A Comparison of Crowd Types: Idea Selection Performance of Students and Amazon Mechanical Turks

Victoria Banken
Universität Innsbruck

A Comparison of Crowd Types: Idea Selection Performance of Students and Amazon Mechanical Turks

Victoria Banken[1]

[1] University of Innsbruck, Department of Information Systems, Production and Logistics Management, Innsbruck, Austria
Victoria.Banken@uibk.ac.at

Abstract. Crowdsourcing is an effective means to generate a multitude of ideas in a very short amount of time. Therefore, companies and researchers increasingly tap into the power of the crowd for the evaluation of these ideas. However, not all types of crowds are the equally capable for complex decision-making tasks, which might result in poor selection performance. This research aims to evaluate differences in anonymous crowds and student crowds regarding their information processing, attention and selection performance. A web-experiment with 339 participants was conducted to reveal that 1) undergraduate Information Systems students perform better in idea selection than crowd workers recruited from Amazon Mechanical Turk, 2) attention checks increase selection performance and 3) while crowd workers indicate to process information more systematically, students acquire more information for evaluation than crowd workers.

Keywords: Open Innovation, Crowdsourcing, Crowd Types, Amazon Mechanical Turk, Student Sample, Attention

1 Introduction

Companies increasingly utilize online platforms to kick off innovation contests and thereby tap into the creative power of the crowd to generate new business models, drive innovativeness and enhance competitive advantage [1–4]. In such contests, the crowd easily generates hundreds and sometimes thousands of potentially promising ideas [5, 6] that are typically filtered by domain experts [6]. The complex decision making process, to pick the few most original, unique, useful, and elaborated ideas [7], commonly requires substantial amounts of resources [4]. Google received more than 150,000 ideas and 3,000 employees devoted their time to review the submissions to finally announce 16 winners[1]. Those who filter such large quantities of ideas are not only faced with the challenge of an exceeding cognitive load imposed by this complex task [8], but also by the issue of similar ideas occurring in substantial amounts [9].

In order to reduce cognitive load and to ease the idea selection process, organizations do not only rely on experts for evaluation, but also on small teams, the crowd or

[1] https://www.cnet.com/news/google-announces-project-10100-themes/

16th International Conference on Wirtschaftsinformatik,
March 2021, Essen, Germany

automated idea screening systems [10]. However, the crowd utilized in research tends to differ from the crowd relied upon in practice. In practice, the crowd often consists of internal employees or externals such as potential customers or the ideators themselves that can comment or vote on ideas on the ideation platform [5, 6]. In scientific research, the crowd commonly consist of anonymous crowd workers recruited via crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk) or Figure Eight (formerly known Crowdflower) [11–13], or University students [14, 15] in addition to small expert teams or an internal crowd. Both types of crowds, anonymous crowd workers and students, are used as participant's source in various fields of research [16]. However, the different crowd types also perform disparate tasks. Typical tasks on a crowd working platform are image tagging, relevance feedback or document labeling [17] as well as surveys administered by top researchers [16]. However, crowd platforms rarely offer tasks that require more time and cognitive effort such as idea selection tasks. This is in line with the literature stating that crowd workers deliver high quality work as long as the tasks are not effort-responsive [16]. Students on the other side, are considered unique in terms of their reflective thought [16] and are long accepted as participant source. Multiple studies exist that use students as a proxy for the crowd for a variety of tasks including idea selection [14, 15]. However, a problem remains: How to identify good quality work in idea selection? For classification problems or programming there usually exists one truly good answer, but in innovation contests, it would be very time-consuming and expensive to examine which idea is the best, because essentially, they would all need to be implemented. Hence, researchers developed quality control mechanisms such as attention checks or gold questions for which one truly correct answer exists [18–20].

This paper investigates how crowd types differ in their attention, information processing style and performance when accomplishing complex decision-making tasks such as idea selection. An online experiment with a crowd recruited from Amazon Mechanical Turk and a crowd of European undergraduate students was conducted.

2 Theoretical Background

2.1 Crowd Tasks

Crowdsourcing means bringing people in from outside the company and involving them in a creative, collaborative process [21]. Crowdsourcing has been gaining increasing interest, because the "wisdom of the crowd", the independent judgements of a large and diverse group of individuals, has been proven to be relatively accurate [22]. Following that, a wide variety of tasks with different levels of complexity have been passed over to the crowd. These tasks cover activities in all phases of the value chain including but not limited to crowd testing, funding, ideation, logistics, production, promotion and support [23]. Cognitively less demanding tasks such as data annotation, image tagging, accessing content on the web or finding information online [24] were shown to be completed pretty accurately by the crowd [e.g., 25]. However, complex tasks that require strenuous effort like creating content, generating or evaluating ideas provide mixed results [4]. While many studies show that the crowd is able to quickly

generate hundreds or thousands of ideas [5, 26], selection performance may not be considerably higher than chance [11, 12, 27, 28]. One reason is the high cognitive demand that is imposed by the task of comparing very similar ideas [26] and processing multiple idea attributes [29]. Another reason might be related to the characteristics of the crowd. Thus, to better understand this issue, this paper first investigates which types of crowd exist.

