Computers in Human Behavior 49 (2015) 213–219

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Research Report

Quanty: An online game for eliciting the wisdom of the crowd

Wahida Chowdhury [a,*], Christopher Burt [b], Ahmad Aakkaoui [b], Jim Davies [a]

[a] Institute of Cognitive Science, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada
[b] School of Information Technology, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada

ARTICLE INFO

Article history:

Keywords:
Human computation
Wisdom of the crowd
Games
Quantitative magnitudes

ABSTRACT

Quanty is an online game that anonymously pairs players to estimate distances, weights, sizes, frequencies and such from photographs. The degree to which players agree determines the number of points they receive. We hypothesized that this game would generate more accurate aggregated estimates than would singular estimates by exploiting the wisdom of the crowd. Ninety-six participants (50 in group 1 using the metric system, and 46 in group 2 using the non-metric system) estimated height, weight, and distance of various objects; aggregated estimates of each group were more likely to approach accurate answers than were individual estimates, especially when the aggregates were calculated using medians and median absolute deviations. Also, the majority of participants thought that the game was as fun as the popular game Tetris. The results suggest that Quanty can be used to improve the judgment accuracy of professionals.

1. Why we need quantitative data about everyday things

Professionals increasingly rely on computer software to assist them in their practices. For example, doctors often rely on computers to make a diagnosis (see Gilbert & Lemke, 2014), and real estate agents might use computers to estimate the cost of properties (see Von Zur Gathen & Gerhard, 2013). Much of the software they employ also requires input of information. For example, a doctor might be asked to input a patient's weight, or a real estate agent might be asked to input the size of a house. Although professionals can often retrieve the requested information in existing databases, sometime they must resort to estimates based on their own mental models.

Suppose an engineer wants to investigate whether a land is good for installing wireless antennas; she is required to estimate the size of potholes from the photographs of the land. A particular study might be available by a company or a scientist that measured the approximate number of potholes in a given city, but such data are scattered across computers. Also many measures, such as the size of one specific pothole, seem too trivial to collect. The engineer in question might end up guessing, but how can she improve the accuracy of guessed estimates?

One possibility is to utilize *the wisdom of the crowd*, that is, to collect and aggregate estimates of several people. In fact, the Internet is a common platform that utilizes the wisdom of the crowd to gather estimates of things, such as oil, gas, or stock prices in the next month, the likelihood of a terrorist attack, the width of a brain tumor, or who would win the next presidential election. The first purpose of the present study was to design an online game that could be motivating and fun for people to provide their best estimates of various objects shown in photographs, and the second purpose was to analyze how the estimates could be aggregated to elicit the wisdom of the crowd and provide the most accurate answers.

1.1. How wise is the wisdom of the crowd?

The very idea about using the wisdom of the crowd came when a renowned British elitist, Francis Galton, discovered in 1906 that the median guess (1207 lb) of a randomly selected 800 uneducated commoners, about the weight of an about-to-be slaughtered ox at a village fair, was within 1% of the ox's true weight, 1198 lb (Surowiecki, 2005). Since then, many researchers attempted to utilize the wisdom of the crowd to solve complex problems. For example, Nickerson et al. (2009) investigated whether or not crowdsourcing, that is gathering people's independent judgments, could effectively match solutions to problems. The researchers asked students and Internet users to match several problem situations (for example, job search) with likely solutions (for example, social networks), and found that the aggregated wisdom of their crowd was as good as experts. Similarly, Steyvers, Lee, Miller, and Hemmer (2009) elicited the wisdom of the crowd by asking a group of people to order problems, such as listing chronologically the US Presidents, or ranking cities according to their populations.

* Corresponding author. Tel.: +1 613 883 6248.
E-mail address: wahida_chowdhury@carleton.ca (W. Chowdhury).

http://dx.doi.org/10.1016/j.chb.2015.03.004
0747-5632/

Surowiecki (2005) outlines four characteristics of the crowd that generates close to accurate wisdom:

1. *Diversity:* The crowd is formed of diverse people with their own knowledge or bias.
2. *Independence:* People in the crowd provide their judgments independently of others.
3. *Decentralization:* People in the crowd are allowed to draw on their own knowledge.
4. *Aggregation:* A mechanism exists to aggregate individual judgments into a collective decision.

