Aaron Brown

Try This at Home, Kids

Testing the wisdom of crowds. For real

Hi folks, captain Numa here again... I was wondering if we could get a show of hands on the best RATE OF DESCENT...?

One of those random memories that stick with me when more important – and certainly more useful – items slip away into the mists is from a high school philosophy class. Although the teacher was anti-inspiring, the material was reasonably interesting, consisting of the standard chapters on the headline philosophers of the Western tradition. The textbook writer had honed complex ideas into easily digestible catch phrases.

The memory is of the teacher ridiculing Aristotle for believing that heavier things fall faster than light things, saying: "*I can't understand why he didn't just try it.*" No doubt, the memory's persistence is explained by the Zeigarnik effect because I did nothing. My immediate thought was to reply: "*Have you tried it?*" then to propose an immediate test. Of course, we would find that heavy objects do fall faster than light ones, at least for the easily available classroom objects of significantly different weights. By the way, it was my inspiring high school physics teacher who showed me the invisible string logical demonstration that heavy objects do not fall faster than light ones, another and more pleasant fresh memory for me.

Perhaps in compensation, I have always had a compulsion to try things for myself. One field that is particularly fertile for this purpose is behavioral economics. It's one thing to read a study; it's another to observe it yourself. Therefore, at the Global Association of Risk Professionals conference in New York this month, I decided to try out four behavioral experiments relevant to risk management, one of which I describe here. Donna Howe, Michael Miller, and Kent Osband were helpful co-presenters in this effort.

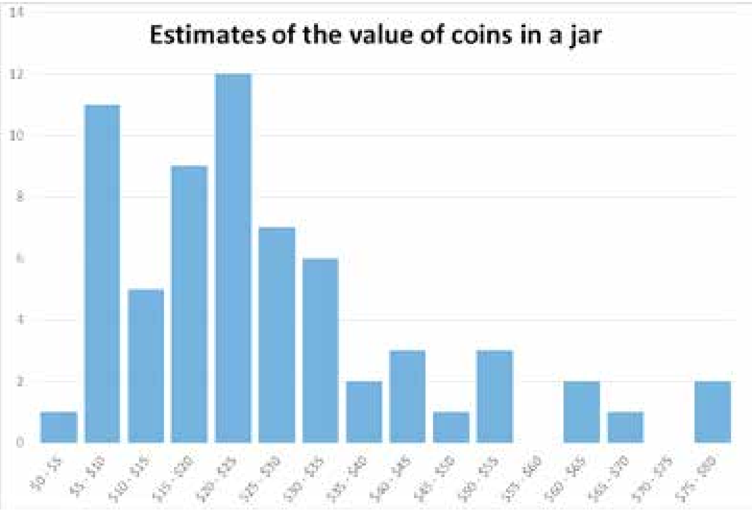
The main motivation was to teach the effects to the audience in a more meaningful way than reading a bunch of small-print PowerPoint slides to them, before running out of time a third of the way through. We did the experiment, then discussed both the academic literature on the subject and our own group experience. A secondary motivation was to see how a group of professional risk managers behaved compared to the university sophomores and others whose results are documented in the literature. I claim no scientific validity to the results; participants were self-selected and may have known of the effects beforehand. No controls were employed, and there was no blinding. I offer the results as anecdotal and suggestive only.

James Suroweicki popularized *The Wisdom of Crowds* effect in his book of that name. Groups can form surprisingly accurate judgments under certain conditions, even on questions about which no individual member has much knowledge. We began by asking 67 people to individually and silently guess the total value of coins in a glass jar (we told participants that the coins consisted only of US currency: quarters, dimes, nickels, and pennies). Figure 1 shows a histogram of their guesses. Five other guesses, up to $150, are omitted from the right tail.

The mean of the guesses was $28.35, quite close to the actual value of $28.19. As is often the case, no individual guess was closer to the correct value than the group mean was. Only four people guessed within $1.00 of the correct amount. In fact, the histogram in Figure 1 shows that individuals had no idea of the correct amount. Guesses ranged from $3 to $150, and there is no obvious clustering around the correct value. Even if you exclude bins with fewer than five individual guesses, the range of common guesses extends from $5 to $35. It does not seem likely that the people who got close were exceptionally accurate coin-value assessors, any more than lottery winners have insight into which numbers will be drawn. I suspect they were guessing with as much uncertainty as anyone else but were a bit luckier in the result. The median was $22.00, much farther from the true value than the mean was. So, somehow the group average extracted accurate information from individuals who didn't know the answer.