2.2 Crowd Types in Idea Selection

Specific tasks call for domain-specific or company internal knowledge, hence, companies do not only ask externals but also their employees to make suggestions. Consequently, the crowd can be distinguished into being either internal or external to the crowdsourcer [23]. In practice, the evaluation of ideas is done by three types of raters that are the crowd, a jury of experts, and self-assessments, which can also be used in combination [10, 30]. In research, the "crowd" is a widely used term and can refer to anonymous crowd workers from crowd platforms such as Amazon Mechanical Turk or FigureEight, but also a University student crowd, user crowd or an internal employee crowd. Student samples were used to compare different evaluation mechanisms [14, 31]. Related research suggests that students who are evaluating ideas based on a multi-criteria rating scales outperform students that were evaluating ideas in prediction markets [31]. Furthermore, a student sample was utilized to show that rating scales invoke higher ease of use than preference markets and that perceived ease of use mediates the role between the evaluation mechanism and decision quality [14]. Additionally, a study found that higher decomposition of information load (fewer ideas per screen) leads raters to acquire more information on ideas and to eliminate more ideas, which improved choice accuracy [28]. Online consumer panels were found to represent a better way to determine a "good" idea than are ratings by experts [33]. And significant agreement was found between theatre projects that were funded by the funding crowd and experts [34]. Anonymous crowd workers have been recruited, because a multitude of responses can be generated in a short time. The ratings for novelty of an anonymous crowd (MTurk) are highly correlated with those of experts [35]. The evaluations of an MTurk crowd were also used to develop an expertise prediction heuristic to automatically identify experts within the crowd [13]. Crowd workers of MTurk that evaluate sets with similar ideas have higher elimination performance and lower cognitive effort than those crowd workers that evaluated sets with random ideas [11]. Idea selection done by users was relatively successful when compared to expert assessments and even technically naïve users recruited from Amazon Mechanical Turk yielded satisficing results [36]. Contrary to previous studies of crowd evaluations for simple aesthetic tasks, one study also provides first evidence of the limitations of anonymous crowd evaluations (Crowdflower), and warns that crowd evaluations are not adept to the expert ratings when more complex submission such as business models are evaluated [12]. While crowds were frequently compared to experts, little is known about whether one crowd type might be better able at selecting high quality ideas than another. Hence, this research aims to evaluate differences in

anonymous crowds and student crowds regarding their information processing, attention and selection performance.

2.3 Information Processing

It is important to understand how raters process the ideas and decide on their quality to better deal with challenges related to the complex and effort intensive selection process. When making decisions, people engage in disparate types of cognitive processes that can be distinguished into intuition [37] and reasoning [38], also referred to as System 1 and System 2. System 1 represents intuition and denotes fast, automatic, and effortless information processing. System 2 represents reasoning, being a slow, controlled, and effortful information processing [39]. System 1 thinking consists of subsystems which include autonomous behaviors and domain-specific knowledge obtained through domain-general learning mechanisms [40]. When utilizing System 1 cognitive processes to make decisions, individuals tend to use shortcuts in their decision making [41] and adopt rules of thumb stored in their long-term memory to process information [42]. System 2 information processing makes use of the central working memory system [40]. When individuals engage in System 2 cognitive processes, all available options are objectively compared until a decision is made. Usually, individuals are expected to make decisions as objectively as possible, since rational decision making is supposed to lead to accurate choices and, thus, good decisions [43]. However, as the information processing capacity of a human cognitive system is limited, it is impossible to evaluate all possible outcomes [44, 45]. Hence, due to their limited rationality choices lose objectivity.

2.4 Attention and Quality Control in Crowdsourcing

Crowdsourcing platforms such as Amazon Mechanical Turk or Figure Eight allow to collect large amount of responses in a very short amount of time. Unfortunately, the process of verifying the quality of submitted results is not that easy and often workers take the chance to submit low quality work [17]. Hence, quality control is essential for requesters of the crowdsourced tasks and it comes in various forms. First, requesters rely on redundant task assignment and ask multiple crowd workers the same questions [17, 46]. Further, financial incentives such as performance-based payments are used to increase the quality of submissions [46]. Next, over time attention check questions or gold questions were developed, which are a small set of tasks for which the requester knows the correct answer and, thus, is able to directly assess the quality of the submission [18]. These questions should be unique for each task or study in order to reduce the probability for a crowd worker to be familiar with the attention check questions and hence, to increase their effectiveness [16]. One type of these attention checks are instructional manipulation checks (IMC), where participants demonstrate that they were reading and following the instructions [19]. IMCs typically consist of a text in which the participants are instructed to answer in a specific way to a question that is posted below. When a participant does not read the text, s/he would answer the question incorrectly and hence, would fail the IMC. Factual manipulation checks are

questions with an objective, matter-of-fact answer. The problem with factual manipulation checks is that participants can easily search the internet for the correct answer and they do so, if researchers do not intervene with the simple instruction to not look up the answers [16]. Another attention check is the affirmation form in which crowd workers indicate whether they paid attention and answered the questions honestly [47]. Keith et al. review crowd studies and identified that only 22.8% of the studies report on using attention checks, among which are direct, archival and statistical attention checks such as instructed items (e.g. "Please select strongly disagree, if you are paying attention.", bogus items (e.g., "My friends are all mermaids."), questions to recall information from the instructions or an article, or measuring the time spent on the task [48].

2.5 Research Model and Hypotheses Development

It is commonly noted that there are differences between various participant sources with respect to their attention, cognitive processing styles and task performance. The crowd in general was found to be a good proxy for experts' in idea evaluation [36]. This includes both, the student crowd as well as the anonymous crowd. However, one study found that crowd workers from Figure Eight were not as good as commonly assumed [12]. This is in line with the literature stating that crowd workers deliver high quality work as long as the tasks are not effort-responsive [16]. Students on the other side, are considered unique in terms of their reflective thought [16]. Hence, anonymous crowd workers are assumed to have lower selection performance than students.

H1: Crowd workers from anonymous crowd working platforms will have lower selection performance in terms of a) lower accuracy, b) higher false negative rate and c) higher false positive rate than a student crowd.