If these four characteristics of the crowd are met, people's estimates can be modeled as a probability distribution with a central tendency near the true value of the quantity to be estimated. Large numbers of people are often not needed to improve estimates. For example, in the popular game show *Who Wants To be a Millionaire*, studio audiences are often enough to derive more accurate answers than the answer of an assumed expert. Also, one study showed that the average value of independent guesses of as few as three people produced estimates reliably closer to reality than did just one (Lee & Shi, 2010).

Surowiecki proposes several kinds of problems that could be solved by such a crowd, but for our purposes we were interested in how best the crowd could solve cognitive problems, where each person provides independent estimates of things. But how can we gather diverse people at one place and collect independent and decentralized estimates from them? How can we motivate each person in the crowd to provide their best estimates, and at the same time, make sure that the experience is fun so that they spend time in providing estimates in the first place? One possibility is to construct and deploy an online, serious game.

1.2. Can serious games elicit the wisdom of the crowd?

Serious games refer to those games that are fun, and at the same time, "educational, engaging, impactful, meaningful, and purposeful" (Ritterfeld, Cody, & Vorderer, 2009, p. 3). Ritterfeld et al. reported that the five factors of games that are consistently found to influence the experience of fun are "overall game design, visual representation, audio representation, complexity, and diversity, and control" (p. 36). Most serious games however involve educational or skill training; the serious games that involve coordination or cooperation of players to solve a social problem is less prominent (Ritterfeld et al., 2009).

Von Ahn and Dabbish (2008) attempted social, serious games simply by asking regular Internet users to play an online game that had no external rewards, but were presumably intrinsically entertaining. The authors were able to gather huge amounts of data (labels for images) from diverse Internet users. Others picked up this idea and attempted to elicit the wisdom of the crowd simply by asking amateur Internet users, for example, to map the world (www.openstreetmap.de), or to create geospatial data by playing location-based games (Matyas, Kiefer, Schlieder, & Kleyer, 2011). Instead of explicitly performing quality checks, these games seemed to motivate people to play, by taking into account others decisions to solve a common problem. So can we deploy an online game to generate independent estimates of different objects?

Inspired by Von Ahn and Dabbish's (2008) Games With a Purpose, especially his ESP and Peekaboom games (see also Von Ahn, Liu, & Blum, 2006), we designed a serious, social game called *Quanty* as an enjoyable and competitive way to collect quantitative estimates of physical properties of objects in photographs. Previous studies show that Online Games are a fast way to elicit user's preferences while making it hard for the users to cheat (Hacker & Von Ahn, 2009). Quanty is deployed online, and to further ensure that users do not cheat, Quanty randomly pairs players to estimate quantities such as height and weight; If only one player is available in the game website, Quanty waits for a second player without starting the game. When two players are successfully paired, Quanty starts the game and instructs players that the closer their estimates are to the guesses of their partners, the higher the scores they will receive. Thus the players are assumed to be motivated to produce an estimate in a competitive situation. After players complete the game, estimates are statistically aggregated.

1.3. How to aggregate individual answers

The next question we considered is how to aggregate estimates of a diverse crowd so that the aggregated estimates could be close to accuracy. Most previous studies calculated simple mean or median of estimates (for example, how many jelly beans in a jar); Yi, Steyvers, Lee, and Dry (2012) developed aggregation methods that either combined individuals' judgments into a grand judgment, or identified judgments that is most similar to other individual judgments. Leys, Ley, Klein, Bernard, and Licata (2013) showed when aggregating judgments from a diverse group, it is best to calculate median absolute deviations of estimates (averaging after removing the estimates two standard deviations away from the median). Which method of aggregating is better?