An important point, which is a common finding in these types of experiment, is that the deviant right-tail guesses improved the mean estimate. Without the single optimist who thought there was $150 in the jar, the average of the remaining 66 individual guesses was $26.51, an error of $1.68 versus $0.16.

Figure 1: Histogram of the respondents' guesses



Some of the wisdom of crowds is just a mathematical consequence of averaging the kinds of numbers you get from independent guesses. You expect about half the guesses to be on the opposite side of the average from the true value, and also the mean should be closer than some of the tail guesses on the same side. If all the individual guesses are independent draws from a Normal distribution whose mean is the correct value, you expect about 2/π times the square root of the total number of people to have guesses closer than the mean of all the guesses. In this case, with 67 people guessing, we would expect 5 to come closer than the group mean. Therefore even with no wisdom of crowds effect, the mathematics of the situation tend to make the group mean a better estimate than almost all of the individual guesses, in reasonably large groups.

There are three potential problems with the above theory, and they argue for conflicting adjustments to the mean. The first problem is that the individual guesses may not be unbiased estimates of the true value. There may be false consensus. In the extreme, if the bias in individual estimates is large relative to the dispersion of estimates, you might get all estimates on the same side of the true value, in which case you expect half of the individual guesses to be closer to the true value than the mean is. This is a dangerous situation as we will tend to use the dispersion of guesses as an indicator of the probable accuracy of the mean, so not only will the mean be little better than any random individual guess, but we also may have an underestimate of its probable error.

The second problem arises if the accuracy of different individuals varies widely. The accuracy of the average will be dominated by the accuracy of the least accurate guessers, and the effective sample size will be much reduced. In the extreme, one highly deviant estimate can move the mean a lot, and possibly cause the mean to be a worse estimate than all but one of the individual guesses.

Perhaps in compensation, I have always had a compulsion to try things for myself

The third problem is that guesses that are not independent. We tried to guard against this by having people look at the jar individually and write down guesses, and we asked them not to discuss the value of coins with anyone in the room. However, there can still be dependence; for example, different people might be influenced by the same framing or availability heuristic, or perhaps some people are from the same foreign country in which coins are significantly less or more valuable than in the US. For real-world estimates as opposed to contrived experiments, dependence is the rule rather than the exception.

If we suspect we have the first problem, we will value deviant members of the group as they make it more likely that the range of our guesses includes the true value. For the second problem, we would want to exclude outlying guesses. If lack of independence is the issue, we might be tempted to exclude or combine many of the similar observations and overweight the outliers.

The question we investigated for this demonstration was whether groups of individuals could do a better job aggregating their information than could be achieved by a simple average. The 67 people were divided into 12 tables of five or six. We had people go through a structured process that has been shown to produce good results in a variety of situations:

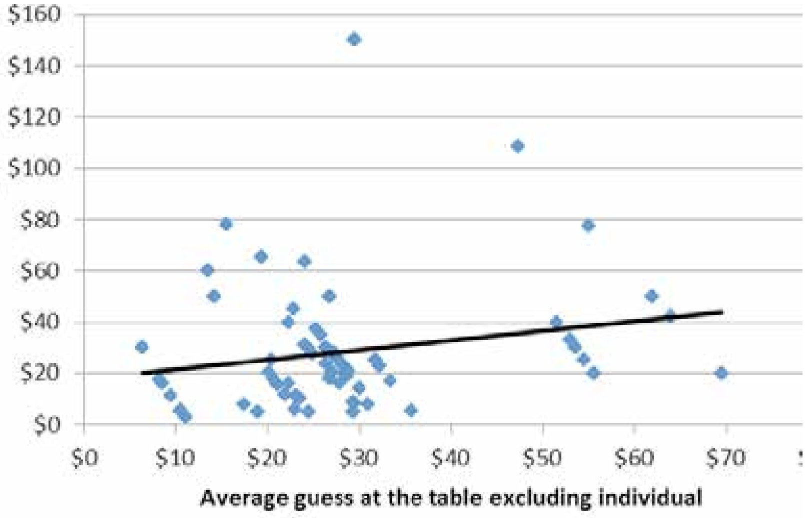
- A designated person at the table names a figure such that he thinks there is about a 10 percent chance that the value of the coins in the jar is below it.
- The person to his left either accepts the figure or names another figure.
- The person on her left either accepts the figure or names a new one. This continues until you get successive acceptances from all but one person at the table (the person who named the last figure is considered to have accepted it). That figure becomes the 'low guess,' L.
- The person who went second for the low guess process names a figure such that he thinks there is about a 10 percent chance that the value of the coins in the jar is above it. Again, the figure goes around the table until someone names a figure for which there are successive acceptances from all but one person at the table. This is the 'high guess,' H.
- Everyone who thinks the true value is greater than (L + H)/2 raises a hand. If exactly half