Crowd workers have learned to be attentive to specific types of questions such as attention questions. They tend to search for information that help them to quickly come to a decision as some of the crowd workers make a living of these short and often ill paid crowd task. Whereas students like to engage in cognitively demanding tasks as they also selected to enroll in a University program. Hence, the following hypotheses regarding the crowd types' cognitive load and information processing styles can be formulated:

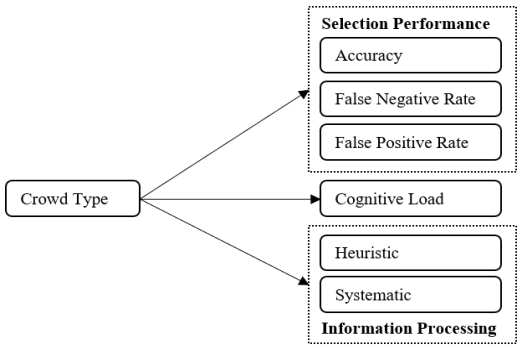

Figure 1. Research Model

H2: Crowd workers from anonymous crowd working platforms will have lower cognitive load than a student crowd.

H3: Crowd workers from anonymous crowd working platforms will process information a) more heuristically and b) less systematically than a student crowd.

Combining the arguments mentioned above, a research model is proposed that compares the relationships between two crowd types (anonymous crowd and student crowd) and their selection performance, cognition and information processing (see Figure 1).

3 Methodology

This study compares two different crowds, i.e., an anonymous crowd and a student crowd, with regards to their attentiveness, information processing styles and their resulting selection performance using a web-experiment consisting of a pre-survey, an idea selection task and a post-survey.

3.1 Idea Set

In the idea selection task, participants were presented with 35 ideas from the "Gratitude at the Workplace" Challenge hosted on openIDEO[2]. The contest was selected because the ideas covered a broad range of topics that did not require any technical or domain-specific knowledge. The ideas were accessible and easily comprehensible for individuals that have a basic understanding of appreciation and workplaces. The original ideas were adapted and shortened to control for the idea length and possible effects on the selection (e.g., shorter ideas are easier to comprehend and therefore selected). The ideas were randomly allocated to subsets. Ideas and subsets were allocated to participants in random sequence to control for order bias using the Smart Idea Allocation method [49]. Ideas were presented with their title, description and the number of likes they received on the platform.

3.2 Subjects

Data was collected from 284 crowd workers recruited from Amazon Mechanical Turk (using the platform cloudresearch.com) and 55 undergraduate students enrolled in an introductory course to Information Systems (IS) at a European University (via the online course forum). Participants that failed the reCaptcha on the first page (to identify bots or machines) or the first simple instructional attention check ("Click the radio button for strongly agree.") were excluded to ensure a representative sample. After eliminating all participants that failed at least one attention check question, 87 MTurks and 49 students remained. The reward consisted of a fix and a variable, performance-based payment as recommended for effort-responsive tasks [46]. While MTurks

[2] The original ideas of and information about the contest can be found on the following website: https://challenges.openideo.com/challenge/gratitude-in-the-workplace/brief

received 2.50 USD, students received 3.6 points as course credit for successful completion of the whole task as a fixed reward. The variable amount consisted of a bonus for every good idea they selected (+0.30 USD for MTurks and +0.3 points for students) minus a deduction for every bad idea they selected (-0.10 USD for MTurks and -0.1 points for students). The payment model for MTurks was chosen to comply with the minimum wage for the United States, as the expected duration to complete the task was about 20-30 minutes. The reward was special for both participant groups, while MTurks received an above average payment compared to other tasks on the platform, students had the chance to receive course credits. Participation was voluntary for students and MTurks. Furthermore, students had the opportunity to choose between two different tasks to receive course credit similar to MTurks who could move on to another Human Intelligence Task (HIT). Only MTurks that completed at least 100 HITs and had an approval rate of minimum 80% (i.e., 80% or more of that participant's previous submissions were approved by requesters) were allowed to participate in the task. MTurks were, with on average 38 years (SD = 10.8 years) about 16 years older than students that were on average 22 years old (SD = 2.9 years). Among the MTurks 56% indicated to be male, 43% female and 1% others; students indicated to be 45% male and 55% female. All participants graduated from high school. Additionally, the majority of MTurks (51.7%) and some students (4.1%) possess a Bachelor's degree. Undergraduate IS students are expected to have some basic understanding of human resources and workplace innovation. MTurks themselves have some form of employment relationship with the requesters of the HITs and more than 60% of the crowd workers in previous studies participate on MTurk to generate a second source of income [50]. Participants were also asked to rate to what extent they usually experience or express gratitude "while collaborating with colleagues", "by receiving or giving donations", "from your leader or as a leader", "via platforms and applications", "via e-mail", "during business trips and travels", "during meditation", "in or to specific groups of people (e.g., healthcare, farmers, police)", and "through handcrafted objects (e.g., handwritten notes, paintings, collages)" (7-point-Likert scale from 1 = "strongly disagree" to 7 = "strongly agree"). On average, MTurks and students indicated a level of experience with gratitude of 4.78 and 4.44 with a standard deviation of .98 and .72, respectively. Both crowd types more often experienced or expressed gratitude while collaborating with colleagues ($M_{\text{crowd worker}} = 5.38$, $M_{\text{students}} = 5.24$) and from their leader or as a leader ($M_{\text{crowd worker}} = 5.05$, $M_{\text{students}} = 5.29$). To conclude, students as well as MTurks should have sufficient experience with "Gratitude at the Workplace" to evaluate the ideas.