We investigated which of four methods of statistical aggregation might produce the most accurate estimate: (1) simple averaging of all estimated values; (2) averaging estimated values after removing outliers (estimates two standard deviations away from the mean); (3) calculating the simple median estimate; and (4) Calculating Median Absolute Deviation or MAD.

2. Quanty: game design

Quanty is a web-based game, playable at the following URL: http://www.quantygame.com/. See Appendix A for the starting page with the 'how to play' instructions shown to a player. When clicked to start playing, players are randomly paired into teams of two. Players are matched into pairs randomly for the sake of anonymity. A given player does not know who his or her teammate is, nor can teammates communicate, so that they cannot cheat by agreeing on which numbers to enter. For example, if two players knew they would be partners, then they could agree to put the same number in for every question, giving them high scores and worse, bad data for professionals.

Both players are presented with a photograph. One or more objects in the photograph are outlined, each with a different color. The players are prompted with a question regarding some quantitative magnitude associated with the shown object(s), such as "how much does the radiator weigh?" or "what is the distance between the sidewalk to the building?" See Appendix B for an example question with the screenshot. Each player inputs his or her answer. The closer the two answers are, the more points both players get. This mechanism not only gives the game a way to generate a score, but also encourages the players to be as accurate as possible to score higher. This completes a single round. Then a new photograph is shown, and another round begins. Players continue with as many rounds as they can get through before the timer runs out after 3 min.

2.1. Photographs

Quanty in its current version uses a subset of the photographs and data freely available on the website LabelMe (Russell, Torralba, Murphy, & Freeman, 2005), (http://labelme.csail.mit.edu/) which offers an activity where users view photographs, click to trace the outline of an object in the photographs, and label that

outline with a name. Thus Quanty collects a series of photographs with outline objects. In each round, Quanty randomly chooses a new photograph, a new number of objects to ask about (1–3), and then an appropriate question about the objects chosen.

The photographs are presented in black and white, except for the target object or objects, which are in color and surrounded by an outline. If the software is asking about more than one object (as when asking about the distance between two things) then each object gets a different colored outline. In the text of the question, the name of the object has the same color as the outline. This makes the question easier to understand, and is necessary when the two distinct areas of the photographs have the same label (e.g., "Is the leaf above the leaf?" where two distinct leaves are highlighted in corresponding colors in the text, and in the photograph). In Appendix B, for example, the wheel rim image outline and text is yellow.

2.2. Questions

In each round, the questions are chosen randomly. First, a random photograph is chosen from the database. If there are only two outlined objects in the photograph, then there is a 50% chance that the question will be about one of them, and a 50% chance that it will be about two of them. If there are three or more, then there is a 50% chance it will choose a question about one of the objects, a 25% chance it will choose a question about two objects, and a 25% chance it will ask about three objects. These percentages were based on the fact that there are a lot more questions available for one object than for multiple objects. The measure used in the question (e.g., height, distance) is chosen randomly. When players input responses, they may do so with a variety of units. For example, a question about width may be answered in meters, centimeters, feet, inches, or other applicable units. All answers are then converted to standard units listed below for calculation.

2.2.1. Single object variables

For single objects, such as a bucket, the Quanty software randomly chooses a single object variable to ask about. For example, "What is the width of the bucket?" The software converts whatever units are used by the player into standard units, as indicated in the lists below in parentheses. Note that the game also can ask about more "subjective" qualities, such as beauty (as measured by a five choice Likert scale) and brightness (in terms of a percentage).

- Continuous
  - Height (meters)
  - Weight (grams)
  - Length (meters)
  - Width (meters)
  - Distance from Camera (meters)
  - Volume (cubic meters)
  - Depth (meters)
  - Temperature (degrees Celsius)
- Subjective ("fuzzy")
  - Beauty (Bipolar adjective scale 1–5)
  - Brightness (Percentage)

2.2.2. Multi-object variables

If two or three objects are outlined, they are each outlined in a different color. The question in these cases will be relational-either the distance between the two objects (e.g., "What is the distance between the owl and the leaf?") or a subjective Likert rating for English prepositions, such as "occluding" (e.g., "How much does the owl occlude the leaf?").