>

the people raise hands (or as close to half as possible if there are an odd number of people), skip to the next step. If more than half the people raise hands, L is replaced by (L + H)/2. If fewer than half the people raise hands, H is replaced by (L + H)/2. In either of these cases, the step is repeated with the new L or H until a figure is reached that gets half the votes.

- Here's the one that makes people mad. After going through all of this, you throw away the result and agree on a consensus figure by unstructured discussion. The purpose of the first five steps, plus the prior step of writing down an individual guess before any discussion, was to frame the question, make sure everyone in the group contributed, and explore the variety of beliefs and attitudes before anyone stated a direct estimate of the value of coins in the jar.

Figure 2: The 67 individual guesses versus the average of the five other guesses



Overall, discussion was not helpful. Only three of the 12 groups had a better consensus estimate than the average of their individual guesses. This is a common finding (although the process described above is supposed to improve the group decision).

> **It is interesting, however, to note which groups improved their estimates**

The group decision tends to be dominated by the most confident people, who are the least accurate, and also by the highest status people, who tend to be the most insulated from error correction.

It is interesting, however, to note which groups improved their estimates. Two of them were the groups with the second and third least accurate individual estimates. The group of six with the $150 optimist would have had an accurate group average of $29.55 without him, but with him they had a second-worst group average of $49.63. In discussion, they gave less-than-equal weight to his input and settled on $32.50. Although the $150 estimate improved the full 67-person group average, it was too extreme for a group of six, and reducing weight on it made the group consensus better than the average of individual estimates.

The group with the third least accurate individual average included the lowest estimate among all 67 ($3.00). All of the other five guesses were below the correct total, so their average was $9.76. In discussion, they reduced weight on the $3.00 guess and improved slightly to an estimate of $11.25. Although they moved in the correct direction, they didn't gain much because their group lacked diversity. The third and final group to improve started with five members guessing from $18.00 to $40.00 for an accurate average of $29.35, pulled down to $25.29 by the second lowest individual guess of $5.00. Like the first group that improved, they reduced weight on the outlying opinion to get a group consensus of $28.75.

So, three groups with a single outlying opinion reduced the weight given to that opinion and improved their estimate. Two other groups had a single outlier, but gave increased weight to that individual. One group had an average of $21.24, pulled down from $24.48 by a single $5.00 estimate. In discussion, they moved down even further to $20.00, essentially giving extra weight to the outlier. Another group had an average of $20.22, pulled down from $23.06 by a single $6.00 estimate, and moved down in discussion to $12.50.

The other single-outlier group had five individual guesses from $42.00 to $108.37, all of which would have been outlier guesses in other groups, and one $20.00, which was an outlier for this group but would not have been in any other group. Like the three groups that improved their estimates, this group reduced weight on the outlier and had a group consensus of $62.00 versus an individual average of $59.55.

Of the remaining six groups, three had two outliers on opposite sides of the group mean, and three had no outliers. All six of these groups picked consensus numbers worse than the average of their individual guesses. The moral seems to be that discussion usually reduces the weight on an outlier, which seems to help more often than not (three times out of four). When discussion goes in the opposite direction, toward the outlier, it hurts (at least two times out of two). Unless there is a single outlier, group discussion seems to hurt (six times out of six).

You may have noticed that high and low guesses seem to be concentrated at different tables. You are correct. Figure 2 shows each of the 67 individual guesses versus the average of the five other guesses at the same table. There is a strong relationship. Individual guesses average $18 plus 0.4 times the average of the rest of the table. If your five tablemates averaged a guess of $10, on average you will guess $22, but if the other five averaged a guess of $60, on average you will guess $42.