3.3 Experimental Procedure and Task Instructions

Once participants accepted the task on their specific platform (cloudresearch.com for MTurks and online course forum for students), they were redirected to the pre-survey. On the welcome screen, participants were informed about the task, the reward scheme and the approval criteria. Specifically, they were informed about the expected minimum work duration for the task to be 8 minutes with an average about 20-30 minutes. Furthermore, they were notified to pay attention to answer all attention questions

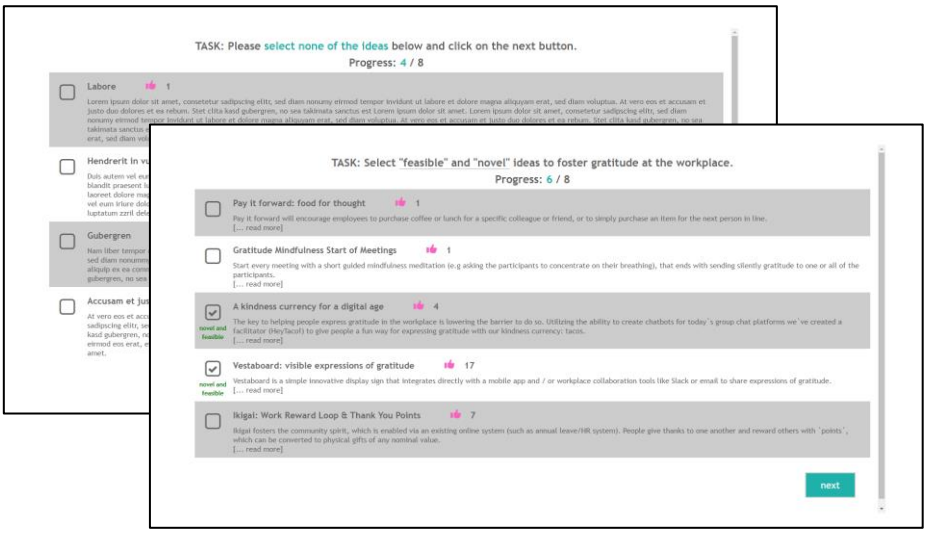

Figure 2. Screenshots of Idea Selection Platform

correctly to receive the fixed reward (see section 3.4 Attention Checks). Afterwards, participants answered some perception-based questions and were informed about the task setting: "Imagine you are a Human Resource (HR) Manager. The organization you work for wants to foster gratitude at the workplace. Research shows that too many people are feeling unappreciated and taken for granted at work. Gratitude strengthens our relationships, improves our health and motivates us. Hence, you organized an external innovation contest about gratitude at the workplace and received 39 ideas from the crowd. You know that you want to assess the ideas as objectively as possible and not according to your own preferences." Participants then selected categories of their interest and were further introduced to the selection environment: "Click the Select-Button if you deem an idea novel and feasible. Click the Read-more button to see the full idea description. You can select zero, one or multiple feasible and novel ideas from each set. The progress tracker bar shows you how far along you are in the task. Click the next button to get a new subset; there is no back button." The binary assessment can be understood as a holistic rating scale, which means that only one score with a single trait is collected [51]. The meaning of "feasible and novel" was further explained in order to guide the attention to relevant quality criteria: "An idea is feasible, if it can be easily implemented and is socially acceptable. An idea is novel, if it is new and original; not like anything seen before." Participants agreed that they have understood the task setting and the selection environment and were then directed to the selection platform. On each of the next seven screens (see Figure 2), four to seven ideas were presented where participants could check boxes to select feasible and novel ideas indicated by check mark and "novel and feasible". Note that after three screens four Latin dummy text ideas were presented as attention check. The experiment ended with a survey that collected perception-based variables and demographic data. During the task, the author

included seven different attention checks. When participants failed an attention check question they were notified and could not proceed with the task.

3.4 Measures and Operationalization

Performance Measures. The binary nature of the idea quality (low quality vs. high quality) allows to use performance metrics from the field of Information Retrieval (e.g., [11, 52]). The selection of each participant is compared to the gold standard in a confusion matrix (see Table 1). To assess selection performance in innovation contests, three particular measures are relevant, which are the selection accuracy, false negative rate and false positive rate. Selection accuracy (ACC) is the proportion of all correct predictions (true positives and true negatives) divided by all predictions [53]. The more ideas are correctly classified as being high or low quality, the higher is the measure. As contest managers might be concerned with fear of missing out [54], the false negative rate (FNR), which is the fraction of ideas that have been incorrectly classified as being low quality [53], should be low. Furthermore, having low quality ideas in the consideration set increases subsequent evaluation effort, which is at best avoided [55]. Hence, the false positive rate (FPR), which represents the fraction of ideas that have been incorrectly classified as being high quality [53], should be low.

In scientific research, the gold standard is usually established through multiple raters with domain knowledge (e.g., [9, 14]). Hence, seven Human Resources experts were asked to rate the ideas according to their feasibility and novelty. Based on the experts aggregated assessments, six ideas were defined as high quality ideas and the remaining 29 ideas as low quality. The ratio of 17% good ideas is in line with the literature, which states that 10-30% of user generated ideas are of high quality [31].

Attention Checks. Seven different attention check questions were included. Two simple instructional attention checks were included in the pre-survey and in the post-survey, where participants were asked to "Click the radio button for strongly agree/disagree." A memory attention check question was included that consisted of two question, one was asked in the pre-survey and one in the post-survey. Participants were supposed to select the same answers in both questions. In the first multiple-choice question, they were notified to remember their choice for a later stage of the task. Specifically, participants were asked "What would you like to have for your birthday?"

