The Wisdom of Crowds: Impact of Collective Size and Expertise Transfer on Collective Performance

Christian Wagner
School of Creative Media and
Department of Information Systems
City University of Hong Kong
Kowloon, Hong Kong
c.wagner@cityu.edu.hk

Ayoung Suh
School of Creative Media and
Department of Information Systems
City University of Hong Kong
Kowloon, Hong Kong
ayoung.suh@cityu.edu.hk

Abstract

We explore the performance of collectives on judgments tasks along two dimensions: crowd-size and expertise transferability. Based on an experimental investigation of the judgment ability of 478 subjects and sub-samples thereof, we identify that crowd-size matters, but only within limits. We also identify limited transferability of expertise, largely based on an anchor-and-adjust heuristic. Collective size and expertise transfer effects are moderated by task difficulty and are strongest for tasks in a medium difficulty range.

1. Introduction

By now, ample evidence exists that crowds, or collectives, are able to provide high quality judgments on difficult judgment and choice tasks, frequently outperforming individual experts [e.g., 27, 38]. This capability has correspondingly been termed collective intelligence, or wisdom of crowds, to reflect the collective ability of producing results of surprising, expert-like quality [16, 45]. While group decision making and problem solving have long been recognized as valuable methods to improve decision quality and solution-finding, collective intelligence is different, in that it seeks the (aggregate) insights of collectives, not groups [25].

Having a large pool of judges can obviously reduce errors and noise in predicting, but proponents [e.g., 31] suggest that collective intelligence comes down to more than simply the law of large numbers. In estimating tasks, the supposed special characteristic of collective intelligence is that different individuals will apply different "theories", the aggregate of which results in highly precise judgment [31]. This raises important questions concerning the operationalization of collective intelligence.

First, if collective intelligence is, at least in part, a matter of aggregating personal "theories" about a judgment task, then the collective has to be large enough to provide sufficient theory diversity [36], while at the same time creating enough replications to reduce the size of any random errors. The term "theory" in this context does not represent a true theory, but can reflect a different heuristic to solving the problem, or even a sub-heuristic including estimation of individual problem constructs. For example, if asked to guess the weight of the Eiffel Tower, one person may consider the weight of a miniature Eiffel Tower statue and "scale it up", while someone else may consider the weight of steel, then the height, density, and so on to compute an estimate. These would be considered two different heuristics or "theories". Hence, the question arises of how large the collective has to be. This question is not all new, having been raised in very early studies of individual vs. group judging, such as Gordon [14], or more recently in the realm of prediction markets [34]. Proponents of prediction markets have illustrated that a functional collective can consist of less than 10 individuals [17], an estimate in line with research elsewhere [e.g., 43]. This minimum number would likely though not create the best results, just acceptable results. Collectives, large or small, need to generate a high level of diversity among participants, so as to create a multiplicity of theories and thus a well balanced outcome. The impact of size is thus an important consideration.

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.80

IEEE computer society

Second, if the possession of a diverse set of theories is important for the crowd to function well, a crowd, although able to create a diverse set of responses for one problem type, may fail to do so. In other words, expertise transfer from one problem type to another may face considerable hindrances, thus resulting in very unstable collective performance across problems. This again is an important consideration for organizations, which seek to maintain a think-tank of problem solvers to draw on for a range of problem solving tasks.

Correspondingly, within this article we seek to answer two questions: First, how many people are needed to create intelligent collective? Second, does collective intelligence transfer between problems?

The remainder of the article is organized as follows. In the background section we review the relevant prior research for this work. Thereafter we explore the research model, constructs, and research hypotheses. Then we discuss findings, followed by discussion and limitations. We end the article with conclusions and suggestions for future work.

2. Background

2.1. Collective intelligence and judgement quality

Collective intelligence typically refers to the high task performance of large aggregates of loosely connected individuals. In a broader sense, it encompasses a host of meanings, related to the behavior of a "complex adaptive system" and the distributed knowledge or capability in human systems where the whole is greater than the sum of the parts [3, 43]. Recently, research has focused on how digital networks and information technologies are enabling collectives to create valuable knowledge and to make better decisions [4, 29]. The ability of the Internet to connect large numbers of diverse individuals, combined with electronic collaboration technologies that facilitate aggregation of responses from many parties are opening a new era of collective intelligence [12, 19].

While many studies support the notion of the *wisdom of crowds*, it does not mean that collectives always outperform individuals in terms of judgment quality [1, 2]. In particular, diversity and expertise are necessary properties for a collective to be wise [4, 25]. Diversity is important so that the collective has access to different individual skill sets and different sources of information in order to avoid bias [17]. Furthermore, the collective must have expertise in that individuals understand the problem and collectively make positive contributions to solving it [4]. If everyone guesses randomly, due to a lack of expertise, no valuable information can be extracted from the responses. In addition, no amount of diversity and expertise will matter if individuals are completely unfamiliar with the task at hand, thus being unable to apply their information and expertise [38]. Accordingly, we posit that the co-existence of diversity, expertise, and task familiarity are key to tapping into collective intelligence. In the following sections, we discuss how those three elements operate and interplay to maximize the judgment quality of a collective. We thereby focus on collective size, transferability of expertise, and task difficulty, respectively.