- Continuous
  - Distance between (meters)
- Subjective ("fuzzy")
  - Occluding (percentage)

2.3. Scoring

The game lasts for an allotted time, and consists of a series of rounds. In each round, the pair of players answers a question. The players may complete as many rounds as they are able to within the time limit. Afterwards, the total score is the sum of the score of each round. The maximum score for a given round is 100. The minimum score for a round is 10, implemented by awarding the players 10 points even if they actually scored lower.

When the software receives a response message from both players, the values are used to calculate a score. The idea is to award more points for closer responses. This is done in two ways, depending on whether this question has been answered before for the same objects. If previous answers exist, the standard deviation (SD) of the set of answers is calculated. The difference between the two current responses is calculated. Points are then subtracted from the maximum score depending on how many SDs apart the two responses are. First the system calculates the standard deviation:

$$\sigma = \sqrt{\left(\sum (r - r_2)/N\right)} \tag{1}$$

where r is each response, r is the mean of responses, and N is the number of responses.

Then the system calculates the distance, defined as the number of standard deviations, between participant answers:

$$\text{distance} = |(r_1 - r_2)|/\sigma \times m \tag{2}$$

where r_1 and r_2 are the player responses, $m = 5$ for quantitative questions, and 25 for Likert scale responses, so that the limited range or answers (1–5) spans the 0–100 score range. The final score is 100 − distance.

If no previous answers are found, the SD cannot be calculated. Instead, the percentage value of one response over the other is calculated and then multiplied by 100 to find the score value.

A special case is where there are no preexisting answers, and one of the current responses is zero. In this case, a percentage is not appropriate because no matter how close the other responses are, the score would be calculated as zero. To get around this problem, a score of 50 is automatically given as a reward for being the first players to answer the question.

Different functions are used to evaluate Measurement and Likert responses. The functions differ because Measurements are more complex to evaluate, in the following way: The process for Measurement responses (as opposed to Likert-scale responses) must account for a variable "scale" of the questions. For instance, it is harder to guess the exact length of a whale than the exact length of a pencil. As in the formula given above for finding the distance, the difference between responses is divided by the SD. The distance is therefore the number of standard deviations between responses, multiplied by a constant. Computing score in this way is expected to be adaptive to each question because the SD is assumed to depend on the difficulty of the question. The difficulty may be due to physical scale, or other factors. Using the same example, the SD for responses about the whale will, in all likelihood, be much larger than the SD for responses about the pencil. The scoring for both questions will be fair because of the SD's involvement in the calculation.

3. Evaluation of the Quanty game

The purpose of this evaluation was to assess the usability and accuracy of measurements obtained by playing Quanty. We hypothesized that Quanty would produce accurate estimates when

aggregated statistically, and would be rated highly for being fun to play. To test the hypotheses, in March 2014, an online survey titled "Play an online guessing game and get paid!" was advertised in an online crowd sourcing service, Crowdflower.com, offering $0.50 per participant. The advertisement requested participants able to use the Internet and fluent in English. Participation was entirely anonymous and voluntary, and participants had the right not to answer certain questions, or to withdraw by clicking the 'Discard' button at the bottom of each survey page.

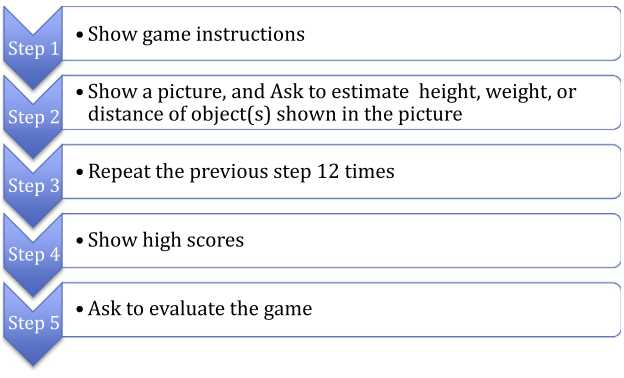
The advertisement was broadcasted through American other International channels; Different channels were chosen because Americans use non-metric system (feet, inches, and pounds) and most other countries use the matric system (centimeter, meter, and kilogram) to estimate. This allowed us to check if our method elicits the wisdom of the crowd regardless of how people provide estimates. Because of the uneven numbers of participant in each group, only within group (and not between group) comparisons are considered to make conclusions.

