. In *Proc. DUX 2007*, ACM (2007), 27:1–27:1.