2.2. Collective size

To create a collective of sufficient diversity and expertise, collective size would seem an important factor. Hence, how large does a collective have to be? Prior studies of collective intelligence demonstrated meaningful results already with less than 10 participants [14, 17, 43], and yet performance improvements should be expected for larger collectives, which would benefit from increased diversity, expertise, as well as replication of knowledge. Page [31] argues that collectives perform well because of the multiplicity of theories represented. Irrespective of the number of theories, prior research [11, 35] found that problem solvers, experts as well as non-experts, generally used few factors (e.g., 2-3) and rarely more than 5 in their judgments. Is the number of factors dependent on the complexity of the problem? It could be, but it may well be more of a problem solver characteristic (limitation) than a problem characteristic. Past research has recognized human capabilities being limited to the consideration of only 7±2 stimuli [30], or possibly even as few as four [7]. Thus, if problem solvers were approached with a judgment task that required consideration of a large number of stimuli, they would need to simplify, or would fail in the task. Consequently then, collectives are expected to operate on a small set of task characteristics. With a small number (e.g., 2 to 3) of problem solving heuristics and a similarly small number of relevant parameters (e.g., 1 to 4) per heuristic, the overall set of parameters considered by the collective overall might be 10 or less, and possibly closer to 5 (i.e., 2 × 2 = 4) than to 10.

While proponents of collective intelligence de-emphasize the impact of error reduction due to the law of large numbers [21, 31], one would expect that the estimate quality for each parameter would also rise with the replication of judgments. The NASA clickworker task [26], for instance, relies on replication to improve judgment accuracy. How many replications are needed? The desired sample size (for each parameter) would depend on the effect to be achieved, or the desired *power of the test* [6]. Given that the power of the test is logarithmic with respect to sample size, marginal returns will diminish. In practice, samples of 20, or even 10 lead to convergence.

Overall then, a collective representing multiple theories with multiple parameters and large enough sample size per each parameter would still only require about one hundred subjects.

2.3. Transferability of expertise

Beyond collective size, another important issue is the transferability of expertise. Our research connects back to prior studies by Farnsworth and Williams [13] and Klugman [24], which raised the question of the impact of familiarity on the quality of group judgment, namely whether familiarity would promote transfer of expertise. Farnsworth and William hypothesized that familiarity would improve collective performance. Klugman, who had subjects guess the number of objects in a jar (e.g., lima beans or marbles), found that subjects systematically over-judged object quantities, yet that collectives outperformed individuals for unfamiliar objects, contrary to expectation. Wagner and Suh [41] shed more light on the finding by demonstrating the existence of a range of difficulty with relatively high collective performance. This range, at a medium level of task difficulty, is thus referred to as the *collective range*. Outside of this range, both individuals and collectives perform well on easy tasks, thus yielding little advantage for the collective. For difficult problems, collectives and individuals both tend to perform poorly, again resulting in no advantage for the collective [28].

Correspondingly, we should expect some transferability of expertise, as long as the task to transfer to would not be too difficult or unfamiliar. In consideration of Dawes [8], we would thus expect transferability to be limited by two criteria, namely the transferability of the judgment model and the ability to apply the model well. Faithful transfer of the judgment model would lead to high judgment consistency, but not necessarily to accurate results. Accurate results would only be achieved if the model were applied appropriately, in particular, without bias [4, 33]. For example, a real estate appraiser with London expertise should predict prices consistently and accurately for London, and consistently, but not necessarily accurately for Liverpool or Manchester (in case of a possible "London bias"). When asked to estimate the price of a racehorse, the appraiser should perform both inconsistently and inaccurately, for lack of any judgment model.

In consequence, we would expect some transferability of judgment ability and corresponding collective intelligence from one problem to another, which should reveal itself in the use of one or more judgment models and accuracy of results, as long as the task difficulty did not exceed the collective range.

2.4. Task difficulty

The task nature appears to also play an important role in creating collective intelligence [32, 41, 43]. In particular, task difficulty has been considered as an important factor that can shape the collective performance [28, 42]. Accordingly, we posit that the collective performance should vary depending on task difficulty. Drawing on the previous literature, we conceptualize task difficulty as the task attributes that raise individuals' cognitive processing needed to reach a solution [5, 23, 37], reflecting both the subjective and objective difficulty individuals perceive in performing tasks [41].

3. Research Model and Hypotheses

3.1. Research model

The research employed a relatively simple model, with judgment quality (of the collective) as dependent variable, and collective size and expertise as independent variables. Task difficulty was modeled as moderating the impact of expertise and collective size (See Figure 1).