3.1. Participants

Forty-six Americans completed the survey and formed the group 1 for subsequent analysis. Fifty other participants, from 23 different countries, formed group 2. Their countries included: Russian Federation, Bangladesh, Belarus, Germany, Turkey, Mexico, Latvia, Argentina, Austria, Ukraine, Greece, Hungary, Philippines, Pakistan, Spain, Indonesia, Netherlands, Portugal, Vietnam, Canada, Romania, and India. All of these countries use the metric system to estimate.

3.2. Method

A participant entered the survey by typing in the survey URL. The survey first showed the informed consent form; if participants consented they were shown the screenshot of the Quanty game home screen that described how to play (see Appendix A). Then the survey showed 12 photographs to a participant, one photo per page, and asked participants to guess lengths, weights, or distances of objects shown in each picture (see Appendices C and D for examples). After participants estimated the requested quantity on a page, they were shown another page with high scores, as shown in the Quanty game. Finally, the survey asked participants to rate the game for fun. See the following diagram for the research design:



3.3. Results: Analysis of accuracy

Unlike the real Quanty game the actual answers to each measurement question asked in the survey was known to the researchers. Because Group 1 had 46 participants there were 46 estimates by the non-metric system for each of 12 pictures shown to Group 1. Because Group 2 had 50 participants there were 50 estimates by the metric system for each of 12 pictures shown to Group 2. Visual scan of the estimates showed that they were widely variable, and there was no way for any researcher to know before hand which participant was better at giving close to accurate estimates. So we wanted to investigate whether or not aggregating the estimates for each picture could give close to accurate answers. Thus there could be a total of 12 aggregated estimates for Group 1, and a total of 12 aggregated estimates for Group 2.

We reasoned ratios of aggregated estimates to correct answers, rather than other measures, could standardize the estimates for comparison across different units of measurements. The closer a ratio was to 1.00, the more accurate was the aggregated estimate. A ratio of 5.00 would mean the estimate was 5 times larger than the real value. A ratio of 0.20 would mean the estimate was five times smaller. We calculated the ratios by four statistical analyses to investigate which method produced aggregated estimates closest to the real value.

3.3.1. Statistical analysis 1

For each of the 12 pictures shown, the average answer across all the participants of a group was calculated. So for example, the average answer for the total weight of fruits shown in Appendix D was 5.6 lb by Group 1, and the average answer was 4.4 kg by Group 2. Then ratios of average answers to correct answers were calculated for each picture. For example, the correct answer for the total weight of fruits shown in Appendix D was 5.5 lb (2.5 kg); so the ratio of average to correct answer was 1.0 for Group 1 and it was 1.8 for Group 2. Thus calculated, the ratios of group 1 participants ranged from 0.04 to 24.0; the average of these ratios indicated that aggregated estimates by Group 1 were on average 4.3 times higher than correct. The ratios of group 2 participants ranged from 0.5 to 14.0; the average of these ratios indicated that aggregated estimates by Group 2 were on average 4.55 times higher than correct.