Table 1. Confusion Matrix and Performance Measures

		Gold Standard	
		High quality	Low quality
Prediction of participant	High quality	True positive (TP)	False positive (FP)
	Low quality	False negative (FN)	True negative (TN)
Performance Measures	Accuracy:	$ACC = \frac{\sum TP + \sum TN}{\sum TP + \sum FP + \sum TN + \sum FN}$	
	False Negative Rate:	$FNR = \frac{\sum FN}{\sum TP + \sum FN}$	
	False Positive Rate:	$FPR = \frac{\sum FP}{\sum FP + \sum TN}$	

and could choose among “Birthday cake”, “Health for family and friends” and/ or “Laptop”. Another memory attention check, this time without prompting, was included after the idea selection task in the post-survey and asked participants to “Please select those ideas that you have been presented with in the previous idea selection task.” Five options were available in this idea recognition task from which four were self-invented ideas about Virtual Reality apps that were not presented before and one option said “None of the above”. Participants were supposed to select “None of the above” as the other ideas were not related to the “Gratitude at the Workplace” topic of the contest. Furthermore, a task-related attention check was included during the idea selection task. After completing the first half of idea sets, participants were presented with four Latin dummy text ideas. One dummy text idea title was “Hendrerit in vulptate” and the corresponding short description “Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros.” As these ideas did not have any meaning, participants were supposed to not select any of the ideas. The last attention check question for both groups was the completion time, which was expected to be more than eight minutes. MTurks were also asked to submit their individual completion code that they received at the end of the survey. The author refrained from including attention checks that test factual knowledge as it was shown that crowd workers would use the internet to solve these questions (e.g., [16]).

Cognition and Information Processing Styles. All measurements to operationalize our research variables are based on previously validated operationalizations and have been adapted to the context of our study. Four items were used to deduce Extraneous Cognitive Load (ECL), that is the cognitive load imposed by the task presentation [56]. Finally, the items for heuristic (HEU) and systematic (SYS) information processing were adapted from Novak and Hoffman’s experiential and rational situation-specific thinking style scales, defined as the experiential or rational thinking style or momentary thinking orientation adopted by a consumer in a specific situation. [57]. See Table 1 in Online Appendix[3] for the adapted survey items. All items were measured on a 7-point-Likert scale (from 1 = “strongly disagree” to 7 = “strongly agree”).

4 Data Analysis and Results

This study investigates the differences between an anonymous crowd and a student crowd in terms of attention, information processing styles and selection performance when selecting ideas for an innovation contest.

Statistical Assumptions. First, data was checked against violation of statistical assumptions for analysis of variance. For normal distribution, data was visually inspected with Q-Q plots, boxplots and histograms as well as skewness and kurtosis statistics for each group. For the selection performance measures Accuracy, FNR and FPR and the perception-based variables systematic processing and heuristic processing, boxplots and histograms indicated a close to bell curve; skewness and kurtosis are mostly close to 0. Homogeneity of variance was tested with Levene’s statistics, which

[3] https://tinyurl.com/y2xl2rtv

turned out to be satisfactory for most variables (ACC: F = 1.784, p = .184; FNR: F = 0.943, p = .333; FPR: F = 0.639, p = .425; SYS: F = 2.486, p = .117; HEU F = .130, p = .719) as p-values should be greater than .05 [58]. For ECL, Levene's test was significant and hence, the assumptions of homogeneity of variance did not hold [58]. To conclude, the data are sufficiently normally distributed and homogeneity of variance is satisfactory, hence, multiple analysis of variance is conducted.

Reliability and Validity. To test convergent and discriminant validity, exploratory factor analysis with Promax (kappa = 4) rotation was performed. Most of the items of the perception-based constructs loaded well on three of the resulting four factor solutions with factor loadings higher than .5. One item (SYS7) loaded on the fourth additional factor. However, this was the only one and hence, it was kept for analysis. Cross-loadings were low and MSA-values higher than .5. All these values exceeded the recommended thresholds [59] and therefore convergent and discriminant validity are deemed satisfactory. Reliability analyses with Cronbach's Alpha were performed for extraneous cognitive load (Cronbach's α = .911), heuristic processing (Cronbach's α = .799) and rational processing (α = .762). All perception-based constructs reached the recommended threshold of .7 [59].

4.1 Attention

To start with, 284 MTurks and 55 students passed the first (reCaptcha) and second ("Click strongly agree") attention check (see Table 2). The task-related attention check followed and only 37.0% of MTurks answered it correctly, whereas 90.9% of the students were able to correctly not select any of the Latin dummy text ideas. From the remaining 105 MTurks and 50 students, 101 MTurks correctly answered the second simple instructional attention check ("Click strongly disagree") while all students followed that instruction correctly. The memory attention check with prompting (birthday present) was answered correctly by 99 of the remaining MTurks and again all students remembered their choice from the multiple-choice question from the pre-survey correctly. Whereas the memory attention check without prompting (idea recognition test) was answered correctly by 88 of the remaining MTurks and by 49 of the remaining students. The expected completion time of at least eight minutes was met

Table 2. Exclusion of Participants Based on Attention Checks

	MTurks	Students	Total
Participants	284	55	339
Excluded from analysis	197	6	203
Failed task related AC	179	5	184
Failed simple instructional AC	4	0	4
Failed memory AC with prompting	2	0	2
Failed memory AC without prompting	11	1	12
Failed completion time	1	0	1
Included in analysis (Success Rate)	87 (30.6%)	49 (89.0%)	136 (40.1%)

by 87 of the remaining MTurks and 49 of the remaining students. The average completion time of the remaining MTurks is 23:08 minutes and is significantly shorter than the completion time of the students with 45:31 minutes, $F(1, 134) = 61.243$, $p < .001$, partial $\eta^2 = .314$. In total, 89.0% of the students and only 30.6% of the MTurks were able to successfully complete the complex selection task and all attention checks, indicating that students are more attentive to complex decision-making tasks.

Attention and Selection Performance. As crowd workers seem to be rather inattentive to the attention checks, the author analyzed whether there are differences in selection performance over time, i.e., before and after the task-related attention check. The performance measures accuracy, false negative rate and false positive rate were calculated for the first half and for the second half of idea sets. A within-subject MANOVA of all participants (N = 339) reveals statistically significant differences for all three performance measures over time, Wilks $\lambda = 0.769$, $F(5, 130) = 7.822$, $p < .001$. Specifically, selection accuracy was on average 55.4% for the first half and for the second half with 58.5% significantly higher ($F(1, 338) = 19.040$, $p < .005$). Furthermore, the false positive rate was 41.5% for the first half and significantly lower for the second half with 37.3% ($F(1, 338) = 19.040$, $p < .005$). These results indicate that the task-related attention check increased selection performance.