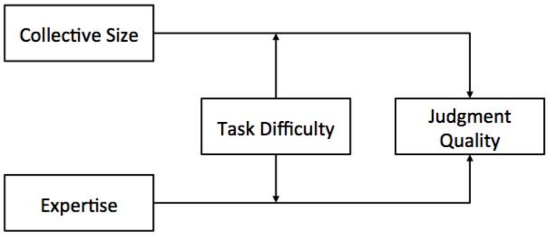

Figure 1: Research model

3.2. Effect of collective size

Based on prior research, we expected that the quality of collective estimates would depend on the size of the collective, with diminishing marginal returns for size increases. Quality improvements should eventually become insignificant, so that an increase in sample size should not lead to any more increase in variance explained. We estimated the upper bound for quality improvements to be in the range of 50 to 200, considering the possible presence of 5 to 10 relevant parameters, and the need for 10 to 20 replications. Our hypotheses are:

[H1] The size of collective will be positively associated with the collective performance (judgment quality) in general.

[H2] Diminishing marginal returns will set an upper bound for the effectiveness of collective size increases.

3.3. Effect of transferability of expertise

Transferability of expertise refers to the degree to which performance, here judgment quality, in one task affects the performance on other tasks [cf. 22]. Since expertise is domain and task specific, experts in a domain should perform very well in that domain, but not necessarily elsewhere. Even then, experts may fail in performing novel tasks within the same domain, and sometimes also similar tasks [10]. Accordingly, we cautiously assume that expertise would narrowly transfer to tasks in the same domain, with similar complexity, and similar information requirements. For example, someone able to judge the daily high temperature for a particular city 1 or 2 days from today, should also be good at judging it for 5-6 days from today, simply by extrapolating results through a process of anchor-and-adjust [9, 39]. However, some of the judging expertise may be lost, if we ask for the judgment concerning a different city, due to the specificity of information.

[H3a] Expertise of a collective will transfer across problems.

[H3b] Expertise transfer takes place through a process of anchor-and-adjust.

3.4. Task difficulty

For the conceptualization of task difficulty, we included both objective (task complexity) and subjective (problem solver confidence) considerations, to derive an overall difficulty formulation [cf. 41]. As we were interested only in the moderating effects of task difficulty, we applied task difficulty however at a much less granular level, differentiating only between collective range (medium) difficulty tasks, versus easy / difficult tasks.

The assumption was in all cases, that task difficulty would moderate outcomes whereby the collective would have an advantage when problems were not so easy that each individual to find the right result and not so difficult that everyone would succumb to random guessing, thus preempting the extraction of meaningful information. Hence we formulated two hypotheses:

[H4a] Task difficulty will moderate the effect of collective size on judgment quality.

[H4b] Task difficulty will moderate the effect of expertise transfer on judgment quality.

3.5. Conceptualizing the assessment of collective intelligence (Judgment Quality)

For the measurement of collective intelligence, the challenging question is "how close (to the true value) is close enough"? The answer cannot be given in absolute terms. For example, a collective judgment 50% higher than the true value may be considered poor, but if all individuals are further from the true value, then the collective intelligence is still superior. Thus, for the measurement of collective intelligence quality, we used two separate relative measures, independent of the details of individual problems. One measure was based on the overall size of the collective error vs. individual errors, the other based on the number of times the collective performed better than individuals.

3.5.1. Collective intelligence quality. The first measure relied on the formulation by Page [31], which identifies two measures, collective error (CE) and individual error (IE). Collective error, the squared error of the collective prediction, represents the difference between the true value and the averaged estimates of all individuals. The individual error aggregates the squared errors (difference between their estimates and the true value) of all the participants. It thus captures the average inaccuracy of individual guesses. Individual error (squared-average) is a benchmark to compare the performance of the collective.

$$\textbf{CE} = (\bar{x} - x_{true})^2 \qquad \textbf{IE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_{true})^2$$

CE: Collective Error | IE: Individual Error

Using IE and CE, one can define a single measure, collective intelligence quality CIQ = IE/CE, as an overall measure of the collectives' superiority [41]. The theoretical minimum for CIQ is 1.0. At 1.0, the collective's aggregate error is equal to the sum of all individual errors, thus reflecting no collective intelligence. CIQ values < 10 identify moderate quality improvements. CIQ is useful when comparing different judgment problems whose mean results and thus errors differ widely, and as an expertise benchmark [41]. We instead standardized judgment results and judgment errors, and thus used CE only.

3.5.2. Win ratio. When the aggregate judgment of a number of N individuals in a collective is compared to the judgment of each of the individuals in that collective, we can compute how many times out of the N comparisons the individual judgment is (as good or) better and how many times the collective is better. Thus we can define a "win ratio" (WR), following earlier research, such as [20, 40]. A win ratio of 50% would reflect equal performance of the collective and individuals, and thus no performance improvement; a WR of 75% would indicate that 3-of-4 times the collective performed better.