3.3.2. Statistical analysis 2

This analysis considered the outliers that were very different from other estimates. For example, one player guessed the height of the building in Appendix C as 80,000 feet, when it is actually 2443 feet. The outliers greatly skewed the average estimates. As a standard practice in experimental psychology (Leys et al., 2013), in analysis 2, values that fell outside the average value plus or minus two standard deviations were removed. For each picture, zero to five estimate(s) was thus removed and a new average was calculated from the remaining estimates. For example the new average for the total weight of fruits shown in Appendix D was 4.4 lb by Group 1, and it was 2.3 kg by Group 2. Finally, the ratios of the new averages to correct answers were calculated for each picture. Thus ratios of Group 1 participants ranged from 0.02 to 17.0; the average of these ratios indicated that the estimates were 3.08 times higher than correct for Group 1. The ratios of Group 2 participants ranged from 0.4 to 6.0; the average of these ratios indicated that the estimates were on average 1.59 times higher than correct by Group 2.

3.3.3. Statistical analysis 3

Because a median rather than the mean is less sensitive to outliers (Leys et al., 2013), in statistical analysis 3, the median estimate for each picture was calculated. Then the ratios of the medians to correct answers were calculated to gauge the accuracy of estimates. For example, the median answer for the total weight of fruits shown in Appendix D was 4.0 lb by Group 1, and the average answer was 1.5 kg by Group 2. Then ratios of median answers to correct answers were calculated for each picture. For example, the correct answer for the total weight of fruits shown in Appendix D was 5.5 lb (2.5 kg); so the ratio of median to correct answer was 0.7 for Group

1 and it was 0.6 for Group 2. The median estimates of Group 1 participants ranged from 0.01 to 12.0 of correct answers; the average of these ratios indicated that the estimates were on average 2.34 times higher than correct answers for Group 1. The median estimates of Group 2 participants ranged from 0.2 to 2.0 of correct answers; the average of these ratios indicated that the estimates were on average 1.29 times smaller than correct answers.

3.3.4. Statistical analysis 4

First, the median of all the original estimates for a picture was calculated. Second, each original estimate of the picture was transformed by subtracting the median. Third, a new median was calculated from the transformed values. Fourth, this new median was multiplied by a constant value, 1.5 (Leys et al., 2013, p. 3). Fifth, the original estimates that were greater than the multiplied value was removed. Sixth, the median of the remaining original estimates (MAD) was calculated. This method of calculating MAD was repeated for each of the 12 pictures shown to each Group. Finally, the ratio of each of resulting 12 MADs to correct answers was measured for each group. According to the MAD/correct answer ratios, the accuracy of measurements by Group 1 participants ranged from 0.01 to 10.83 (the average ratio being 2.12 times correct answers), and by Group 2 participants ranged from 0.05 to 1.56 (the average ratio being 0.61 times correct answers).

Because both means and standard deviations are sensitive to outliers, statistical analyses 3 and 4 were expected to produce the greatest accuracy for not considering means and standard deviations. Fig. 1 shows the average of the ratios calculated by each statistical analysis. The first pair of bars show the average of the 12 Mean/Correct answers by Group 1 and 2 participants; the second pair of bars show the average of the (Average + −2sd)/Correct answers by group 1 and 2 participants; the third pair of bars show the average of the 12 Median/Correct answers by group 1 and 2 participants; and the fourth pair of bars show the average of the 12 Mad/Correct answers by group 1 and 2 participants.

The average ratios of both the statistical method 3 (Median/Correct) and statistical method 4 (MAD/Correct) were better at approaching accuracy of the estimates (approaching the ratio 1) compared to other two statistical methods. We did not find any significant difference between the ratios calculated by Median/Correct and MAD/Correct methods. Perhaps that is because both the median and MAD do not use mean or standard deviation, and are robust measures in the presence of outliers.

Additionally, we explored why participants were more accurate in estimating the measurements of some objects than they were for other objects. People seemed to be more accurate in estimating measurements of objects that they were likely to have personal experience with (for example, the weight of fruits) than they were for estimating measurements of unknown objects (for example, the weight of a military tank). However, the differences were not statistically significant, and are therefore not reported.

3.4. Results: Analysis of fun

Fig. 2 shows how the participants rated the game played in the survey (1 indicated a low rating and 7 indicated a high rating).