4.2 Selection Performance, Cognition and Information Processing

To examine the effect of the crowd type on selection performance, cognitive load and information processing styles, the author performed multiple analyses of variance. The crowd type had a significant effect on all tested variables, Wilks $\lambda = 0.769$, $F(3, 336) = 12.760$, $p < .001$, partial $\eta^2 = .231$. The mean values, standard deviation and median for each crowd type and each variable can be found in Table 3. The results of the MANOVA are presented in Table 4. The anonymous crowd worker have a lower selection accuracy (57.8%), indicating that they are not as good as the student crowd (64.7%) at identifying the truly good and truly bad ideas as suggested by the gold standard ($F(1, 134) = 9.529$, $p < .005$, partial $\eta^2 = .066$). While no significant effect was found for the false negative rate, MTurks have a higher false positive rate (38.3%) than students (29.4%) ($F(1, 134) = 9.105$, $p < .005$, partial $\eta^2 = .064$), which means that MTurks define more ideas as high quality even though they are categorized as low quality by the experts, inducing higher subsequent evaluation effort.

The anonymous crowd experiences significantly lower extraneous cognitive load (Mean = 3.22) than the student crowd (Mean = 4.20) ($F(1, 134) = 15.034$, $p < .005$, partial $\eta^2 = .101$). With regards to information processing, MTurks reports significantly higher values for heuristic processing (Mean = 5.15) than the students (Mean = 4.61) ($F(1, 134) = 10.322$, $p < .005$, partial $\eta^2 = .072$). Interestingly, MTurks simultaneously report higher values for systematic processing (Mean = 5.29) than the students (Mean = 4.83) as well ($F(1, 134) = 10.727$, $p < .005$, partial $\eta^2 = .074$).

Due to the surprising finding that MTurks also outperformed students in terms of systematic processing, the author tested the extent of systematic processing with behavioral data gathered on the selection platform. Participants could click on the „read more“ button to read the full idea description, which is an indicator of how much

information was acquired to make the decision whether or not to select an idea. Hence, the variable information acquisition is the sum of clicks on the "read more" button. An ad-hoc analysis revealed that MTurks clicked on the read more button on average 20.1 times and students 26.0 times. This difference in information acquisition between MTurks and students was found to be significant, $F(1, 134) = 13.515$, $p = .000$, partial $\eta^2 = .092$. Interestingly, MTurks reported that they systematically processed the ideas, but they acquired less information about the idea than the students.

Table 3. Descriptive Statistics for Performance Measures, Cognition and Information Processing

	ACC		FNR		FPR		ECL		HEU		SYS	
	C	S	C	S	C	S	C	S	C	S	C	S
N	87	49	87	49	87	49	87	49	87	49	87	49
M	.578	.647	.609	.639	.383	.294	3.22	4.20	5.15	4.61	5.29	4.83
SD	.131	.111	.252	.234	.174	.149	1.56	1.06	0.95	0.95	0.85	0.67
Mdn	.600	.629	.667	.667	.345	.278	3.00	4.00	5.20	4.80	5.43	4.71

M = Mean, SD = Standard Deviation, Mdn = Median, C = Crowd, S = Student

Table 4. MANOVA for Crowd Type

Source	DF	Mean square	F	p-value	partial η^2
MANOVA Dependent variable: *Elimination accuracy*					
Treatment	1	0.148	9.529	**.002**	.066
Error	134	0.016			
MANOVA Dependent variable: *FNR*					
Treatment	1	0.029	0.474	.493	.004
Error	134	0.061			
MANOVA Dependent variable: *FPR*					
Treatment	1	0.249	9.105	**.003**	.064
Error	134	0.027			
MANOVA Dependent variable: *Extraneous Cognitive Load*					
Treatment	1	29.788	15.034	**.000**	.101
Error	134	1.981			
MANOVA Dependent variable: *Heuristic Processing*					
Treatment	1	9.340	10.322	**.002**	.072
Error	134	0.905			
MANOVA Dependent variable: *Systematic Processing*					
Treatment	1	6.662	10.727	**.001**	.074
Error	134	0.621			

5 Conclusion

This study compares two different crowds, i.e., an anonymous crowd and a student crowd, with regards to their attentiveness, information processing styles and their

selection performance using a web-experiment. It was found that crowd workers recruited from Amazon Mechanical Turk have lower selection performance in terms of lower selection accuracy and higher false positive rate. Indicating that the student crowd is better at identifying high quality and low quality ideas correctly and produces less subsequent evaluation effort as fewer low quality ideas are included in the set for further consideration. Furthermore, MTurks experience lower extraneous cognitive load as they are more familiar with crowd tasks than undergraduate students from the Information Systems discipline. MTurks reported to process information more heuristically than students. Surprisingly, they also outperformed students in terms of systematic processing. Even though MTurks indicate to process information in depth, an ad-hoc analysis of their click behavior revealed that they acquire less information about the ideas. This study expands our understanding of two crowd types, examines their suitability for complex decision-making tasks and offers three main contributions. First, the IS student crowd selects ideas more accurately and with a lower false positive rate than the anonymous MTurk crowd. Second, this study confirms that crowd types process information differently in terms of heuristic and systematic processing as well as in terms of their actual processing behavior. Third, this study also provides a methodological contribution as it explores diverse attention checks and finds that using a task-related attention check increases selection performance of the crowd.