4. Data Collection and Findings

4.1. Data collection overview

To carry out the empirical investigation, we recruited problem solvers through an online panel service. 500 subjects were solicited and instructed to complete our task set without the use of external aids. Tasks were structured so that subjects were unable to revise already completed answers, in order to avoid backward changes resulting from process learning. Each subject received approximately $2 for the task completion, a typical amount for the participants of this online panel service. Subjects were 45% women vs. 55% men, ranged in age from 21 to 59, possessed a high school (18%), college (72%), or graduate school education (9%), and held a range of occupations (67% were office workers). Overall, we considered the collective to have considerable diversity. As they had been solicited online from an anonymous pool of thousands of panel members, we expected subjects to not know each other, to draw on individual data sources and to act independently from each other. Subjects were advised to not attempt to look up data from the Internet. Given the relatively low remuneration and the challenge to look up relevant data on the Internet, we believed they did not. Further, the data identified no individuals who consistently generated good answers, as would be suggested by Internet look-up of information (especially when perfect information was available).

22 subjects out of 500 were excluded from the evaluation because of non-sensical data, including text input when numbers were required, or other inconsistencies, which disqualified them from meaningful evaluation. The remaining 478 subjects were then used in the evaluation.

4.2. Experimental tasks

Overall, seven principal tasks in two sets were chosen for the study: Three temperature guessing tasks, and four tasks to guess the weight of substances. All tasks were non-random. The temperature guessing tasks had uncertain, yet predictable (heuristic) outcomes, the weight guessing tasks were completely deterministic, but required some analysis and lacked data availability.

Temperature tasks asked for judgments of city temperatures today (T0), plus one (T1) and 6 days (T6) ahead, for a city well known to the collective, plus two other less familiar cities. At the low end of complexity was the question "what is the high temperature today in Seoul?" (Seoul being the respondents' home country capital).

The other task set asked for an estimate of weights of specific amounts of coffee, milk, gasoline, air, and gold.

Tasks had been pilot tested and modified to create a broad difficulty spectrum, based on familiarity and thus information availability, as well as task complexity.

We added a highly complex and difficult problem as a control task. We asked "if all poor people in the world gave you US$ 1 each, how much sustainable monthly income could you derive from the resulting amount?" This control task was to observe subject behavior at that level of difficulty. Results are not shown here, but are consistent with the findings described for difficult problems in this experiment.

To measure task difficulty, we focused on both task and problem solver related characteristics, by combining subjects' perceived complexity ratings for the problem [cf. 5], and their perceived confidence (reversed) in the quality of their solutions. Subjects would provide these ratings on a 7-point Likert scale, which in the case of perceived confidence would be reversed to measure lack of confidence. The ratings were then combined for an aggregate measure of difficulty, which we categorized into medium (collective range) vs. easy/difficult to enable the subsequent assessment of moderation effects.

4.3. Impact of collective size

We sought to determine whether an increase in collective size improved collective judgment quality, expecting some effect, but diminishing returns as the collective size increased. To test the hypotheses, we compared the performance of collectives of different size, by sub-sampling from the original pool of 478 subjects. The sub-sampling used a bootstrapping approach, which generated 10 sub-samples of size N = 10, 20, 50, 100, and 200 subjects with replacement.

Within each sub-sample we calculated win ratios (WR) and collective errors (CE) and computed mean WR and CE values across each 10 subsamples. Using the $\log_{10}$ sub-sample size as the independent variable, with mean WR and mean CE as dependent variables, we carried out regression analyses to identify possible relationships between sample size and performance. Regression analyses differentiated between problems of medium difficulty (collective range), and problems outside that range (easy or difficult tasks).

According to prior research [41] and expressed in H4a, problems in the collective range were expected to yield better judgment quality, because of the moderating effect of task difficulty. To measure this, regression functions needed to be separated. Results of the regression analyses for WR revealed two distinctly different regression functions.

Collective Range Problems: $WR_{\mathbf{CR}} = 0.695 + 0.074 \times \log_{\mathbf{10}}(\text{Size})$

Easy/Difficult Problems: $WR_{\mathbf{ED}} = 0.480 + 0.023 \times \log_{\mathbf{10}}(\text{Size})$

For the collective range, the regression function was significant at $p = 0.008$. For easy/difficult problems, the regression function was significant at $p = 0.067$. $R^2_{\mathbf{CR}}$ ("R-square collective range") was 0.928, $R^2_{\mathbf{ED}}$ was 0.725.

Problems in the collective range yielded a base win ratio for the collective of almost 70% (0.695), with an increase of 7.4% for each 10-fold increase in the collective size. This suggests that a collective of 10 would outperform an individual approximately 3 out of 4 times (0.769). Problems outside the collective range, i.e., either difficult or easy, showed relatively even win ratios (48%), with only 2.3% increase for each 10-fold increase in crowd size. Thus, a collective of 10 would have a mere 0.503 win expectation.

Results for collective error (CE) were similar, with CE dropping for larger collectives, but only to a point. Figures 1 and 2 represent the relationship between collective size and collective performance in terms of WR and CE, respectively.

The considerable differences between win ratio and collective error functions based on task difficulty (Figures 1 and 2) were significant according to t-test results, with $t = 21.80$ ($p = 0.000$) for WR and $t = 7.11$ ($p = 0.002$) for CE, thus indicating the moderating effect of task difficulty.