The high ratings in response to all questions suggested that the Quanty game could be as fun and engaging as the game 'tetris'. In addition, to measure whether or not participants paid attention to the game played in the survey, they were asked to recall their measurements of objects shown in three pictures. They were able to recall their answers almost 100% of time.

4. Discussion

The purpose of the paper was to propose and evaluate an online game that collects quantitative estimates from a diverse group of people in a fun and competitive way. The idea was to aggregate players' responses to elicit the wisdom of the crowd so that professionals could utilize responses by such a game confidently. Four statistical methods were used to investigate which method produced the greatest accuracy in aggregated estimates.

Participants were asked to estimate measurements of objects shown in 12 pictures. We found that answers gained through the wisdom of the crowd were often accurate, but the accuracy depended on which statistical method was used to aggregate answers. For our sample we found that medians and median absolute deviations of aggregated responses were able to elicit close to accurate responses. Our results also show that the participants wanted to play the game again and it was rated highly as being fun.

Furthermore, we found that the accuracy of estimates varied across pictures. Estimates were better for magnitudes related to objects we handle daily than were for very large, small, or distant objects. Thus for quantitative magnitudes, we can expect to get accuracy if we could get relevant people to provide us with guesses. For example, it seems reasonable to predict that the aggregated estimates of the weight of a bridge will be more accurate if we aggregate guesses of engineers than of psychologists, and ratings of how depressive a person looks will be more accurate if we aggregate guesses of psychologists than of engineers, etc.

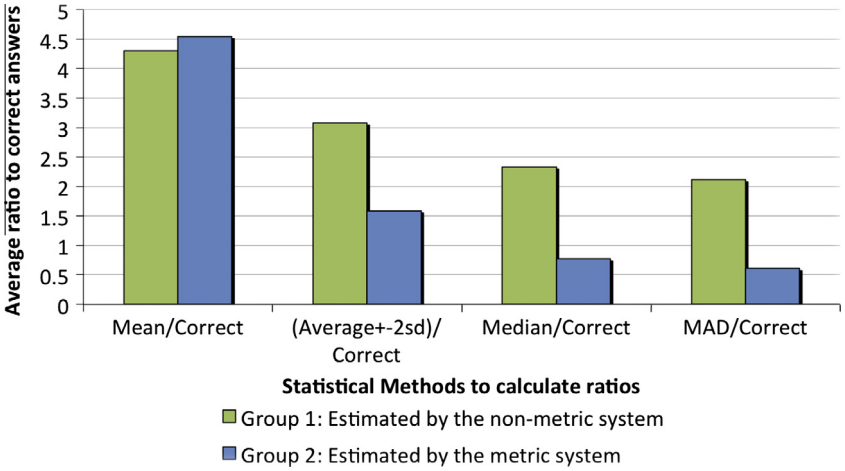


Fig. 1. Average accuracy of aggregated estimates by four statistical analyses.

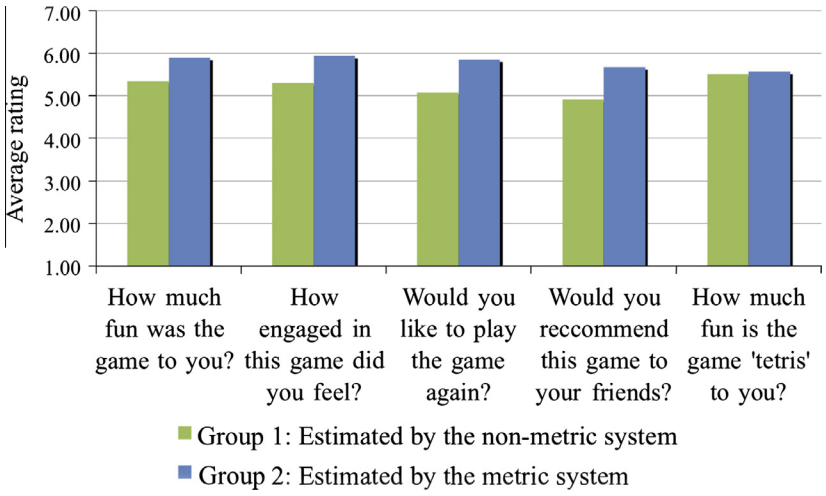


Fig. 2. The average ratings of the game by group 1 and group 2 of participants.