Like any other study, this study has its limitations, which, in turn, opens the door for future research. First, the crowd reported high levels of heuristic and systematic processing, which could not yet be fully explained. One attempted explanation could be that processing information, independent of whether heuristically or systematically, is socially desirable. Furthermore, heuristic and systematic processing are subjective perception variables and hence, do not necessarily reflect the participants' behavior. While the inclusion of mouse tracking behavior acts as a means to validate the information processing style, it does not yet suffice and further hard data would be desirable. Future research could examine potential biases and eye tracking could expand the existing database to better understand the crowds' information processing. Second, while this paper demonstrates that the student crowd performs better than the MTurks, our understanding of why is limited to students being more attentive. Future research could aim at identifying causal mechanisms that explain this effect. Third, while this study included only two external crowd types, namely undergraduate IS students and MTurks, future research could include contrasting crowds to enhance generalizability. An internal employee crowd, students from another discipline or anonymous crowd workers from crowd platform with a focus on more complex tasks might perform better in selecting ideas from a "Gratitude at the Workplace" contest. While all participants are expected to have a general understanding of human resources and workplace innovation, little is known about the participants' experience with the complex task of selecting good ideas from an innovation contest. Finally, students and MTurks received a different reward. MTurks received a financial reward whereas students received course credits, which might have had an impact on their motivation to accurately perform the task. Future research could consider the same incentive to rule out that there is an effect on information processing, attention and selection performance.

6 Acknowledgments

The research leading to the presented results was funded by the Austrian Science Fund (FWF): P 29765.

References

1. Chesbrough, H.W.: The Era of Open Innovation. MIT Sloan Manag. Rev. 127 (2003).
2. Du Plessis, M.: The role of knowledge management in innovation. J. Knowl. Manag. 11 (2007).
3. Gassmann, O., Enkel, E.: Towards a theory of open innovation: three core process archetypes. R&D Manag. Conf. (2004).
4. Nagar, Y., Boer, P. de, Garcia, A.C.B.: Accelerating the Review of Complex Intellectual Artifacts in Crowdsourced Innovation Challenges. In: 37th International Conference on Information Systems (2016).
5. Bjelland, O.M., Wood, R.C.: An Inside View of IBM's 'Innovation Jam.' MIT Sloan Manag. Rev. 50, (2008).
6. Jouret, G.: Inside Cisco' s Search for the Next Big Idea. Harv. Bus. Rev. 87, 43–45 (2009).
7. Dean, D.L., Hender, J.M., Rodgers, T.L., Santanen, E.L.: Identifying good ideas: constructs and scales for idea evaluation. J. Assoc. Inf. Syst. 7, (2006).
8. Sweller, J.: Cognitive load during problem solving: Effects on learning. Cogn. Sci. 12, (1988).
9. Kornish, L.J., Ulrich, K.T.: Opportunity Spaces in Innovation: Empirical Analysis of Large Samples of Ideas. Manage. Sci. 57, (2011).
10. Merz, A.: Mechanisms to Select Ideas in Crowdsourced Innovation Contests - A Systematic Literature Review and Research Agenda. In: European Conference on Information Systems (2018).
11. Banken, V., Seeber, I., Maier, R.: Comparing Pineapples with Lilikois: An Experimental Analysis of the Effects of Idea Similarity on Evaluation Performance in Innovation Contests. In: 52nd Hawaii International Conference on System Sciences (2019).
12. Görzen, T., Kundisch, D.: Can the Crowd Substitute Experts in Evaluating Creative Jobs? An Experimental Study Using Business Models. In: 24th European Conference on Information Systems (2016).
13. Burnap, A., Gerth, R., Gonzalez, R., Papalambros, P.Y.: Identifying experts in the crowd for evaluation of engineering designs. J. Eng. Des. 28, (2017).
14. Blohm, I., Riedl, C., Füller, J., Leimeister, J.M.: Rate or Trade? Identifying Winning Ideas in Open Idea Sourcing. Inf. Syst. Res. 27, (2016).
15. Riedl, C., Blohm, I., Leimeister, J.M., Krcmar, H.: Rating scales for collective intelligence in innovation communities: Why quick and easy decision making does not get it righ. In: 31st International Conference on Information Systems. (2010).
16. Goodman, J.K., Cryder, C.E., Cheema, A.: Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. J. Behav. Decis. Mak. 26, (2013).
17. Ipeirotis, P.G., Provost, F., Wang, J.: Quality management on Amazon Mechanical Turk. Work. Proc. - Hum. Comput. Work. 2010, (2010).