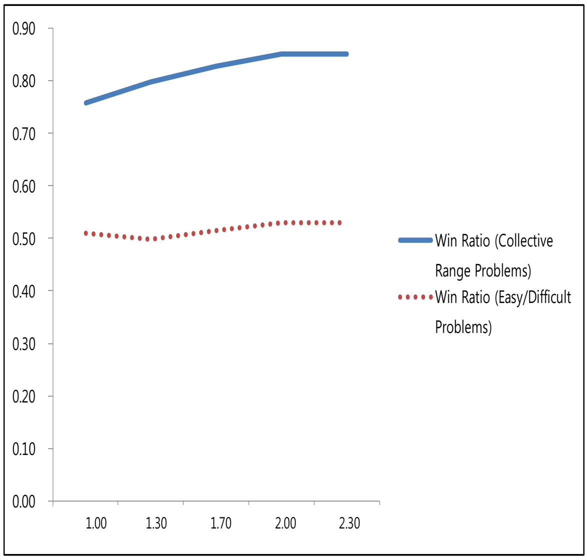

Figure 1. Relationship between collective size and win ratio (WR)

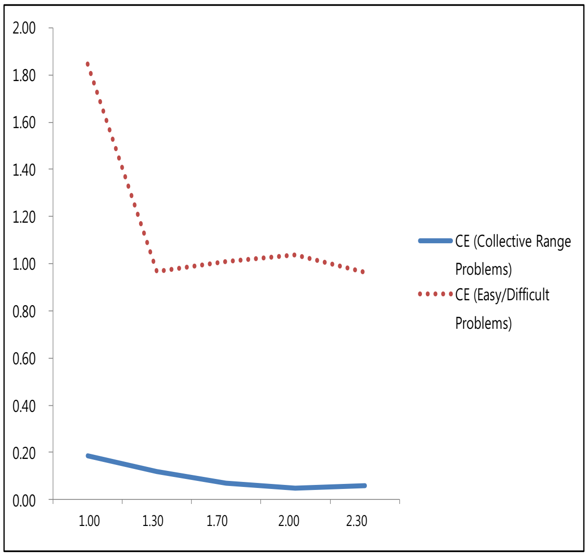

Figure 2. Relationship between collective size and collective error (CE)

To further determine the impact of size, we computed different $R^2_{\mathbf{CR}}$ and $R^2_{\mathbf{ED}}$ values in the collective range of difficulty, namely for collective sizes up to 50 (i.e., 10, 20, 50), up to 100, and up to 200. R^2 for win ratios grows with sample size increases up to 100, R^2 for CE increases only until sample size 50. This suggests that beyond sample size of 100 (or even 50), a larger collective did not matter. For problems outside the collective range, collective size mattered up to $N = 200$. Table 1 presents the relationship between collective size and R^2 with consideration of task difficulty.

Table 1: Impact of collective size on R^2 with Consideration of Task Difficulty

	Collective Range		Easy/Difficult Tasks	
Collective Size from 10 up to	WR	CE	WR	CE
50	.970	.976	.146	.639
100	.980	.951	.392	.508
200	.928	.833	.725	.487

4.4. Transfer of expertise

This test sought to answer the questions "does collective performance extend from one problem to another of the same type?", and "can we observe particular mechanisms of the transfer of expertise?" (which would explain the transfer of expertise). Our expectations were that expertise would be transferred, that expertise transfer would use an anchor-and-adjust and a heuristic, yet that expertise transfer would deteriorate with increased problem difficulty.

4.4.1. Transfer of expertise: Weight. The weight judgment tasks provided a useful scenario to assess transfer of expertise. Subjects began by judging the weight of a cup of a highly familiar substance, coffee (water), and then moved on to judge the weight of the same quantity of milk, gasoline, air, and gold. The physics of the situation were simple. Weight (mass) was determined by the product of the container size and the density of the substance. Coffee (water) provided the base with density 1.00 ($1.00g/cm^3$). Milk's true weight was 1.03 times as much, gasoline 0.75 times, air about 0.001 times, and gold had a multiple of 19.3. Of course, subjects were not told this. Would subjects then be able to transfer expertise from one task (coffee weight) to the next (milk weight), and would familiarity with the substance matter?

The mean estimate for a cup of coffee (water) across all subjects was 225 grams, thus providing a meaningful starting value for other estimates (suggesting a mind model where a cup equaled 225 cm^3 or 7.6oz). Hence we carried out regressions to test whether subjects considered the mathematical relationship between the substance weights. A first regression was computed without a constant term, assuming that subjects would (correctly) adjust proportionally, but not make a linear shift. The second, with a constant, allowed for both linear and proportional shift.

Tables 2 and 3 below explain the estimation of various weights in relation to the weight of a cup of coffee (water). The first table assumes a model without constant term, the second table with constant term. For all substances, the model without constant (proportional model) provided a better fit, yielding higher R^2 values and higher F values. Thus, we focused on the (correct) model without constant. For expertise transfer to milk and gasoline weight, F values were 2 to 3 orders of magnitude larger than for gold and air. Similarly, R^2 values were high for milk and gasoline (.944 and .761) but low for gold and air (.006 and .129).