4.1. Comparisons to other serious games

Previous studies show that some serious games, such as the ESP Game (Von Ahn & Dabbish, 2008) and Peekaboom (Von Ahn, Lui, & Blum, 2006), can be deployed online and elicit wide responses. Previous studies also show that the wisdom of the crowd can be elicited by aggregating judgments (Surowiecki, 2005). Furthermore, guessing things appears to be fun, perhaps in the same way guessing answers in trivia is fun.

Quanty uses previous ideas to deploy an online game, in which an aggregation approach is used to find close to accurate solutions to quantitative problems. These are the problems that are difficult to solve by computational means but nonetheless can be solved reasonably well, by a diverse group of independent people.

4.2. Limitations

Similar to all studies, however, ours is not without limitations. The sample size was relatively small and the survey was conducted with online paid volunteers. Nevertheless, we have designed, implemented, and deployed a web game that will collect quantitative information about our everyday world and can be aggregated to approach accuracy. Anyone can and may export Quanty's data from the Quantygame.com website. Professionals can improve their estimates by showing their own pictures to several known or unknown people, elicit the wisdom of the crowd in the same way as we did, and be more confident about what they estimate in their practice.

Acknowledgements

The project is funded by Carleton University Foundry Program, and National Science and Engineering Research Counsel (NSERC) Discovery Grant.

Appendix A. 'How to play' instructions shown to a player

Appendix B. An example question shown in Quanty

Appendix C. What is the height of the building?

Appendix D. What is the total weight of fruits?

References

Gilbert, F. J., & Lemke, H. (2014). Computer-aided diagnosis. *The British Journal of Radiology, 78*(Suppl. 1), S1–S2. http://dx.doi.org/10.1259/bjr/23717382.

Hacker, S., & Von Ahn, L. (2009). Matchin: Eliciting user preferences with an online game. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1207–1216). ACM.

Lee, M. D., & Shi, J. (2010). The accuracy of small-group estimation and the wisdom of crowds. In R. Catrambone & S. Ohlsson (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society*. Cognitive Science Society.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766.

Matyas, S., Kiefer, P., Schlieder, C., & Kleyer, S. (2011). Wisdom about the crowd: Assuring geospatial data quality collected in location-based games. In *Entertainment computing – ICEC 2011* (pp. 331–336). Berlin, Heidelberg: Springer.

Nickerson, J. V., Zahner, D., Corter, J. E., Tversky, B., Yu, L., & Rho, Y. J. (2009). Matching mechanisms to situations through the wisdom of the crowd. *ICIS 2009 proceedings*, paper 41.

Ritterfeld, U., Cody, M., & Vorderer, P. (Eds.). (2009). *Serious games: Mechanisms and effects*. UK: Routledge.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2005). Label me: A database and web-based tool for image annotation. *Memo AIM-2005-025*, MIT AI Lab.

Steyvers, M., Lee, M. D., Miller, B., & Hemmer, P. (2009). The wisdom of crowds in the recollection of order information. In J. Lafferty & C. Williams (Eds.). *Advances in neural information processing systems* (Vol. 23, pp. 1785–1793). Cambridge, MA: MIT Press.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM, 51*(8), 58–67.

Von Ahn, L., Liu, R., & Blum, M. (2006). Peekaboom: A Game for locating objects in images. In *Proceedings of the SIGCHI conference on human factors in computing systems (Montreal, April 22–27)* (pp. 55–64). New York: ACM Press.

Von Zur Gathen, J., & Gerhard, J. (2013). *Modern computer algebra*. Cambridge University Press.

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive Science, 36*(3), 452–470.