Although we did not run a controlled experiment, we did take precautions to prevent influence. Groups who arrived together were broken up, and people were instructed to move if they knew anyone at their table. Participants were asked to write their guesses silently and privately, and not to discuss anything about the coins or their value with anyone. Although it's possible that some people ignored the rules, it is not uncommon to find evidence of unconscious dependence. Perhaps someone at the table mentioned buying an umbrella on the street for $4; we know that is enough to move everyone's guess toward $4, even though the information is irrelevant to the value of the coins in the jar. Perhaps the view from one table contained visual cues that cause people to be more or less optimistic. Or there could have been communication by body language that influenced guesses. Whatever the reason, take this as an object lesson that assuming independence in any situation is dangerous – however unlikely the possibility of information transmission seems.

Given that only three of 12 tables improved on the average of their individual guesses, it is no surprise that the average of table consensus ($24.42) was further from the true value than the average of all individual guesses; in fact, no table consensus was closer to the true value than the individual average was. But the average table consensus was only $0.80 worse than the average of its individual guesses, while the average of the table consensuses was $3.61 less accurate than the average of the individual guesses. The problem was that tables improved by reducing weight on their outliers, yet these same outliers that made individual tables inaccurate made the overall group of 67 more accurate.

This is a major problem that afflicts rigid and hierarchical organizations. Subgroups function better by ignoring or marginalizing people with deviant opinions, but the overall organization would be improved by averaging the deviant opinions in with everyone else's beliefs.

For the final set of observations we had to resort to a lie. The ethics of lying in human experiments are murky and controversial. Not only is lying wrong in itself, but knowledge that experimenters lie can make future experimentation more difficult. On the other hand, there are things you can't learn if you only speak the truth. We did not consult a human experimentation committee; we just figured the whole thing was informal enough that no harm would be done.

When we collected the individual guesses initially, we did not ask people to write anything except the amount of their guess. But the ushers collecting the papers were instructed to note the table and place number. After the end of the group discussions, I announced that we had failed to collect table and place information from the original guesses. I asked everyone to write their guess a second time, this time with their table and place number on the slip. I was emphatic that we just wanted the number written the first time, not their current guess.

Prior research led me to expect about half the people would write a different number on the second slip than the first, and in almost all cases the change would be in the direction of their table consensus. Although some people might deliberately change, there is zero incentive for that. Careful research seems to indicate that people genuinely misremember prior beliefs after either group discussions or events unfold, and that the memory lapses are usually in the direction of the group consensus or actual events.

Subgroups function better by ignoring or marginalizing people with deviant opinions, but the overall organization would be improved by averaging the deviant opinions in with everyone else's beliefs

The risk managers we sampled turned out to have slightly sharper memories than I expected; 37 of the 67 participants turned in the same number both times. However, all but two of the 30 changes were in the direction of the table consensus. We did observe that round dollar amount initial guesses were less likely to change than guesses with non-zeros to the right of the decimal point, suggesting that it is more an issue of memory than honesty. Only four of the 33 people who guessed round dollar amounts changed their guess, while 26 of the 34 people who guessed non-integer amounts of dollars switched (in 12 cases, switched to an even dollar amount, although rarely the even dollar amount closest to the original guess). For what it's worth, the guesses using pennies – that is, the ones that did not have a zero or five in the second decimal place – were also unlikely to switch. Perhaps the people who picked these numbers had some reason for their choice, or some fondness for certain digits.

We concluded the session with an open discussion. The crowd was pretty negative about our discussion structure, feeling that it led to a contentious and inflexible discussion. In a way, that's the point; human social instincts, which are far stronger than conscious calculation or the desire for good organizational decisions, push us to consider good feelings and an answer everyone can accept over getting to the truth. Structured discussions are designed to frustrate those instincts; they aim to reveal honest individual opinions without prejudice, and to let the group determine who is in step with the majority and who is not. However, those theoretical advantages did not produce a good result – 75 percent of the consensus estimates were less accurate than simple averages of individual guesses – so I have to side with the participants here.

Although there is a strong aspect of fun and games to this demonstration, I think it illustrates serious points about group decision making. There is a wisdom in crowds, but only under the right conditions. Simple exercises like this are important for learning and building intuition. If you found this article informative, you will learn much more by trying this or a similar experiment for yourself. Try this at home, kids.