The results suggest that subjects used an estimation process consistent with proportional adjustment (no constant term) for the weight of other substances. As long as subjects were familiar with substances, they estimated the weights quite consistently, resulting in R^2 values of .947 and .800, respectively, and fairly accurately ($b_{\mathbf{true}}$ within 25% of b). For gold and air, the knowledge transfer process deteriorated. R^2 values dropped to .006 and .129 respectively, suggesting a change in judgment processes, indicating random guessing. Weight judgments were quite inaccurate, underestimating the weight of gold by a factor of 4.1 ($b_{\mathbf{true}}$ / b), and overestimating air by a factor of 50.

Table 2: Judgment model without constant (N=478)

	b	b_{true}	R^2	F	p <
Milk	1.00	1.03	.944	7,978.394	.0000
Gas	0.99	0.75	.761	1,521.029	.0000
Gold	4.69	19.30	.006	4.045	.0450
Air	0.15	0.00	.129	70.711	.0000

Table 3: Judgment model with constant (N=478)

	b_0	b_1	R^2	F	p <
Milk	23.77	0.93	.851	2,710.789	.0000
Gas	30.41	0.91	.502	480.686	.0000
Gold	934.81	2.09	.001	0.299	.5850
Air	17.88	0.10	.023	12.194	.0010

4.4.2. Transfer of expertise: Temperature. We approached the temperature-judging task with a similar logic as the weight task, exploring whether subjects would make judgments with their present day estimates as an anchor. The hypothesized model was slightly different from the model for weights, in that subjects were believed to anchor on the current temperature and then make linear, not proportional adjustments. Hence we tested

$$T_{\mathbf{Days}} = b_{\mathbf{0}} \times T_{\mathbf{0}} + b_{\mathbf{1}} \times \text{Days} ;$$

with the expectation that if subjects were anchoring on temperature $T_{\mathbf{0}}$, then $b_{\mathbf{0}}$ would become 1.00, and $b_{\mathbf{1}}$ would reflect the collective's expectation of daily temperature changes in future. R^2 would inform us of the quality of the model, or in other words the collective's precision (but again not its accuracy, which would be identified by accurate b-values). We regressed temperatures for each city separately, using models with and without constant (intercept). Tables 4 and 5 below summarize the collective estimation results for temperatures in different cities. Again, the models without constant term (Table 4) were superior, so we will refer to them only.

Once again, the models were highly significant, explaining between 88% and 97% of the variance, and validating a model what would anchor on present day estimated temperatures (b_0 of 0.90, 0.96 and 0.98 within 10% of the true value 1.00), and yield temperature gradients (slopes) of between 0.18 and 0.30 per day. The positive slopes indicated that subjects knew that the two less familiar cities (Rome and Vladivostok) were in an upward temperature trend (from Winter to Spring), while their prior estimates for T_0 revealed knowledge about the relative climates of Rome and Vladivostok ("Vlad").

Table 4: Linear model of collective temperature estimates without constant (N = 478)

	b_0	b_1	R^2	F	p <
Seoul	0.90	0.28	0.88	16,867.73	0.000
Rome	0.98	0.30	0.97	13,659.07	0.000
Vlad	0.96	0.18	0.91	4,754.58	0.000

Table 5: Linear model of collective temperature estimates with constant (N=478)

	b	b_0	b_1	R^2	F	p <
Seoul	-0.63	0.96	0.35	0.86	925.49	0.000
Rome	0.48	0.96	0.25	0.87	3305.46	0.000
Vlad	-0.93	0.93	0.32	0.67	2929.38	0.000

5. Discussion

First, our analysis reconfirmed earlier research on a collective range of difficulty, where collectives perform best. For our findings, this differentiation was very important, as judgment quality in the collective range was 3:1 (win ratio), whereas it was 1:1 outside that range. Effects for expertise transfer disappeared as well for difficult tasks. Second, we confirmed that collective size mattered, but only in limited ways. Specifically, beyond a collective size we estimate to be around 100, independent of task, but dependent on judges' model complexity (and thus the number of factors for which collective "consensus" needs to be achieved), there appears little improvement and little economic value in enlarging the collective. Collective sizes have to increase 10-fold to achieve marginal judgment quality gains. Finally, we confirmed a transfer of expertise, which can be explained surprisingly well by a simple anchor-and-adjust model. This does not mean that the adjustment model is trivial. In our case, for instance, temperature gradient adjustments not only were in the right direction, but also showed a slope (b_1) ranging from 0.18 to 0.30) relatively close to true results for the locations and time of year. In other words, not only were the collective estimates good, but so also was the estimation mechanism, as long as problem difficulty was in the collective range. Overall, our hypotheses were confirmed (see Table 6).