18. Checco, A., Bates, J., Demartini, G.: Adversarial Attacks on Crowdsourcing Quality Control. J. Artif. Intell. Res. 67, (2020).
19. Oppenheimer, D.M., Meyvis, T., Davidenko, N.: Instructional manipulation checks: Detecting satisficing to increase statistical power. J. Exp. Soc. Psychol. 45, (2009).
20. Hauser, D.J., Schwarz, N.: Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. Behav. Res. Methods. 48, (2016).
21. Howe, J.: The Rise of Crowdsourcing. Wired Mag. 14, (2006).
22. Surowiecki, J.: The Wisdom of the Crowds. Anchor Books, New York (2005).
23. Durward, D., Blohm, I., Leimeister, J.M.: Crowd Work. Bus. Inf. Syst. Eng. 58, (2016).
24. Difallah, D.E., Catasta, M., Demartini, G., Ipeirotis, P.G., Cudré-Mauroux, P.: The Dynamics of Micro-Task Crowdsourcing. In: 24th International Conference on World Wide Web - WWW '15 Companion. (2015).
25. Bentivogli, L., Federico, M., Moretti, G., Paul, M.: Getting Expert Quality from the Crowd for Machine Translation Evaluation. In: 13th Machine Translation Summit. (2011).
26. Di Gangi, P., Wasko, M., Hooker, R.: Getting customers' ideas to work for you: Learning from Dell how to succeed with online user innovation communities. MIS Q. Exec. 9, (2010).
27. Rietzschel, E., Nijstad, B., Stroebe, W.: Productivity is not enough: A comparison of interactive and nominal brainstorming groups on idea generation and selection. J. Exp. Soc. Psychol. 42, (2006).
28. Santiago Walser, R., Seeber, I., Maier, R.: The fewer, the better? Effects of decomposition of information load on the decision making process and outcome in idea selection. 27th Eur. Conf. Inf. Syst. (2020).
29. Hoornaert, S., Ballings, M., Malthouse, E.C., Van den Poel, D.: Identifying New Product Ideas: Waiting for the Wisdom of the Crowd or Screening Ideas in Real Time. J. Prod. Innov. Manag. 34, (2017).
30. Bullinger, A.C., Moeslein, K.: Innovation Contests – Where are we? In: AMCIS 2010 Proceedings. Americas Conference on Information Systems (2010).
31. Blohm, I., Riedl, C., Leimeister, J.M., Krcmar, H.: Idea Evaluation Mechanisms for Collective Intelligence in Open Innovation Communities: Do Traders Outperform Raters? In: 32nd International Conference on Information Systems. (2011).
32. Wibmer, A., Wiedmann, F.M., Seeber, I., Maier, R.: Why less is more: An Eye tracking study on idea presentation and attribute attendance in idea selection. In: 27th European Conference on Information Systems (2019).
33. Kornish, L.J., Ulrich, K.T.: The Importance of the Raw Idea in Innovation: Testing the Sow's Ear Hypothesis. J. Mark. Res. 51, (2014).
34. Mollick, E.R., Nanda, R.: Wisdom or Madness? Comparing Crowds with Expert Evaluation in Funding the Arts. Manage. Sci. 62, (2016).
35. Kudrowitz, B.M., Wallace, D.: Assessing the quality of ideas from prolific, early-stage product ideation. J. Eng. Des. 24, (2013).
36. Magnusson, P.R., Wästlund, E., Netz, J.: Exploring Users' Appropriateness as a Proxy for Experts When Screening New Product/Service Ideas. J. Prod. Innov. Manag. 33, (2016).
37. Magnusson, P.R., Netz, J., Wästlund, E.: Exploring holistic intuitive idea screening in the light of formal criteria. Technovation. 34, (2014).
38. Riedl, C., Blohm, I., Leimeister, J.M., Krcmar, H.: Rating Scales for Collective Intelligence in Innovation Communities: Why Quick and Easy Decision Making Does Not Get it Right.

In: 31st Interational Conference on Information Systems. (2010).
39. Kahneman, D.: A Perspective on Judgment and Choice. Am. Psychol. 58, (2003).
40. Evans, J.: In two minds: Dual-process accounts of reasoning. Trends Cogn. Sci. 7, (2003).
41. Croskerry, P., Singhal, G., Mamede, S.: Cognitive debiasing 1: Origins of bias and theory of debiasing. BMJ Qual. Saf. 22 (2013).
42. Jahn, G., Chemnitz, D., Renkewitz, F., Kunze, S.: Heuristics in Multi-attribute Decision Making: Effects of Representation Format. CogSci. (2007).
43. Sadler-Smith, E., Shefy, E.: The Intuitive Executive: Understanding and Applying "Gut Feel" in Decision-Making. Acad. Manag. Exec. 18, (2004).
44. Simon, H.: Rational Choice and the Structure of the Environment. Psychol. Rev. 63, (1956).
45. Miller, G.A.: The Magical Number Seven, Plus or Minus Two: Some Limites on out Capacity for Processing Information. Psychol. Rev. 65, (1956).
46. Ho, C.-J., Slivkins, A., Suri, S., Wortman Vaughan, J.: Incentivizing high quality crowdwork. In: International World Wide Web Conference Committee. (2015).
47. Rouse, S.: A reliability analysis of Mechanical Turk data. Comput. Human Behav. 43, (2015).
48. Keith, M.G., Tay, L., Harms, P.D.: Systems perspective of amazon mechanical turk for organizational research: Review and recommendations. Front. Psychol. 8, (2017).
49. Banken, V., Ilmer, Q., Seeber, I., Haeussler, S.: A method for Smart Idea Allocation in crowd-based idea selection. Decis. Support Syst. 124, (2019).
50. Ipeirotis, P.: Demographics of Mechanical Turk. , New York (2010).
51. Harsch, C., Martin, G.: Comparing holistic and analytic scoring methods: Issues of validity and reliability. Assess. Educ. Princ. Policy Pract. 20, (2013).
52. Walter, T.P., Back, A.: A Text Mining Approach to Evaluate Submissions to Crowdsourcing Contests. In: 46th Hawaii International Conference on System Sciences. (2013).
53. Metz, C.E.: Basic Principles of ROC Analysis. Semin. Nucl. Med. 8, October (1978).
54. Sarigianni, C., Banken, V., Santiago Walser, R., Wibmer, A., Wiedmann, F., Seeber, I.: Innovation Contests: How to Design for Successful Idea Selection. In: 53rd Hawaii International Conference on System Sciences. (2020).
55. Riedl, C., Blohm, I., Leimeister, J.M., Krcmar, H.: The Effect of Rating Scales on Decision Quality and User Attitudes in Online Innovation Communities. Int. J. Electron. Commer. 17, (2013).
56. Chang, C.C., Liang, C., Chou, P.N., Lin, G.Y.: Is game-based learning better in flow experience and various types of cognitive load than non-game-based learning? Perspective from multimedia and media richness. Comput. Human Behav. 71, (2017).
57. Novak, T.P., Hoffman, D.L.: The fit of thinking style and situation: New measures of situation-specific experiential and rational cognition. J. Consum. Res. 36, (2009).
58. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E.: Multivariate data analysis; A global perspective. (2010).
59. Nunnally, J.C.: Psychometric Theory. McGraw-Hill, New York (1978).