Table 6: Summary of hypothesis tests

Hypo's	Predicted	Test	Finding
H1	Size effect	Regression function of size effect	True
H2	Upper bound for size effect	R^2 does not increase beyond N = 100	True
H3a	Expertise transfer	Regression, high R^2	True
H3b	Anchor-and-adjust	Regression, collectives adopt correct model, good b-values	True
H4a	Task difficulty moderates size effect	Win-ratios significantly lower (t-test)	True
H4b	Task difficulty moderates expertise transfer	Variance explained (R^2) significantly lower outside collective range	True

6. Conclusions and Future Work

Our work revealed three quite interesting results, all of which will require further exploration in future work.

First, collectives did benefit from larger numbers in their midst, but seemingly less to embed many theories, but to compensate for systematic errors of some individuals who judged without knowledge and thus would have introduced too much bias. This goes somewhat counter the underlying assumptions of collective intelligence, and emphasizes error correction as an important element of crowd problem solving. Nevertheless, the occurrence of clusters in the data sets suggests the presence of multiple theories, which together did generate the final results that were better than individual results.

Second, we cannot over-emphasize the impact of task difficulty. Collective performance differs significantly based on performance, but not as expected based on prior research, which conjectured advantages for very familiar or very unfamiliar tasks. Instead, rather the tasks whose difficulty lies between the familiar and the unfamiliar, i.e., problems of medium difficulty benefit from collective intelligence. For those problems, the impact of collective size is strongest, as is the effect of expertise transfer on judgment quality. Future research on factors that explain task difficulty a priori, and on how difficulty affects the degradation of performance, should thus prove very useful. We expect that high complexity will lead to quasi-random guessing, yet that these guesses will not be truly random, but possibly poorly anchored and adjusted [cf. 9], and thus systematically biased, as we saw with the judgments of weight of air.

Finally, judgment quality appears to deteriorate in stages. When collectives are familiar with the task, their judgment models are highly consistent and accurate. More difficult tasks still provide relatively consistent, but more biased and less accurate results, whereas very difficult tasks lead to random guessing, and inconsistent as well as inaccurate judgments.

7. Acknowledgment

The work described in this paper was supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 150811).

8. References

[1] Anderson, L. R. and Holt, C. A. (1997). “Information Cascades in the Laboratory,” *American Economic Review*, 87, 847-862.

[2] Asch, S.E. (1955). “Opinion and Social Pressure,” *Scientific America*, 193, 5, 31-35.

[3] Bloom, H. (2000). *Global Brain: The Evolution of Mass Mind from the Big Bang to the 21 Century*. John Wiley & Sons, New York.

[4] Bonabeau, E. (2009). “Decision 2.0: The Power of Collective Intelligence,” *MIT Sloan Management Review*, 50, 2, 44-52.

[5] Campbell, D.J. (1988). “Task Complexity: A Review and Analysis,” *Academy of Management Review*, 13, 1, 40-52.

[6] Cohen, J. (1992). "A power primer". *Psychological Bulletin*, 112 (1): 155–159.

[7] Cowan, N. (2001). "The magical number 4 in short-term memory: A reconsideration of mental storage capacity". *Behavioral and Brain Sciences*, 24 (1): 87–114

[8] Dawes, R. (1971). “A Case Study of Graduate Admissions: Application of Three Principles of Human Decision Making”, *American Psychologist*, 180-188.

[9] Epley, N. and Gilovich, T. (2006). “The Anchoring-and-adjustment Heuristic: Why the Adjustments are Insufficient,” *Psychological Science*, 17, 4, 311-318.

[10] Ericsson, K.S. and Charness, N. (1994). “Expert Performance: Its Structure and Acquisition,” *American psychologist*, 49, 725-747.

[11] Ettenson, R., and Shanteau, J. (1987). “Expert Judgment: is more Information Better?” *Psychology Reports*, 60, 227-238.

[12] Faraj, S., Jarvenpaa, S. L., Majchrzak, A. (2011). “Knowledge Collaboration in Online Communities,” *Organization Science*, 22, 5, 1240-1253.

[13] Farnsworth, P.R. and Williams, M.F. (1936). “The Accuracy of the Median and Mean of a Group of Judgments”, *Journal of Psychology*, 7, 237-239.

[14] Gordon, K. (1924). “Group Judgments in the field of Lifted Weights”, *Journal of Experimental Psychology*, 7, 398-400.

[15] Heath, C. and Tversky, A. (1991). “Preference and Belief: Ambiguity and Competence in Choice under Uncertainty,” *Journal of Risk and Uncertainty*, 4, 5-28.

[16] Heylighen, F. (1999). “Collective intelligence and its implementation on the web: Algorithms to develop a collective mental map”. *Computational & Mathematical Organization Theory* 5, 3, 253-280.

[17] Ho, T. H., and Chen, K. Y. (2007). “Discovering and Managing New Product Blockbusters: The Magic and Science of Prediction Markets,” *California Management Review,* 50, 144-58.

[18] Hueffer, K., Fonseca, M., Leiserowitz, A., and Taylor, K. (2013), “The Wisdom of Crowds: Predicting a Weather and Climate-related Event,” *Judgement and Decision Making*, 8, 2, 91-105.

[19] Jeppesen, L. B., Lakhani, K. R. (2010). “Marginality and Problem-solving Effectiveness in Broadcast Research,” *Organization Science*, 21, 5, 1016-1033.

[20] Kawamura, H., and Ohuchi, A. (2000). “Evolutionary emergence of collective intelligence with artificial pheromone communication,” *Proceedings of 26th Annual Conference of the IEEE*, 4, 2831-2836

[21] Kelly, T. L. (1925). “The Applicability of the Spearman-Brown Formula or the Measurement of Reliability,” *Journal of Educational Psychology*, 16, 300-303.

[22] Kimball, D.R., and Holyoak, K. J. (2000). *Transfer and Expertise*, In The Oxford handbook of memory, (Ed.) Tulving, E. and Craik, FIM 109–122. Oxford, UK: Oxford University Press.

[23] Klemz, B., Gruca, T.S. (2003). “Dueling or the Battle Royale? The Impact of Task Complexity on the Evaluation of Entry Threat,” *Psych. Marketing*, 20, 11, 999-1016.

[24] Klugman, S.F. (1945). “Group Judgments for Familiar and Unfamiliar Materials,” *The Journal of General Psychology*, 32, 103-110.

[25] Larrick, R.P., Manners, A. E., and Soll, J. B. (2012). The Social Psychology of the Wisdom of Crowds. In J. I. Krueger (Ed.), *Frontiers in Social Psychology:*

Social Judgment and Decision Making (pp. 227-242). New York: Psychology Press.

[26] Leimeister, J.M. (2010). “Collective Intelligence,” *Business & Information Systems Engineering*, 245-248.

[27] Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D (2011), How social influence can undermine the wisdom of crowd effect, *Proceedings of Natural Academy of Sciences of the United States of America*, 18, 22, 9020-9025.

[28] Lorge, D., Davitz, J., and Brenner, M. (1958). “A survey of Studies Contrasting the Quality of Group Performance and Individual Performance, 1920-1967,” *Psychological Bulletin*, 55, 6, 337-372.

[29] Malone , T. W., Laubacher, R., and Dellarocas, C. (2010), “The Collective Intelligence Genome,” *MIT Sloan Management Review*, 51, 3, 21-31.

[30] Miller, G. A. (1956). "The Magical Number Seven, plus or Minus Two: Some Limits on Our Capacity for Processing Information". *Psychological Review*, 63, 81-97.

[31] Page, S.E. (2007). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press, NJ.

[32] Poole, M. S., Siebold, D. R., and McPhee, R. D. (1985). “Group Decision-making as a Structurational Process,” *Quarterly Journal of Speech*, 71, 1, 74-102.

[33] Reagan-Cinrincione, P. and Rohrbaugh, J. (1992). “Task Bias and the Accuracy of Judgment: Setting a Baseline for Expected Group Performance,” *Journal of Behavior Decision Making*, 5, 233-252.

[34] Servan-Schreiber, E., Wolfers, J., Pennock, D., Galebach, B. (2004). “Prediction Markets: Does Money Matter,” *Electronic Markets*, 14, 3, 243-251.

[35] Shanteau, J. (1992). “How much information does an expert use?” *Acta Psychologica*, 81, 75-86.

[36] Soll, J.B., Manner, A.E., and Larrick, R.P (2011). The wisdom of crowds. In H. Pashler (Ed.), *Encyclopedia of Mind*. Sage Publications.

[37] Speier, C., Vessey, I., and Valacich, J. (2003). “The Effects of Interruptions and Information Presentation Formats on Decision Performance,” *Decision Sciences*, 34, 4, 771-797.

[38] Surowiecki, J. (2005). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*, Random House, New York.

[39] Tversky, A. and Kahneman, D. (1974). “Judgment under Uncertainty: Heuristics and Biases,” *Science,* 185, 1124-1131.

[40] Wagner, C. and Vinaimont, T. (2010), “Evaluating the Wisdom of Crowds,” *Issues in Information Systems*, 11, 1, 724-732.

[41] Wagner, C. and Suh, A. (2013). The Role of Task Difficulty in the Effectiveness of Collective Intelligence, *Proceedings DEST Conference*, Stanford, California, USA.

[42] Wood, R. E. (1986). “Task Complexity: Definition of the Construct,” *Organizational Behavior and Human Decision Processes*, 37, 60-82.

[43] Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., and Malone, T. W. (2010). “Evidence for a Collective Intelligence Factor in the Performance of Human Groups,” *Science*, 29 330, 686-688

[44] Woolley, A. W., and Fuchs, E. (2011). “Collective Intelligence in the Organization of Science,” *Organization Science*, 22, 5, pp. 1359-1367.

[45] Yaniv I, Milyavsky M. (2007). “Using Advice from Multiple Sources to Revise and Improve Judgments,” *Organization Behav Hum Decis Process,* 103, 104-120.